数据竞赛年鉴

2019 年度

出品方：Kaggle 数据竞赛星球

合作伙伴：

Kaggle 竞赛宝典、Coggle、Datawhale

前言

为了更好的对数据竞赛的比赛进行整理、复盘和学习，我们（Kaggle 竞赛宝典、Coggle、Datawhale）对 2019 年国内外常见竞赛平台的竞赛进行了整理，形成此书。

当然由于整理的时间有限，我们不能把所有的比赛细节和方案收集进来，如果遇到任何内容问题，请联系我（微信号：Finlay-LYZ），我们也将持续更新相关内容。

微信公众号：kaggle 竞赛宝典

数据竞赛加分骚操作 & 数据分析方法 & 实践机器学习 & Kaggle + 天池 + 其他

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

微信公众号：Datawhale

一个专注于 AI 领域的开源学习组织，汇聚了众多领域院校和知名企业的优秀学习者，聚集了一群有开源精神和探索精神的团队成员。愿景-for the learner，和学习者一起成长。

机器学习理论与数据竞赛实战：https://zhuanlan.zhihu.com/DataAI

知乎 1W 关注的专栏，总阅读量 100w+

目录

00 竞赛年度发展

0.1 数据竞赛发展趋势

2019 年是不平凡的一年，是 AI 应用爆发和落地的一年，也是数据竞赛开花结果的一年。据不完全统计，中国境内 2019 年举办的数据竞赛超过 150 场，赛题总奖金千万级别。数据竞赛不断开花结果，吸引政府、高校、科研机构和企业参与其中，竞赛主题覆盖面愈来愈广。

（图取自数据竞赛白皮书，和鲸出品）

1. **数据竞赛与科研联系愈加紧密**。学术会议如 CVPR、ICCV 和 WSDM 都开设了 Workshop 竞赛，这些学术竞赛不仅提供了宝贵的学术数据集，也提供了公平的评测环境。2019 年 BERT 模型大放异彩，基本上横扫了所有的自然语言处理竞赛，学术进步推动了数据竞赛的发展，数据竞赛也不断验证和产生新的模型。

2. **数据竞赛的形式和应用领域更加多样。**数据竞赛的形式不单单有算法赛，今年还开设了不少的方案赛和可视化比赛，丰富了参赛的形式。医疗和工业领域的数据竞赛如雨后春笋不断出现。

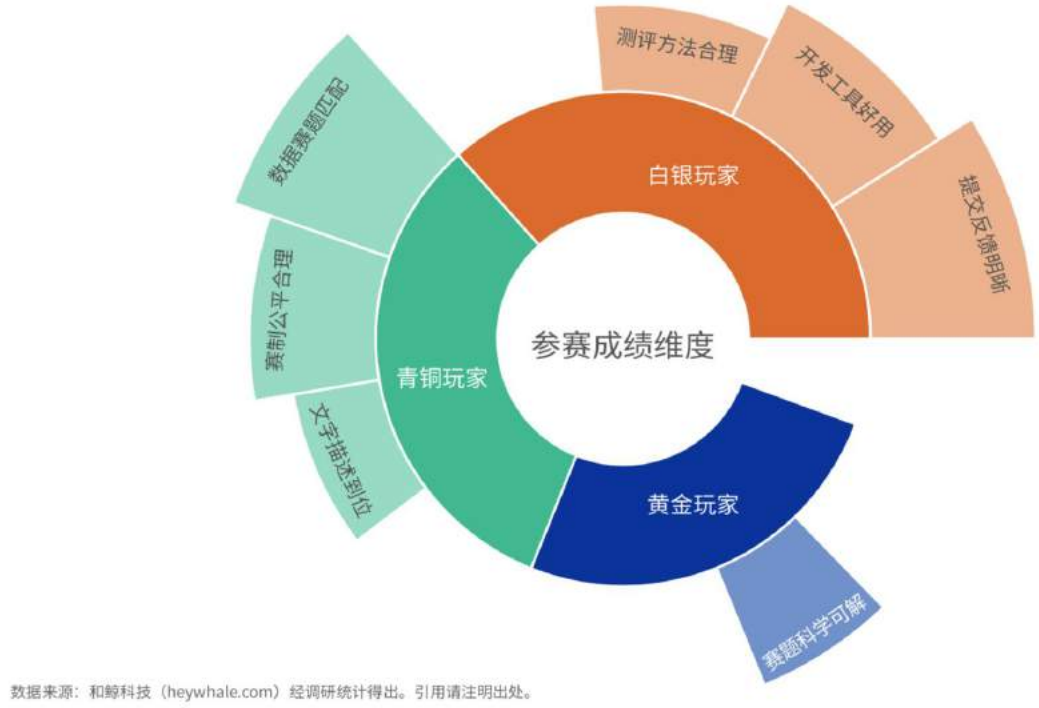
3. **数据竞赛的评测机制更加健全。**今年 Kaggle 很多竞赛都要求使用 kernel 提交，国内天池的一些比赛也要求通过 docker 环境提交。通过虚拟环境的提交方式，不仅对提交代码的模型大小和时间复杂度均有要求，也更加适合后期落地。

4. **数据竞赛平台和组织形式更加多样。**今年数据竞赛平台更加多样，组织比赛的形式更加丰富，有线上赛题、线下赛题和马拉松赛题。竞赛评测形式也更加多样，AI 研习社平台开设了及时评测并分奖金的活动，DataFountain 很多竞赛都开设了周冠军的奖励。

0.2 竞赛平台热度对比

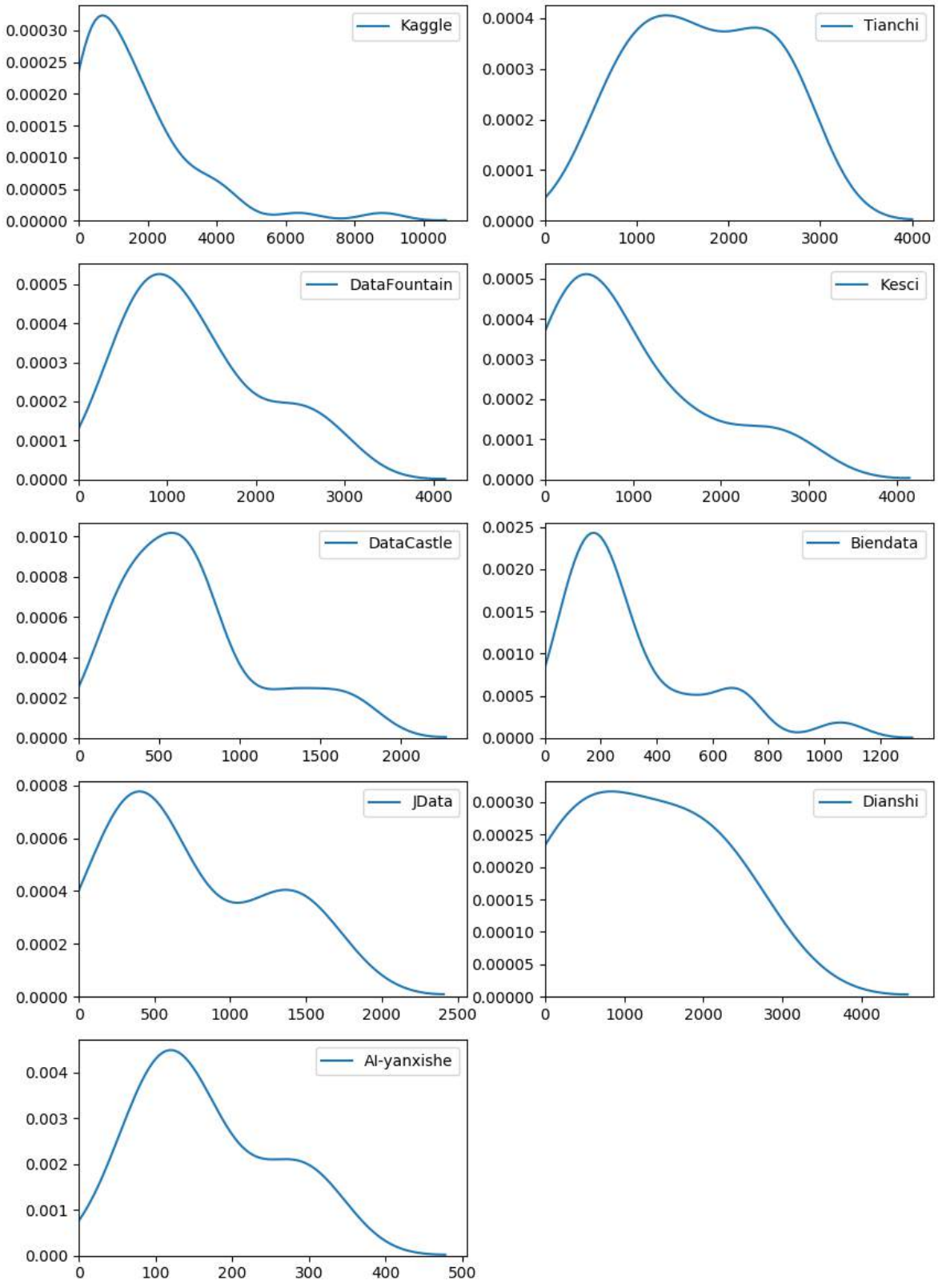
国内竞赛平台的热度（参赛人数）主要受以下几个因素影响：

- 比赛平台的易用性；
- 比赛赛题的内容；
- 比赛赛题的奖金；



（图取自数据竞赛白皮书，和鲸出品）

为了更加直观的对比赛题热度和比赛平台的关系，我们对国内各大竞赛平台的每月参赛队伍数量进行了可视化（按照比赛平台的参数队伍的 sns.distplot 图）：



从上图可以看出，Kaggle 平台上的赛题人数大都集中在 1000-2000 人之间。国内竞赛平台中天池的热度最高，平均每场比赛都有 1000+队伍参加。国内竞赛平台的参赛人数主要与平台曝光量和日常用户量相关，1000 只参赛队伍是一个赛题的分水岭。

0.3 年度竞赛评比

最佳人气奖：Kaggle 平台，Santander Customer Transaction Prediction，参赛队伍：8802，参赛人数：9838。

Santander Customer Transaction Prediction 是结构化比赛，数据量也非常小，非常适合入门学习。

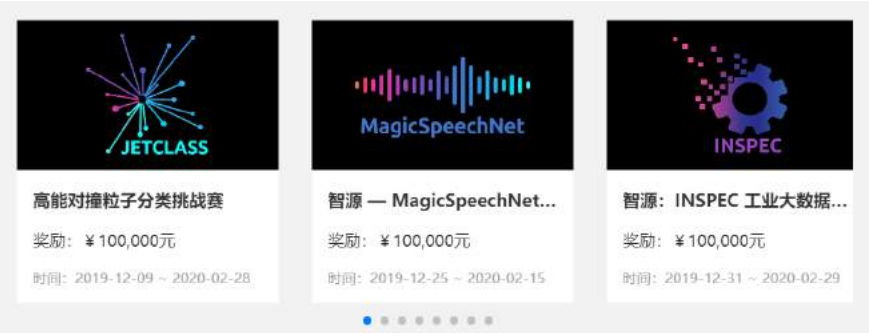
最佳开源奖：DataFountain 平台，CCF 大数据与计算智能大赛（CCF Big Data & Computing Intelligence Contest， CCF BDCI）。

今年 CCF DBCI 将每一个的比赛 Top 选手方案都整理，并在社区中进行了开源分享：https://discussion.datafountain.cn/topics/1077

竞赛方案的开源能够最大程度的促进数据竞赛的成果转化，也为社区保留了最大的价值和贡献。此外 DataFountain 举办的 CCF BDCI 比赛对比赛小号、同 IP 提交和答案相似度校验都做的比较完善，最大程度杜绝了比赛违规的情况，这种做法值得推广和赞扬。

最佳学术奖：biendata 平台，智源人工智能大赛

今年北京人工智能研究院在 biendata 平台举办了智源人工智能大赛，该赛事共包括 10 余项赛事，赛题内容与学术内容联系紧密，有较高的学术价值。



但由于数据集整理比较复杂，biendata 平台今年比赛出现了多次比赛 leak 的现象，比赛数据的整理和清晰工作有待加强。

01 年度数据竞赛汇总

1.1 Kaggle

Human Protein Atlas Image Classification

https://www.kaggle.com/c/human-protein-atlas-image-classification/overview

参赛队伍：2169，比赛类型：Featured，比赛数据：图像

- 比赛背景

In this competition, Kagglers will develop models capable of classifying mixed patterns of proteins in microscope images. The Human Protein Atlas will use these models to build a tool integrated with their smart-microscopy system to identify a protein's location(s) from a high-throughput image.

Proteins are “the doers” in the human cell, executing many functions that together enable life. Historically, classification of proteins has been limited to single patterns in one or a few cell types, but in order to fully understand the complexity of the human cell, models must classify mixed patterns across a range of different human cells.

Images visualizing proteins in cells are commonly used for biomedical research, and these cells could hold the key for the next breakthrough in medicine. However, thanks to advances in high-throughput microscopy, these images are generated at a far greater pace than what can be manually evaluated. Therefore, the need is greater than ever for automating biomedical image analysis to accelerate the understanding of human cells and disease.

Nature Methods has indicated interest in considering a paper discussing the outcome and approaches of the challenge. The Human Protein Atlas team would like to invite top performing teams to join as co-authors in the writing of this paper.

Top performing teams will also be eligible to compete for the special prize. Additional information for both the special prize and co-authoring for Nature Methods will become available through the Discussion posts once the main competition is complete.

Data Competition in 2019

- 评价指标

Submissions will be evaluated based on their macro F1 score.

- 比赛数据

https://www.kaggle.com/c/human-protein-atlas-image-classification/data

- 时间轴：Oct 4, 2018 - January 10, 2019
- 比赛结果

https://www.kaggle.com/c/human-protein-atlas-image-classification/leaderboard

- 赛后分享

3rd Place Solution：https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77320

5th Place Solution：https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77731

7th Place Solution：https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77269

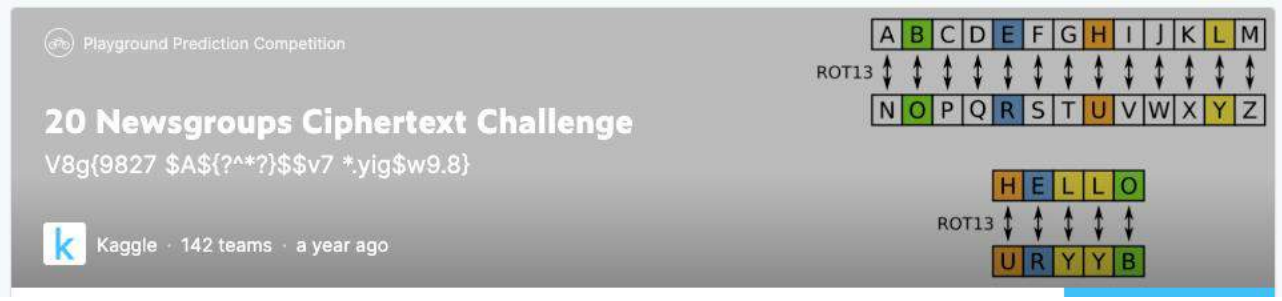
12th Place Solution：https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77325

15th Place Solution：https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77322

20 Newsgroups Ciphertext Challenge

https://www.kaggle.com/c/20-newsgroups-ciphertext-challenge

参赛队伍：142，比赛类型：Playground，比赛数据：文本



- 比赛背景

This isn't your classic decoder ring puzzle found in a cereal box. There's a twist.

Welcome to the Ciphertext Challenge! In this competition, we've encrypted parts of a well-known dataset -- the 20 Newsgroups dataset -- with several simple, classic ciphers. This dataset is commonly used as a multi-class and NLP sample set, noted for its small size, varied nature, and the first-hand look it offers into the deep existential horrors of the 90s-era internet. With 20 fairly distinct classes and lots of clues, it allows for a wide variety of successful approaches.

We've made the problem a little harder to solve.

Data Competition in 2019

Fabulous Kaggle swag will go to the top competitors - the highest-scoring teams (which might be the first to crack the code!), and the most popular kernel. Note that this is a short competition, so use your submissions wisely.

- 评价指标

Submissions will be evaluated based on their macro F1 score.

- 比赛数据

https://www.kaggle.com/c/20-newsgroups-ciphertext-challenge/data

- 时间轴：Dec 14, 2018 - January 16th, 2019
- 比赛结果

https://www.kaggle.com/c/20-newsgroups-ciphertext-challenge/leaderboard

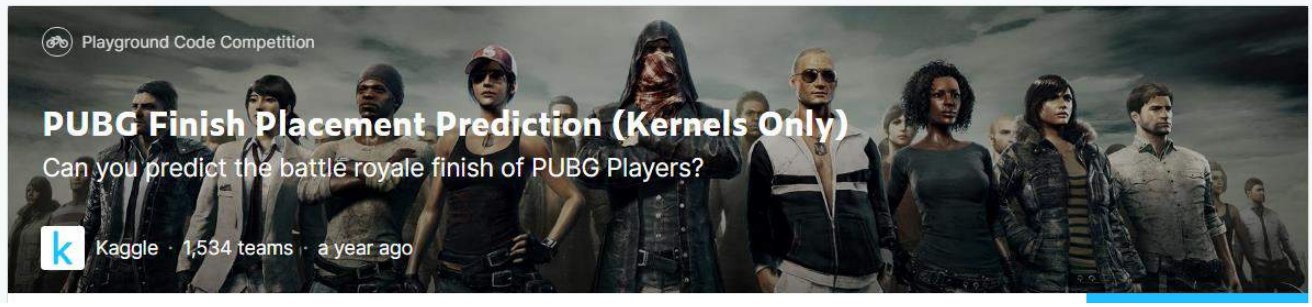
- 赛后分享

1st place solution：https://www.kaggle.com/c/20-newsgroups-ciphertext-challenge/discussion/77894

PUBG Finish Placement Prediction (Kernels Only)

https://www.kaggle.com/c/pubg-finish-placement-prediction

参赛队伍：1534，比赛类型：Playground，比赛数据：结构化



- 比赛背景

So, where we droppin' boys and girls?

Battle Royale-style video games have taken the world by storm. 100 players are dropped onto an island empty-handed and must explore, scavenge, and eliminate other players until only one is left standing, all while the play zone continues to shrink.

PlayerUnknown's BattleGrounds (PUBG) has enjoyed massive popularity. With over 50 million copies sold, it's the fifth best selling game of all time, and has millions of active monthly players.

Data Competition in 2019

The team at PUBG has made official game data available for the public to explore and scavenge outside of "The Blue Circle." This competition is not an official or affiliated PUBG site - Kaggle collected data made possible through the PUBG Developer API.

You are given over 65,000 games' worth of anonymized player data, split into training and testing sets, and asked to predict final placement from final in-game stats and initial player ratings.

What's the best strategy to win in PUBG? Should you sit in one spot and hide your way into victory, or do you need to be the top shot? Let's let the data do the talking!

- 评价指标

Submissions are evaluated on Mean Absolute Error between your predicted winPlacePerc and the observed winPlacePerc.

- 比赛数据

https://www.kaggle.com/c/pubg-finish-placement-prediction/data

- 时间轴：Oct 5, 2018 – Jan 31, 2019
- 比赛结果

https://www.kaggle.com/c/pubg-finish-placement-prediction/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/pubg-finish-placement-prediction/discussion/79161

Reducing Commercial Aviation Fatalities

https://www.kaggle.com/c/reducing-commercial-aviation-fatalities

参赛队伍：178，参赛队伍 193

比赛类型：Playground，比赛数据：结构化

- 比赛背景

Data Competition in 2019

Most flight-related fatalities stem from a loss of “airplane state awareness.” That is, ineffective attention management on the part of pilots who may be distracted, sleepy or in other dangerous cognitive states.

Your challenge is to build a model to detect troubling events from aircrew’s physiological data. You'll use data acquired from actual pilots in test situations, and your models should be able to run calculations in real time to monitor the cognitive states of pilots. With your help, pilots could then be alerted when they enter a troubling state, preventing accidents and saving lives.

Reducing aircraft fatalities is just one of the complex problems that Booz Allen Hamilton has been solving for business, government, and military leaders for over 100 years. Through devotion, candor, courage, and character, they produce original solutions where there are no roadmaps. Now you can help them find answers, save lives, and change the world.

- 评价指标

Submissions are evaluated on the Multi Class Log Loss between the predicted probabilities and the observed target.

- 比赛数据

https://www.kaggle.com/c/reducing-commercial-aviation-fatalities/data

- 时间轴：Dec 20, 2018 – Feb 13, 2019

- 比赛结果

https://www.kaggle.com/c/reducing-commercial-aviation-fatalities/leaderboard

- 赛后分享

8th place solution：https://www.kaggle.com/c/reducing-commercial-aviation-fatalities/discussion/84527

Quora Insincere Questions Classification (Kernels Only)

https://www.kaggle.com/c/quora-insincere-questions-classification

参赛队伍：4037

比赛类型：Featured，比赛数据：文本

- 比赛背景

An existential problem for any major website today is how to handle toxic and divisive content. Quora wants to tackle this problem head-on to keep their platform a place where users can feel safe sharing their knowledge with the world.

Quora is a platform that empowers people to learn from each other. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers. A key challenge is to weed out insincere questions -- those founded upon false premises, or that intend to make a statement rather than look for helpful answers.

In this competition, Kagglers will develop models that identify and flag insincere questions. To date, Quora has employed both machine learning and manual review to address this problem. With your help, they can develop more scalable methods to detect toxic and misleading content.

Here's your chance to combat online trolls at scale. Help Quora uphold their policy of “Be Nice, Be Respectful” and continue to be a place for sharing and growing the world’s knowledge.

- 评价指标

Submissions are evaluated on F1 Score between the predicted and the observed targets.

- 比赛数据

https://www.kaggle.com/c/quora-insincere-questions-classification/data

- 时间轴：Nov 7, 2018 – Feb 14, 2019

- 比赛结果

https://www.kaggle.com/c/quora-insincere-questions-classification/leaderboard

- 赛后分享

3rd place solution：https://www.kaggle.com/wowfattie/3rd-place

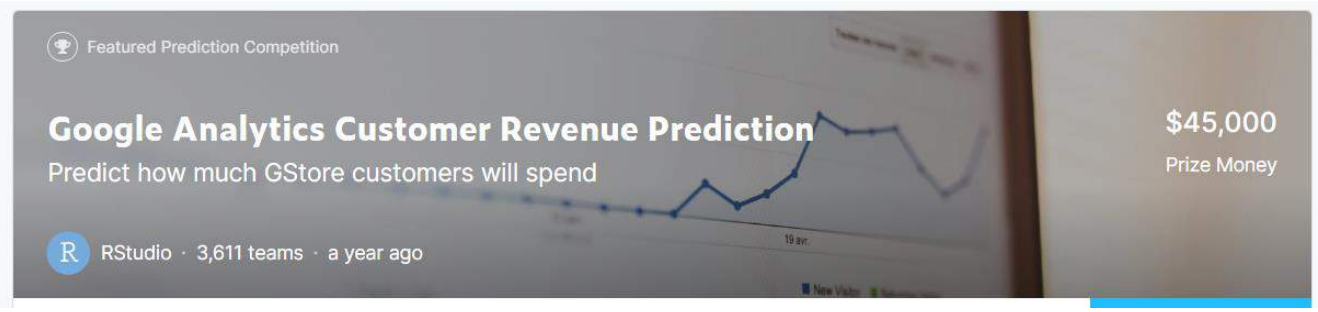
10th place solution：https://www.kaggle.com/tks0123456789/pme-ema-6-x-8-epochs

Google Analytics Customer Revenue Prediction

https://www.kaggle.com/c/ga-customer-revenue-prediction

参赛队伍：3611，参赛人数：4183

比赛类型：Featured，比赛数据：结构化



- 比赛背景

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

RStudio, the developer of free and open tools for R and enterprise-ready products for teams to scale and share work, has partnered with Google Cloud and Kaggle to demonstrate the business impact that thorough data analysis can have.

In this competition, you're challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of GA data.

- 评价指标

Submissions are scored on the root mean squared error.

- 比赛数据

https://www.kaggle.com/c/ga-customer-revenue-prediction/data

- 时间轴：Sep 14, 2018 – Feb 22, 2019
- 比赛结果

https://www.kaggle.com/c/ga-customer-revenue-prediction/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/ga-customer-revenue-prediction/discussion/82614

Elo Merchant Category Recommendation

https://www.kaggle.com/c/elo-merchant-category-recommendation

参赛队伍：4127，参赛人数：4729

比赛类型：Featured，比赛数据：结构化

- 比赛背景

Imagine being hungry in an unfamiliar part of town and getting restaurant recommendations served up, based on your personal preferences, at just the right moment. The recommendation comes with an attached discount from your credit card provider for a local place around the corner!

Right now, Elo, one of the largest payment brands in Brazil, has built partnerships with merchants in order to offer promotions or discounts to cardholders. But do these promotions work for either the consumer or the merchant? Do customers enjoy their experience? Do merchants see repeat business? Personalization is key.

Elo has built machine learning models to understand the most important aspects and preferences in their customers' lifecycle, from food to shopping. But so far none of them is specifically tailored for an individual or profile. This is where you come in.

In this competition, Kagglers will develop algorithms to identify and serve the most relevant opportunities to individuals, by uncovering signal in customer loyalty. Your input will improve customers' lives and help Elo reduce unwanted campaigns, to create the right experience for customers.

- 评价指标

Submissions are scored on the root mean squared error.

- 比赛数据

https://www.kaggle.com/c/elo-merchant-category-recommendation/data

Data Competition in 2019

- 时间轴：Nov 27, 2018 – Feb 27, 2019
- 比赛结果

https://www.kaggle.com/c/elo-merchant-category-recommendation/leaderboard

- 赛后分享

5th place solution：https://www.kaggle.com/c/elo-merchant-category-recommendation/discussion/82314

7th place solution： https://www.kaggle.com/c/elo-merchant-category-recommendation/discussion/82055

10th place solution：https://www.kaggle.com/c/elo-merchant-category-recommendation/discussion/82093

11th place solution：https://www.kaggle.com/c/elo-merchant-category-recommendation/discussion/82127

Humpback Whale Identification

https://www.kaggle.com/c/humpback-whale-identification

参赛队伍：2129，参赛人数：2460

比赛类型：Featured，比赛数据：图像

- 比赛背景

After centuries of intense whaling, recovering whale populations still have a hard time adapting to warming oceans and struggle to compete every day with the industrial fishing industry for food.

To aid whale conservation efforts, scientists use photo surveillance systems to monitor ocean activity. They use the shape of whales' tails and unique markings found in footage to identify what species of whale they're analyzing and meticulously log whale pod dynamics and movements. For the past 40 years, most of this work has been done manually by individual scientists, leaving a huge trove of data untapped and underutilized.

In this competition, you're challenged to build an algorithm to identify individual whales in images. You'll analyze Happywhale's database of over 25,000 images, gathered from research institutions and public contributors. By contributing, you'll help to open rich fields of understanding for marine mammal population dynamics around the globe.

Data Competition in 2019

Note, this competition is similar in nature to this competition with an expanded and updated dataset.

We'd like to thank Happywhale for providing this data and problem. Happywhale is a platform that uses image process algorithms to let anyone to submit their whale photo and have it automatically identified.

- 评价指标

Submissions are evaluated according to the Mean Average Precision @ 5 (MAP@5).

- 比赛数据

https://www.kaggle.com/c/humpback-whale-identification/data

- 时间轴：Dec 1, 2018 – Mar 1, 2019
- 比赛结果

https://www.kaggle.com/c/humpback-whale-identification/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82366

2nd place solution： https://www.kaggle.com/c/humpback-whale-identification/discussion/83885

3rd place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82484

5th place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82369

4th place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82356

7th place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82352

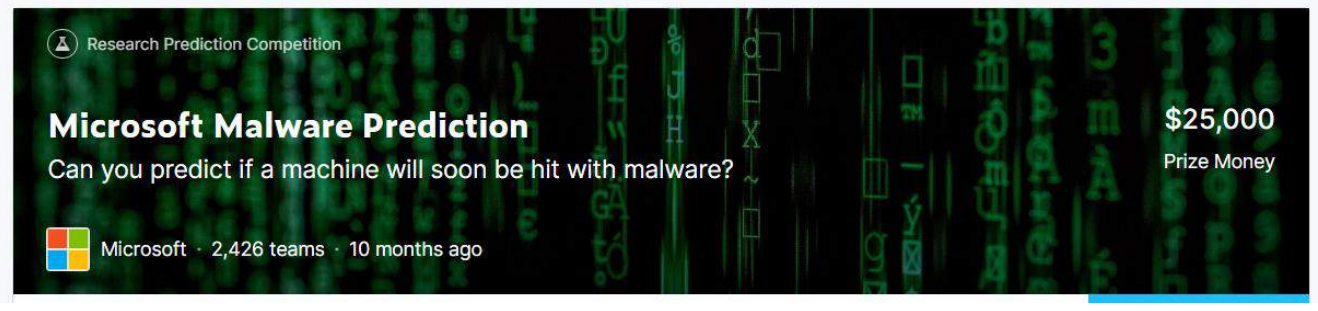
11th place solution：https://www.kaggle.com/c/humpback-whale-identification/discussion/82430

Microsoft Malware Prediction

https://www.kaggle.com/c/microsoft-malware-prediction

参赛队伍：2426，参赛人数：2874

比赛类型：Featured，比赛数据：结构化



Data Competition in 2019

- 比赛背景

The malware industry continues to be a well-organized, well-funded market dedicated to evading traditional security measures. Once a computer is infected by malware, criminals can hurt consumers and enterprises in many ways.

With more than one billion enterprise and consumer customers, Microsoft takes this problem very seriously and is deeply invested in improving security.

As one part of their overall strategy for doing so, Microsoft is challenging the data science community to develop techniques to predict if a machine will soon be hit with malware. As with their previous, Malware Challenge (2015), Microsoft is providing Kagglers with an unprecedented malware dataset to encourage open-source progress on effective techniques for predicting malware occurrences.

Can you help protect more than one billion machines from damage BEFORE it happens?

- 评价指标

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed label.

- 比赛数据

https://www.kaggle.com/c/microsoft-malware-prediction/data

- 时间轴：Dec 14, 2018 – Mar 14, 2019
- 比赛结果

https://www.kaggle.com/c/microsoft-malware-prediction/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/microsoft-malware-prediction/discussion/84065

4th place solution：https://www.kaggle.com/c/microsoft-malware-prediction/discussion/84515

6th place solution：https://www.kaggle.com/c/microsoft-malware-prediction/discussion/84112

7th place solution：https://www.kaggle.com/c/microsoft-malware-prediction/discussion/86888

9th place solution：https://www.kaggle.com/c/microsoft-malware-prediction/discussion/84570

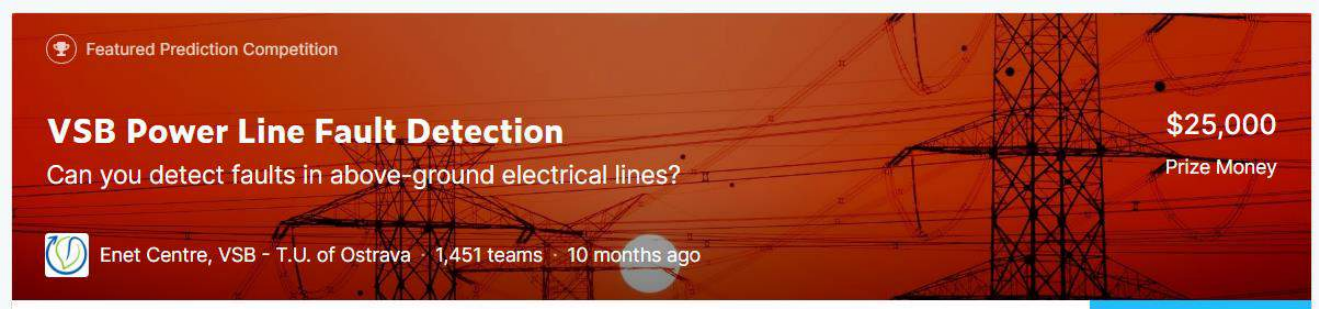
VSB Power Line Fault Detection

https://www.kaggle.com/c/vsb-power-line-fault-detection

参赛队伍：1451，参赛人数：1595

Data Competition in 2019

比赛类型：Featured，比赛数据：结构化



- 比赛背景

Medium voltage overhead power lines run for hundreds of miles to supply power to cities. These great distances make it expensive to manually inspect the lines for damage that doesn't immediately lead to a power outage, such as a tree branch hitting the line or a flaw in the insulator. These modes of damage lead to a phenomenon known as partial discharge — an electrical discharge which does not bridge the electrodes between an insulation system completely. Partial discharges slowly damage the power line, so left unrepaired they will eventually lead to a power outage or start a fire.

Your challenge is to detect partial discharge patterns in signals acquired from these power lines with a new meter designed at the ENET Centre at VŠB. Effective classifiers using this data will make it possible to continuously monitor power lines for faults.

ENET Centre researches and develops renewable energy resources with the goal of reducing or eliminating harmful environmental impacts. Their efforts focus on developing technology solutions around transportation and processing of energy raw materials.

By developing a solution to detect partial discharge you'll help reduce maintenance costs, and prevent power outages.

- 评价指标

Submissions are evaluated on the Matthews correlation coefficient(MCC) between the predicted and the observed response.

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

- 比赛数据

https://www.kaggle.com/c/vsb-power-line-fault-detection/data

- 时间轴：Dec 21, 2018 – Mar 22, 2019

Data Competition in 2019

- 比赛结果

https://www.kaggle.com/c/vsb-power-line-fault-detection/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/vsb-power-line-fault-detection/discussion/86616

6th place solution：https://www.kaggle.com/c/vsb-power-line-fault-detection/discussion/85170

9th place solution：https://www.kaggle.com/c/vsb-power-line-fault-detection/discussion/85258

Histopathologic Cancer Detection

https://www.kaggle.com/c/histopathologic-cancer-detection

参赛队伍：1157，参赛人数：1355

比赛类型：Playground，比赛数据：图像

- 比赛背景

In this competition, you must create an algorithm to identify metastatic cancer in small image patches taken from larger digital pathology scans. The data for this competition is a slightly modified version of the PatchCamelyon (PCam) benchmark dataset (the original PCam dataset contains duplicate images due to its probabilistic sampling, however, the version presented on Kaggle does not contain duplicates).

- 评价指标

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/histopathologic-cancer-detection/data

- 时间轴：Nov 17, 2018 – Mar 31, 2019

Data Competition in 2019

- 比赛结果

https://www.kaggle.com/c/histopathologic-cancer-detection/leaderboard

- 赛后分享

17th place solution：https://www.kaggle.com/c/histopathologic-cancer-detection/discussion/87397

Google Cloud & NCAA® ML Competition 2019-Women's

https://www.kaggle.com/c/womens-machine-learning-competition-2019

参赛队伍：500，参赛人数：527

比赛类型：Featured，比赛数据：结构化

- 比赛背景

As a result of the continued collaboration between Google Cloud and the NCAA®, the sixth annual Kaggle-backed March Madness competition is underway! Another year, another chance to anticipate the upsets, call the probabilities, and put your bracketology skills to the leaderboard test. Kagglers will join the millions of fans who attempt to forecast the outcomes of March Madness during this year's NCAA Division I Men's and Women's Basketball Championships. But unlike most fans, you will pick your bracket using a combination of NCAA's historical data and your computing power, while the ground truth unfolds on national television.

- 评价指标：logloss
- 比赛数据

https://www.kaggle.com/c/womens-machine-learning-competition-2019/data

- 时间轴：Feb 15, 2019 – Apr 8, 2019
- 比赛结果

https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion

- 赛后分享

Data Competition in 2019

1st place solution：https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion/88451

2nd place solution：https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion/88402

3rd place solution：https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion/90156

4th place solution：https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion/88462

5th place solution：https://www.kaggle.com/c/womens-machine-learning-competition-2019/discussion/90305

Google Cloud & NCAA® ML Competition 2019-Men's

https://www.kaggle.com/c/mens-machine-learning-competition-2019

参赛队伍：866，参赛人数：951

比赛类型：Featured，比赛数据：结构化

- 比赛背景

As a result of the continued collaboration between Google Cloud and the NCAA, the sixth annual Kaggle-backed March Madness competition is underway! Another year, another chance to anticipate the upsets, call the probabilities, and put your bracketology skills to the leaderboard test. Kagglers will join the millions of fans who attempt to forecast the outcomes of March Madness during this year's NCAA Division I Men's and Women's Basketball Championships. But unlike most fans, you will pick your bracket using a combination of NCAA's historical data and your computing power, while the ground truth unfolds on national television.

- 评价指标：logloss
- 比赛数据

https://www.kaggle.com/c/mens-machine-learning-competition-2019/data

- 时间轴：Feb 15, 2019 – Apr 8, 2019
- 比赛结果

Data Competition in 2019

https://www.kaggle.com/c/mens-machine-learning-competition-2019/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/mens-machine-learning-competition-2019/discussion/89150

2nd place solution：https://www.kaggle.com/c/mens-machine-learning-competition-2019/discussion/88805

3rd place solution：https://www.kaggle.com/c/mens-machine-learning-competition-2019/discussion/90254

4th place solution：https://www.kaggle.com/c/mens-machine-learning-competition-2019/discussion/89645

5th place solution：https://www.kaggle.com/c/mens-machine-learning-competition-2019/discussion/89942

PetFinder.my Adoption Prediction

https://www.kaggle.com/c/petfinder-adoption-prediction

参赛队伍：2023，参赛人数：3169

比赛类型：Featured，比赛数据：结构化+图像

- 比赛背景

Millions of stray animals suffer on the streets or are euthanized in shelters every day around the world. If homes can be found for them, many precious lives can be saved — and more happy families created.

PetFinder.my has been Malaysia's leading animal welfare platform since 2008, with a database of more than 150,000 animals. PetFinder collaborates closely with animal lovers, media, corporations, and global organizations to improve animal welfare.

Animal adoption rates are strongly correlated to the metadata associated with their online profiles, such as descriptive text and photo characteristics. As one example, PetFinder is currently experimenting with a simple AI tool called the Cuteness Meter, which ranks how cute a pet is based on qualities present in their photos.

Data Competition in 2019

In this competition you will be developing algorithms to predict the adoptability of pets - specifically, how quickly is a pet adopted? If successful, they will be adapted into AI tools that will guide shelters and rescuers around the world on improving their pet profiles' appeal, reducing animal suffering and euthanization.

- 评价指标：

Submissions are scored based on the quadratic weighted kappa, which measures the agreement between two ratings.

- 比赛数据

https://www.kaggle.com/c/petfinder-adoption-prediction/data

- 时间轴：Dec 28, 2018 – Apr 10, 2019
- 比赛结果

https://www.kaggle.com/c/petfinder-adoption-prediction/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/naka2ka/stack-480-speedup-groupkfold-with-no-dict

2nd place solution：https://www.kaggle.com/wuyhbb/final-small

8th place solution：https://www.kaggle.com/adityaecdrid/8th-place-solution-code

9th place solution：https://www.kaggle.com/chizhu2018/final-submit-two-10th-solution-private-0-442

14 place solution：https://www.kaggle.com/baomengjiao/fork-of-v5-fork-v4-change-gpu-add-name-feature

Santander Customer Transaction Prediction

https://www.kaggle.com/c/santander-customer-transaction-prediction

参赛队伍：8802，参赛人数：9838

比赛类型：Featured，比赛数据：结构化

Data Competition in 2019

- 比赛背景

At Santander our mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals.

Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan?

In this challenge, we invite Kagglers to help us identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted. The data provided for this competition has the same structure as the real data we have available to solve this problem.

- 评价指标：

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/santander-customer-transaction-prediction/data

- 时间轴：Dec 28, 2018 – Apr 10, 2019
- 比赛结果

https://www.kaggle.com/c/santander-customer-transaction-prediction/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/santander-customer-transaction-prediction/discussion/89003

2nd place solution：https://www.kaggle.com/c/santander-customer-transaction-prediction/discussion/88939

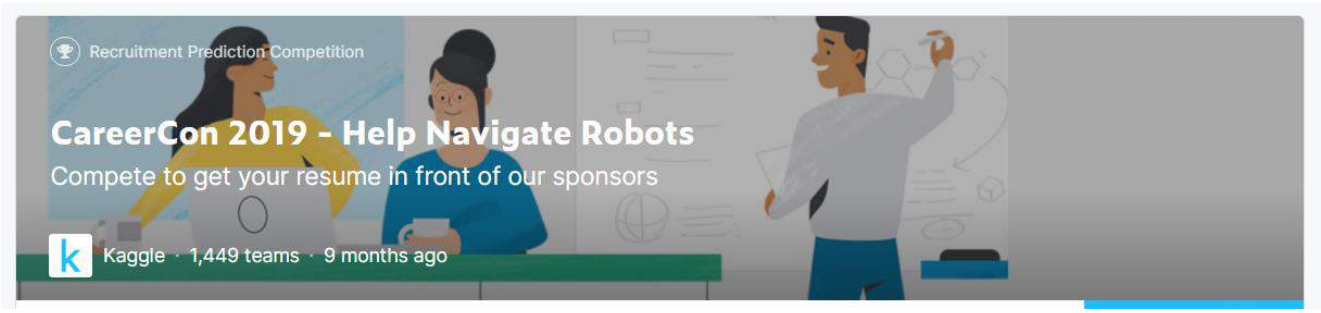
CareerCon 2019 - Help Navigate Robots

https://www.kaggle.com/c/career-con-2019

参赛队伍：1449，参赛人数：1449

比赛类型：Featured，比赛数据：结构化

Data Competition in 2019



- 比赛背景

CareerCon is a digital event all about landing your first data science job — and registration is now open! Ahead of the event, we have a fun competition to get you started. See below for a unique challenge and opportunity to share your resume with select CareerCon sponsors.

Robots are smart… by design. To fully understand and properly navigate a task, however, they need input about their environment.

In this competition, you'll help robots recognize the floor surface they're standing on using data collected from Inertial Measurement Units (IMU sensors).

We've collected IMU sensor data while driving a small mobile robot over different floor surfaces on the university premises. The task is to predict which one of the nine floor types (carpet, tiles, concrete) the robot is on using sensor data such as acceleration and velocity. Succeed and you'll help improve the navigation of robots without assistance across many different surfaces, so they won't fall down on the job.

- 评价指标：

Submissions are evaluated on Multiclass Accuracy, which is simply the average number of observations with the correct label.

- 比赛数据

https://www.kaggle.com/c/career-con-2019/data

- 时间轴：Mar 14, 2018 – Apr 12, 2019
- 比赛结果

https://www.kaggle.com/c/career-con-2019/leaderboard

- 赛后分享

3rd place solution：https://www.kaggle.com/c/career-con-2019/discussion/89181

Gendered Pronoun Resolution

https://www.kaggle.com/c/gendered-pronoun-resolution

参赛队伍：838

比赛类型：Research，比赛数据：图像

- 比赛背景

Can you help end gender bias in pronoun resolution?

Pronoun resolution is part of coreference resolution, the task of pairing an expression to its referring entity. This is an important task for natural language understanding, and the resolution of ambiguous pronouns is a longstanding challenge.

Unfortunately, recent studies have suggested gender bias among state-of-the-art coreference resolvers. Google AI Language aims to improve gender-fairness in modeling by releasing the Gendered Ambiguous Pronouns (GAP) dataset, containing gender-balanced pronouns (50% of its examples containing feminine pronouns, and 50% containing masculine pronouns).

In this two-stage competition, Kagglers are challenged to build pronoun resolution systems that perform equally well regardless of pronoun gender. Stage two's final evaluation will use a new dataset following the same format. To encourage gender-fair modeling, the ratio of masculine to feminine examples in the official test data will not be known ahead of time.

- 评价指标：

Submissions are evaluated using the multi-class logarithmic loss.

- 比赛数据

https://www.kaggle.com/c/gendered-pronoun-resolution/data

- 时间轴：Feb 6, 2019 – Apr 23, 2019
- 比赛结果

https://www.kaggle.com/c/gendered-pronoun-resolution/leaderboard

Data Competition in 2019

- 赛后分享

1st place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90392

3rd place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90424

4th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90484

5th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90668

7th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90334

8th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90344

9th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90421

11th place solution：https://www.kaggle.com/c/gendered-pronoun-resolution/discussion/90483

Ciphertext Challenge II

https://www.kaggle.com/c/ciphertext-challenge-ii

参赛队伍：74，参赛人数 79

比赛类型：Playground，比赛数据：文本

- 比赛背景

In our first ciphertext competition, we hunted the wilds of the '90s-era internet. This time around, we're exploring the dark slow-broadband-y wastelands of 2011, with the Movie Review Dataset. In 2011 most of the internet hadn't even been invented yet*, so wow, you're in for a treat.

Again, simple classic ciphers have been used to encrypt this dataset. Your mission this time: to correctly match each piece of ciphertext with its corresponding piece of plaintext. Daunting! Also, there are some new ciphers in play this time, which will involve some meta-puzzling. Enjoy!

Data Competition in 2019

Swag prizes go to the first three teams to crack all four ciphers OR to the top three teams on the LB (in case the ciphers are not all cracked). Additionally, swag prizes will be awarded to the best competition-related kernels, in both visualization and cryptanalysis, based on upvotes.

- 评价指标：

Submissions are evaluated on Accuracy between the predicted plaintext index and the actual index.

- 比赛数据

https://www.kaggle.com/c/ciphertext-challenge-ii/data

- 时间轴：Mar 28, 2019 – Apr 26, 2019
- 比赛结果

https://www.kaggle.com/c/ciphertext-challenge-ii/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/ciphertext-challenge-ii/discussion/87281

Don't Overfit! II

https://www.kaggle.com/c/dont-overfit-ii

参赛队伍：2330，参赛人数 2456

比赛类型：Playground，比赛数据：结构化

- 比赛背景

This is the next logical step in the evolution of weird competitions. Once again we have 20,000 rows of continuous variables, and a mere handful of training samples. Once again, we challenge you not to overfit. Do your best, model without overfitting, and add, perhaps, to your own legend.

In addition to bragging rights, the winner also gets swag. Enjoy!

- 评价指标：

Submissions are evaluated using AUCROC between the predicted target and the actual target value.

- 比赛数据

https://www.kaggle.com/c/dont-overfit-ii/data

- 时间轴：Feb 9, 2019 – May 8, 2019
- 比赛结果

https://www.kaggle.com/c/dont-overfit-ii/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/dont-overfit-ii/discussion/91766

2nd place solution：https://www.kaggle.com/c/dont-overfit-ii/discussion/91683

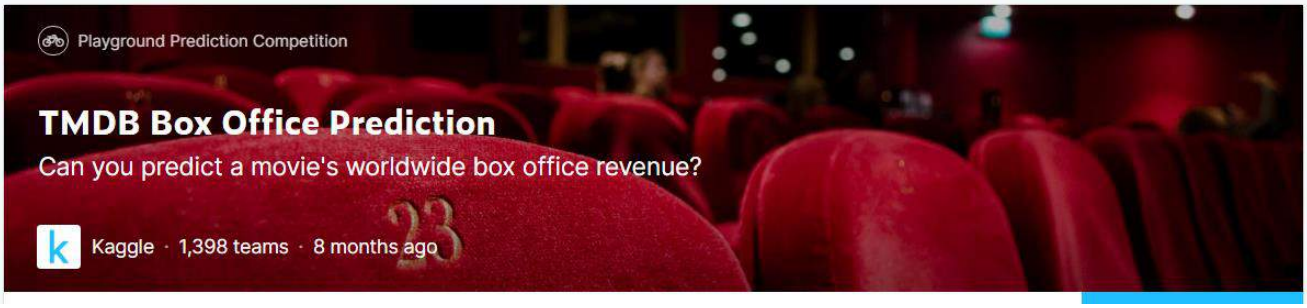
4th place solution：https://www.kaggle.com/c/dont-overfit-ii/discussion/91801

TMDB Box Office Prediction

https://www.kaggle.com/c/tmdb-box-office-prediction

参赛队伍：1398，参赛人数：1618

比赛类型：Playground，比赛数据：结构化



- 比赛背景

We're going to make you an offer you can't refuse: a Kaggle competition!

In a world… where movies made an estimated $41.7 billion in 2018, the film industry is more popular than ever. But what movies make the most money at the box office? How much does a director matter? Or the budget? For some movies, it's "You had me at 'Hello.'" For others, the trailer falls short of expectations and you think "What we have here is a failure to communicate."

Data Competition in 2019

In this competition, you're presented with metadata on over 7,000 past films from The Movie Database to try and predict their overall worldwide box office revenue. Data points provided include cast, crew, plot keywords, budget, posters, release dates, languages, production companies, and countries. You can collect other publicly available data to use in your model predictions, but in the spirit of this competition, use only data that would have been available before a movie's release.

Join in, "make our day", and then "you've got to ask yourself one question: 'Do I feel lucky?'"

- 评价指标：

Submissions are evaluated on Root-Mean-Squared-Logarithmic-Error (RMSLE) between the predicted value and the actual revenue. Logs are taken to not overweight blockbuster revenue movies.

- 比赛数据

https://www.kaggle.com/c/tmdb-box-office-prediction/data

- 时间轴：Feb 7, 2019 – May 31, 2019
- 比赛结果

https://www.kaggle.com/c/tmdb-box-office-prediction/leaderboard

- 赛后分享

Google Landmark Retrieval 2019

https://www.kaggle.com/c/landmark-retrieval-2019

参赛队伍：144

比赛类型：Research，比赛数据：图像

- 比赛背景

Data Competition in 2019

Image retrieval is a fundamental problem in computer vision: given a query image, can you find similar images in a large database? This is especially important for query images containing landmarks, which accounts for a large portion of what people like to photograph.

In this competition, Kagglers are given query images and, for each query, are expected to retrieve all database images containing the same landmarks (if any). The competition will proceed in two phases: The 1st phase will use the same test and index sets as last year, while for phase 2 we will release a completely new dataset that contains 700K images with more than 100K unique landmarks. We hope that this release will accelerate progress in this important research problem.

This challenge is organized in conjunction with the Landmark Recognition Challenge (https://www.kaggle.com/c/landmark-recognition-2019). In particular, note that the test set for both challenges is the same, to encourage participants to compete in both. We also encourage participants to use the training data from the recognition challenge (either from this year's or last year's dataset) to develop models which could be useful for the retrieval challenge.

- 评价指标：

Submissions are evaluated according to mean Average Precision @ 100:

$$mAP@100 = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{min(m_q, 100)}\sum_{k=1}^{min(n_q,100)}P_q(k)rel_q(k)$$

- 比赛数据

https://www.kaggle.com/c/landmark-retrieval-2019/data

- 时间轴：Apr 8, 2019 – Jun 4, 2019
- 比赛结果

https://www.kaggle.com/c/landmark-retrieval-2019/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/landmark-retrieval-2019/discussion/94735

2nd place solution：https://www.kaggle.com/c/landmark-retrieval-2019/discussion/95174

8th place solution：https://www.kaggle.com/c/landmark-retrieval-2019/discussion/94540

9th place solution：https://www.kaggle.com/c/landmark-retrieval-2019/discussion/94581

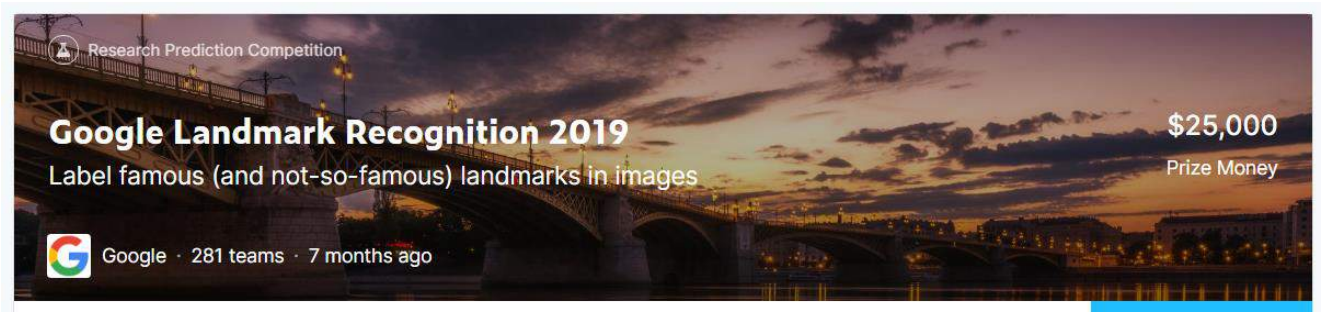
15th place solution：https://www.kaggle.com/c/landmark-retrieval-2019/discussion/94581

Google Landmark Recognition 2019

https://www.kaggle.com/c/landmark-recognition-2019

参赛队伍：281

比赛类型：Research，比赛数据：图像



- 比赛背景

Did you ever go through your vacation photos and ask yourself: What is the name of this temple I visited in China? Who created this monument I saw in France? Landmark recognition can help! This technology can predict landmark labels directly from image pixels, to help people better understand and organize their photo collections.

Today, a great obstacle to landmark recognition research is the lack of large annotated datasets. In this competition, we present the largest worldwide dataset to date, to foster progress in this problem. This competition challenges Kagglers to build models that recognize the correct landmark (if any) in a dataset of challenging test images.

Many Kagglers are familiar with image classification challenges like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which aims to recognize 1K general object categories. Landmark recognition is a little different from that: it contains a much larger number of classes (there are more than 200K classes in this challenge), and the number of training examples per class may not be very large. Landmark recognition is challenging in its own way.

This is the second edition of this challenge. Compared to the first edition, the new dataset is more comprehensive and diverse. See the Data tab for more in-depth discussion on the new released dataset.

This challenge is organized in conjunction with the Landmark Retrieval Challenge (https://www.kaggle.com/c/landmark-retrieval-2019). In particular, note that the test set for both challenges is the same, to encourage participants to compete in both. We encourage participants

to use the training data from the recognition challenge (either from this year's or last year's dataset) to develop models which could be useful for the retrieval challenge.

- 评价指标：

Submissions are evaluated using Global Average Precision (GAP).

- 比赛数据

https://www.kaggle.com/c/landmark-recognition-2019/data

- 时间轴：Apr 8, 2019 – Jun 4, 2019
- 比赛结果

https://www.kaggle.com/c/landmark-recognition-2019/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/landmark-recognition-2019/discussion/95176

8th place solution：https://www.kaggle.com/c/landmark-recognition-2019/discussion/94512

LANL Earthquake Prediction

https://www.kaggle.com/c/LANL-Earthquake-Prediction

参赛队伍：4521，参赛人数：5452

比赛类型：Research，比赛数据：结构化

- 比赛背景

Forecasting earthquakes is one of the most important problems in Earth science because of their devastating consequences. Current scientific studies related to earthquake forecasting focus on three key points: when the event will occur, where it will occur, and how large it will be.

In this competition, you will address when the earthquake will take place. Specifically, you'll predict the time remaining before laboratory earthquakes occur from real-time seismic data.

If this challenge is solved and the physics are ultimately shown to scale from the laboratory to the field, researchers will have the potential to improve earthquake hazard assessments that could save lives and billions of dollars in infrastructure.

This challenge is hosted by Los Alamos National Laboratory which enhances national security by ensuring the safety of the U.S. nuclear stockpile, developing technologies to reduce threats from weapons of mass destruction, and solving problems related to energy, environment, infrastructure, health, and global security concerns.

- 评价指标：

Submissions are evaluated using the mean absolute error between the predicted time remaining before the next lab earthquake and the act remaining time.

- 比赛数据

https://www.kaggle.com/c/LANL-Earthquake-Prediction/data

- 时间轴：Jan 11, 2019 – Jun 4, 2019
- 比赛结果

https://www.kaggle.com/c/LANL-Earthquake-Prediction/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94390

2nd place solution: https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94369

3rd place solution：https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94369

5th palce solution：https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94484

10th place solution：https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94466

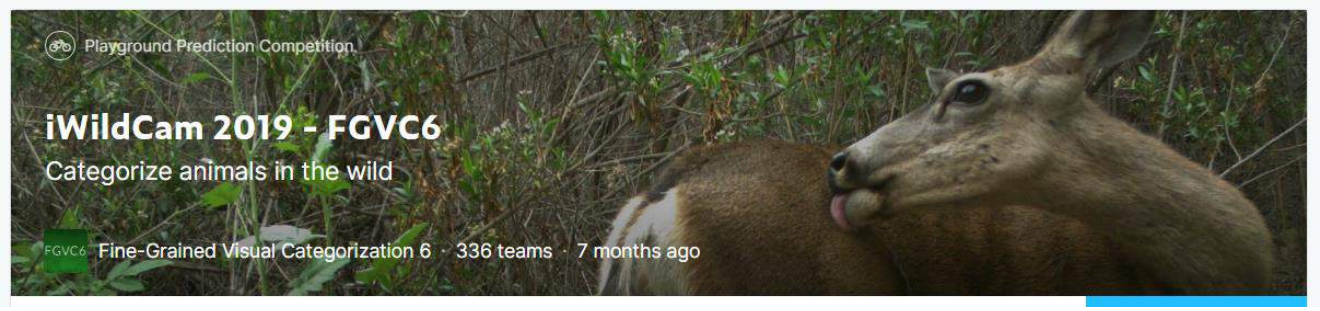
12th place solution：https://www.kaggle.com/c/LANL-Earthquake-Prediction/discussion/94500

iWildCam 2019 - FGVC6

https://www.kaggle.com/c/iwildcam-2019-fgvc6

参赛队伍：336，参赛人数：389

比赛类型：Research，比赛数据：图像



- 比赛背景

Camera Traps (or Wild Cams) enable the automatic collection of large quantities of image data. Biologists all over the world use camera traps to monitor biodiversity and population density of animal species. We have recently been making strides towards automating the species classification challenge in camera traps, but as we try to expand the scope of these models from specific regions where we have collected training data to nearby areas we are faced with an interesting probem: how do you classify a species in a new region that you may not have seen in previous training data?

In order to tackle this problem, we have prepared a challenge where the training data and test data are from different regions, namely The American Southwest and the American Northwest. The species seen in each region overlap, but are not identical, and the challenge is to classify the test species correctly. To this end, we will allow training on our American Southwest data (from CaltechCameraTraps), on iNaturalist 2017/2018 data, and on simulated data generated from Microsoft AirSim. We have provided a taxonomy file mapping our classes into the iNat taxonomy.

This is an FGVCx competition as part of the FGVC6 workshop at CVPR 2019, and is sponsored by Microsoft AI for Earth. There is a github page for the competition here. Please open an issue if you have questions or problems with the dataset.

- 评价指标：

Submissions will be evaluated based on their macro F1 score - i.e. F1 will be calculated for each class of animal (including "empty" if no animal is present), and the submission's final score will be the unweighted mean of all class F1 scores.

- 比赛数据

https://www.kaggle.com/c/iwildcam-2019-fgvc6/data

- 时间轴：Mar 26, 2019 – Jun 8, 2019
- 比赛结果

https://www.kaggle.com/c/iwildcam-2019-fgvc6/leaderboard

Data Competition in 2019

- 赛后分享

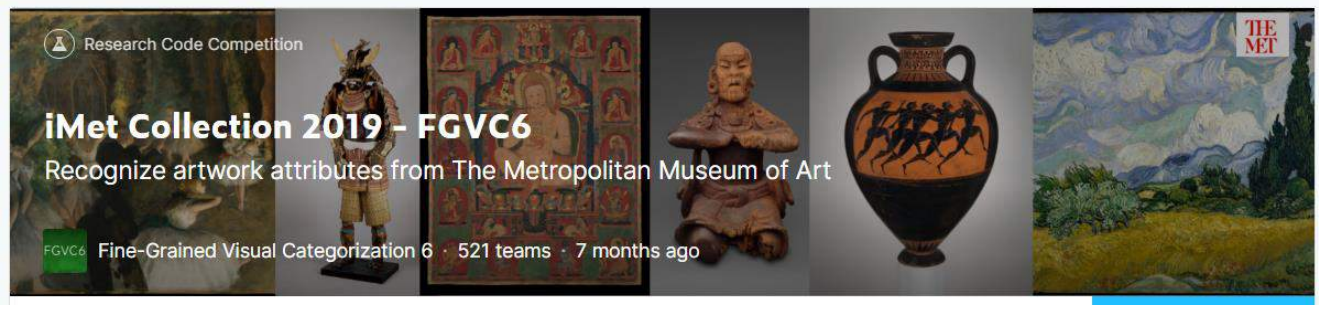
3rd place solution：https://www.kaggle.com/c/iwildcam-2019-fgvc6/discussion/95406

iMet Collection 2019 - FGVC6

https://www.kaggle.com/c/imet-2019-fgvc6

参赛队伍：521，参赛人数：765

比赛类型：Research，比赛数据：图像



- 比赛背景

The Metropolitan Museum of Art in New York, also known as The Met, has a diverse collection of over 1.5M objects of which over 200K have been digitized with imagery. The online cataloguing information is generated by Subject Matter Experts (SME) and includes a wide range of data. These include, but are not limited to: multiple object classifications, artist, title, period, date, medium, culture, size, provenance, geographic location, and other related museum objects within The Met's collection. While the SME-generated annotations describe the object from an art history perspective, they can also be indirect in describing finer-grained attributes from the museum-goer's understanding. Adding fine-grained attributes to aid in the visual understanding of the museum objects will enable the ability to search for visually related objects.

- 评价指标：

Submissions will be evaluated based on their mean F2 score.

- 比赛数据

https://www.kaggle.com/c/imet-2019-fgvc6/data

- 时间轴：Mar 26, 2019 – Jun 8, 2019
- 比赛结果

https://www.kaggle.com/c/imet-2019-fgvc6/leaderboard

Data Competition in 2019

- 赛后分享

1st place solution：https://www.kaggle.com/c/imet-2019-fgvc6/discussion/94687

2nd place solution：https://www.kaggle.com/c/imet-2019-fgvc6/discussion/96149

3rd place solution：https://www.kaggle.com/c/imet-2019-fgvc6/discussion/96424

4th place solution：https://www.kaggle.com/c/imet-2019-fgvc6/discussion/94817

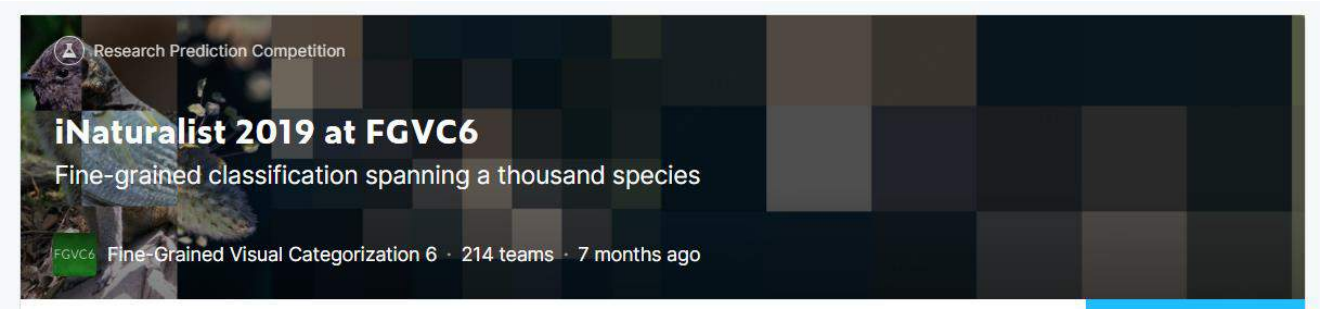
10th place solution：https://www.kaggle.com/c/imet-2019-fgvc6/discussion/95311

iNaturalist 2019 at FGVC6

https://www.kaggle.com/c/inaturalist-2019-fgvc6

参赛队伍：214，参赛人数：249

比赛类型：Research，比赛数据：图像



- 比赛背景

As part of the FGVC6 workshop at CVPR 2019 we are conducting the iNat Challenge 2019 large scale species classification competition, sponsored by Microsoft. It is estimated that the natural world contains several million species of plants and animals. Without expert knowledge, many of these species are extremely difficult to accurately classify due to their visual similarity. The goal of this competition is to push the state of the art in automatic image classification for real world data that features a large number of fine-grained categories.

Previous versions of the challenge have focused on classifying large numbers of species. This year features a smaller number of highly similar categories captured in a wide variety of situations, from all over the world. In total, the iNat Challenge 2019 dataset contains 1,010 species, with a combined training and validation set of 268,243 images that have been collected and verified by multiple users from iNaturalist.

Data Competition in 2019

Teams with top submissions, at the discretion of the workshop organizers, will be invited to present their work at the FGVC6 workshop. Participants who make a submission that beats the sample submission can fill out this form to receive $150 in Google Cloud credits.

- 评价指标：

We use top-1 classification error as the metric for this competition. For each image, an algorithm will produce 1 label. If the predicted label matches the ground truth label then the error for that image is 0, otherwise it is 1. The final score is the error averaged across all images.

- 比赛数据

https://www.kaggle.com/c/inaturalist-2019-fgvc6/data

- 时间轴：Mar 30, 2019 – Jun 11, 2019
- 比赛结果

https://www.kaggle.com/c/inaturalist-2019-fgvc6/leaderboard

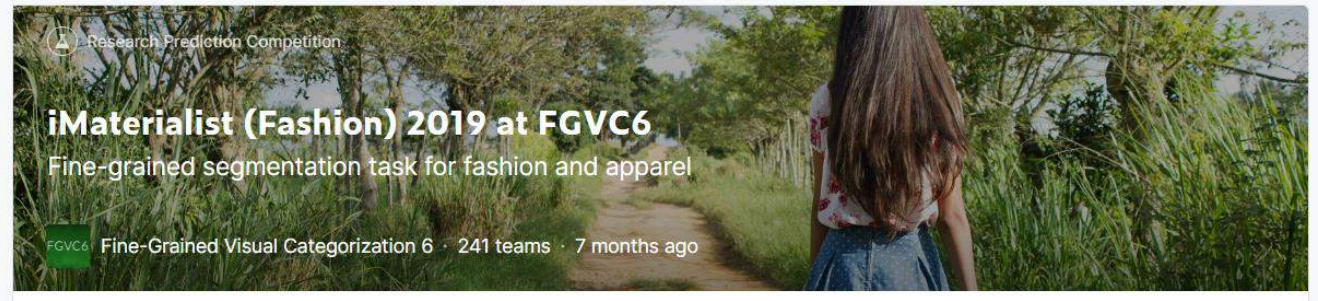
- 赛后分享

iMaterialist (Fashion) 2019 at FGVC6

https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6

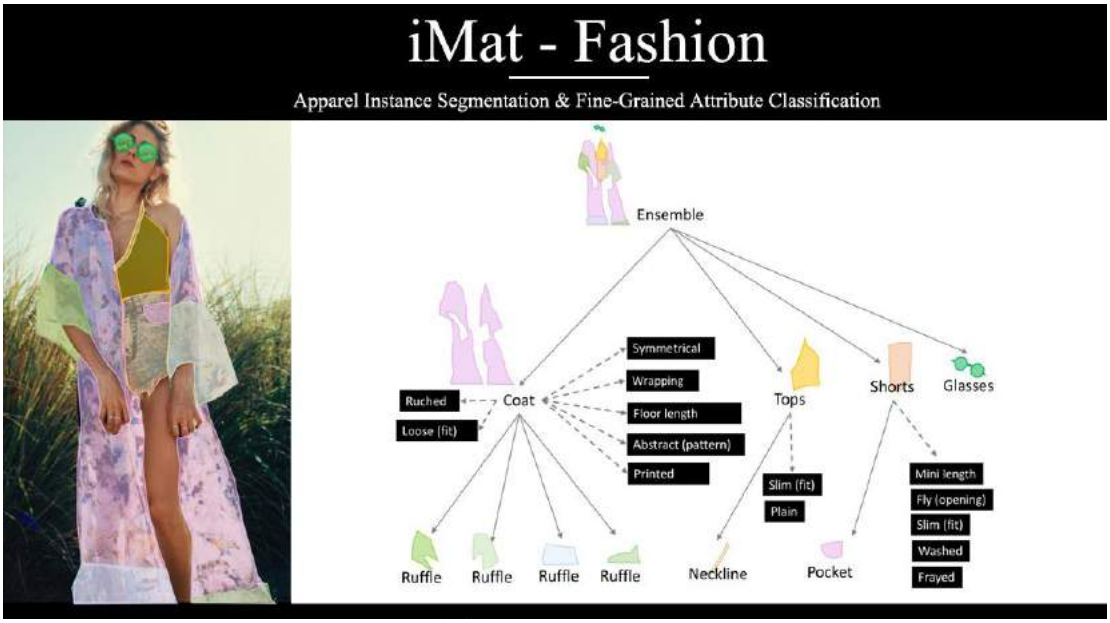
参赛队伍：241，参赛人数：298

比赛类型：Research，比赛数据：图像



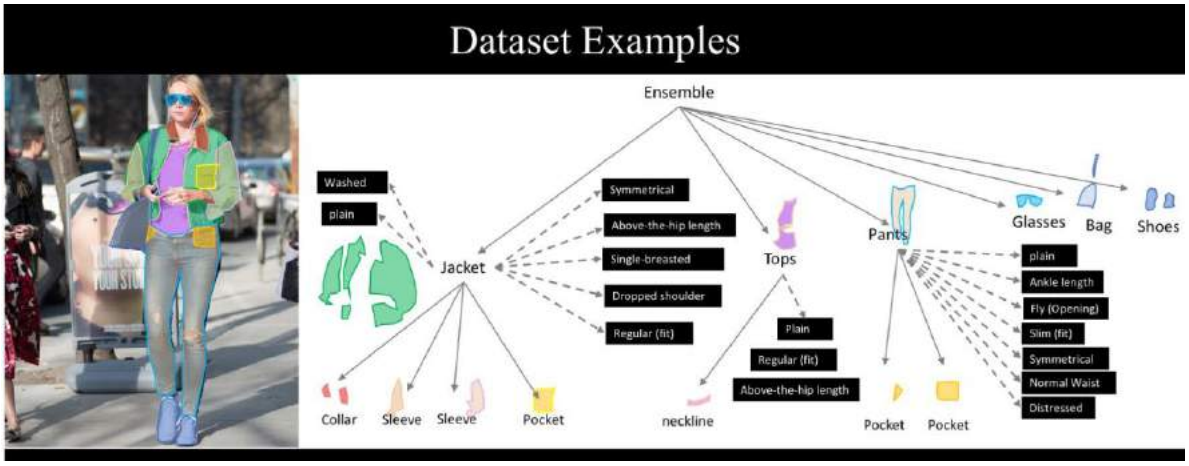
- 比赛背景

Designers know what they are creating, but what, and how, do people really wear their products? What combinations of products are people using? In this competition, we challenge you to develop algorithms that will help with an important step towards automatic product detection – to accurately assign segmentations and attribute labels for fashion images.



Visual analysis of clothing is a topic that has received increasing attention in recent years. Being able to recognize apparel products and associated attributes from pictures could enhance the shopping experience for consumers, and increase work efficiency for fashion professionals.

We present a new clothing dataset with the goal of introducing a novel fine-grained segmentation task by joining forces between the fashion and computer vision communities. The proposed task unifies both categorization and segmentation of rich and complete apparel attributes, an important step toward real-world applications.



While early work in computer vision addressed related clothing recognition tasks, these are not designed with fashion insiders' needs in mind, possibly due to the research gap in fashion design and computer vision. To address this, we first propose a fashion taxonomy built by fashion experts, informed by product description from the internet. To capture the complex structure of fashion objects and ambiguity in descriptions obtained from crawling the web, our standardized taxonomy contains 46 apparel objects (27 main apparel items and 19 apparel parts), and 92

related fine-grained attributes. Secondly, a total of 50K clothing images (10K with both segmentation and fine-grained attributes, 40K with apparel instance segmentation) in daily-life, celebrity events, and online shopping are labeled by both domain experts and crowd workers for fine-grained segmentation.

- 评价指标：

Submissions are evaluated on the mean average precision at different intersection over union (IoU) thresholds.

- 比赛数据

https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/data

- 时间轴：Apr 25, 2019 – Jun 11, 2019
- 比赛结果

https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/discussion/95233

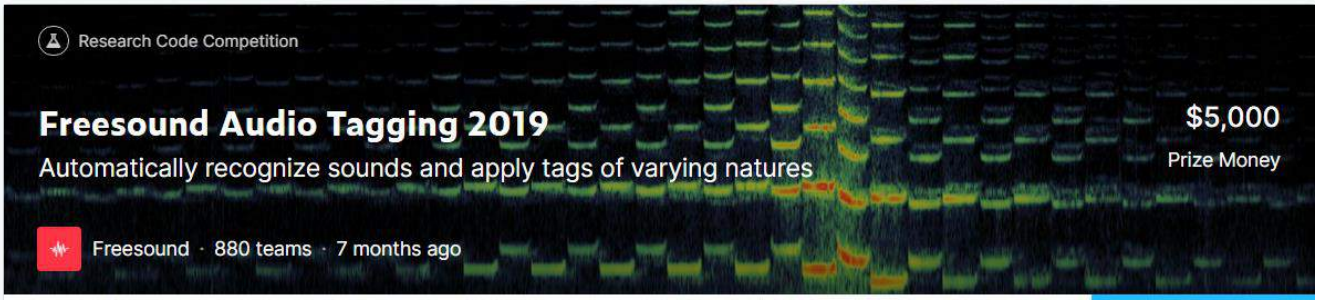
3rd place solution：https://www.kaggle.com/c/imaterialist-fashion-2019-FGVC6/discussion/95234

Freesound Audio Tagging 2019

https://www.kaggle.com/c/freesound-audio-tagging-2019

参赛队伍：880

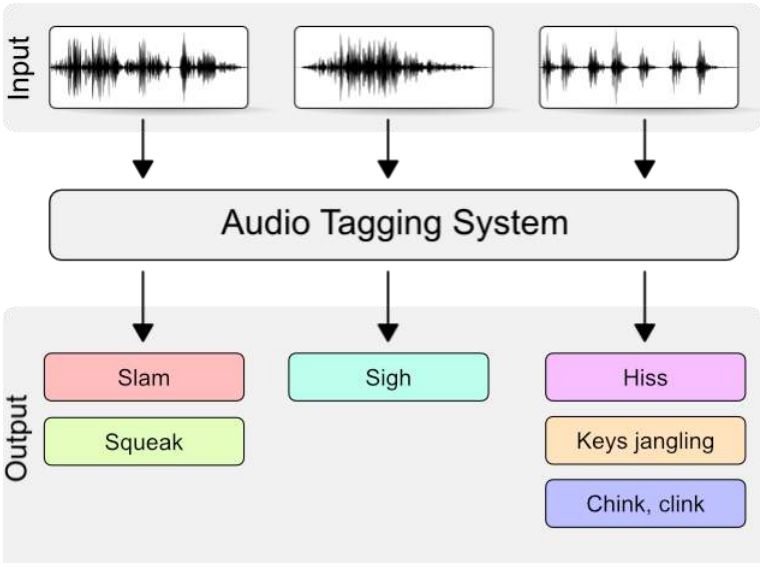
比赛类型：Research，比赛数据：音频



- 比赛背景

One year ago, Freesound and Google’s Machine Perception hosted an audio tagging competition challenging Kagglers to build a general-purpose auto tagging system. This year they’re back and taking the challenge to the next level with multi-label audio tagging, doubled number of audio

categories, and a noisier than ever training set. If you like raising your ML game, this challenge is for you.



Here's the background: Some sounds are distinct and instantly recognizable, like a baby's laugh or the strum of a guitar. Other sounds are difficult to pinpoint. If you close your eyes, could you tell the difference between the sound of a chainsaw and the sound of a blender?

Because of the vastness of sounds we experience, no reliable automatic general-purpose audio tagging systems exist. A significant amount of manual effort goes into tasks like annotating sound collections and providing captions for non-speech events in audiovisual content.

To tackle this problem, Freesound (an initiative by MTG-UPF that maintains a collaborative database with over 400,000 Creative Commons Licensed sounds) and Google Research's Machine Perception Team (creators of AudioSet, a large-scale dataset of manually annotated audio events with over 500 classes) have teamed up to develop the dataset for this new competition.

To win this competition, Kagglers will develop an algorithm to tag audio data automatically using a diverse vocabulary of 80 categories.

If successful, your systems could be used for several applications, ranging from automatic labelling of sound collections to the development of systems that automatically tag video content or recognize sound events happening in real time.

- 评价指标：

The task consists of predicting the audio labels (tags) for every test clip. Some test clips bear one label while others bear several labels. The predictions are to be done at the clip level, i.e., no start/end timestamps for the sound events are required.

The primary competition metric will be label-weighted label-ranking average precision (lwlrap, pronounced "Lol wrap"). This measures the average precision of retrieving a ranked list of relevant labels for each test clip (i.e., the system ranks all the available labels, then the precisions of the ranked lists down to each true label are averaged). This is a generalization of the mean reciprocal rank measure (used in last year's edition of the competition) for the case where there can be multiple true labels per test item. The novel "label-weighted" part means that the overall score is the average over all the labels in the test set, where each label receives equal weight (by contrast, plain lrap gives each test item equal weight, thereby discounting the contribution of individual labels when they appear on the same item as multiple other labels).

We use label weighting because it allows per-class values to be calculated, and still have the overall metric be expressed as simple average of the per-class metrics (weighted by each label's prior in the test set). For participant's convenience, a Python implementation of lwlrap is provided in this public Google Colab.

- 比赛数据

https://www.kaggle.com/c/freesound-audio-tagging-2019/data

- 时间轴：Apr 4, 2019 – Jun 18, 2019
- 比赛结果

https://www.kaggle.com/c/freesound-audio-tagging-2019/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/95924

2nd place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97815

3rd place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97926

6th place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/96680

7th place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97812

9th place solution：https://www.kaggle.com/c/freesound-audio-tagging-2019/discussion/97736

Instant Gratification (Kernels Only)

https://www.kaggle.com/c/instant-gratification

参赛队伍：1832，参赛人数：2035

Data Competition in 2019

比赛类型：Featured，比赛数据：结构化

- 比赛背景

In 2015, Kaggle introduced Kernels as a resource to competition participants. It was a controversial decision to add a code-sharing tool to a competitive coding space. We thought it was important to make Kaggle more than a place where competitions are solved behind closed digital doors. Since then, Kernels has grown from its infancy--essentially a blinking cursor in a docker container--into its teenage years. We now have more compute, longer runtimes, better datasets, GPUs, and an improved interface.

We have iterated and tested several Kernels-only (KO) competition formats with a true holdout test set, in particular deploying them when we would have otherwise substituted a two-stage competition. However, the experience of submitting to a Kernels-only competition has typically been asynchronous and imperfect; participants wait many days after a competition has concluded for their selected Kernels to be rerun on the holdout test dataset, the leaderboard updated, and the winners announced. This flow causes heartbreak to participants whose Kernels fail on the unseen test set, leaving them with no way to correct tiny errors that spoil months of hard work.

Say Hello to Synchronous KO

We're now pleased to announce general support for a synchronous Kernels-only format. When you submit from a Kernel, Kaggle will run the code against both the public test set and private test set in real time. This small-but-substantial tweak improves the experience for participants, the host, and Kaggle:

With a truly withheld test set, we are practicing proper, rigorous machine learning.

We will be able to offer more varieties of competitions and intend to run many fewer confusing two-stage competitions.

You will be able to see if your code runs successfully on the withheld test set and have the leeway to intervene if it fails.

We will run all submissions against the private data, not just selected ones. Participants will get the complete and familiar public/private scores available in a traditional competition.

The final leaderboard can be released at the end of the competition, without the delay of rerunning Kernels.

This competition is a low-stakes, trial-run introduction to our new synchronous KO implementation. We want to test that the process goes smoothly and gather feedback on your experiences. While it may feel like a normal KO competition, there are complicated new mechanics in play, such as the selection logic of Kernels that are still running when the deadline passes.

Since the competition also presents an authentic machine learning problem, it will also award Kaggle medals and points. Have fun, good luck, and welcome to the world of synchronous Kernels competitions!

- 评价指标：

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/instant-gratification/data

- 时间轴：May 18, 2019 – Jun 21, 2019
- 比赛结果

https://www.kaggle.com/c/instant-gratification/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/instant-gratification/discussion/96549

6th place solution：https://www.kaggle.com/c/instant-gratification/discussion/96496

Aerial Cactus Identification (Kernels Only)

https://www.kaggle.com/c/aerial-cactus-identification

参赛队伍：1225，参赛人数：1316

比赛类型：Playground，比赛数据：图像

- 比赛背景

To assess the impact of climate change on Earth's flora and fauna, it is vital to quantify how human activities such as logging, mining, and agriculture are impacting our protected natural areas. Researchers in Mexico have created the VIGIA project, which aims to build a system for autonomous surveillance of protected areas. A first step in such an effort is the ability to recognize the vegetation inside the protected areas. In this competition, you are tasked with creation of an algorithm that can identify a specific type of cactus in aerial imagery.

- 评价指标：

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/aerial-cactus-identification/data

- 时间轴：Mar 9, 2019 – July 9, 2019

- 比赛结果

https://www.kaggle.com/c/aerial-cactus-identification/leaderboard

- 赛后分享

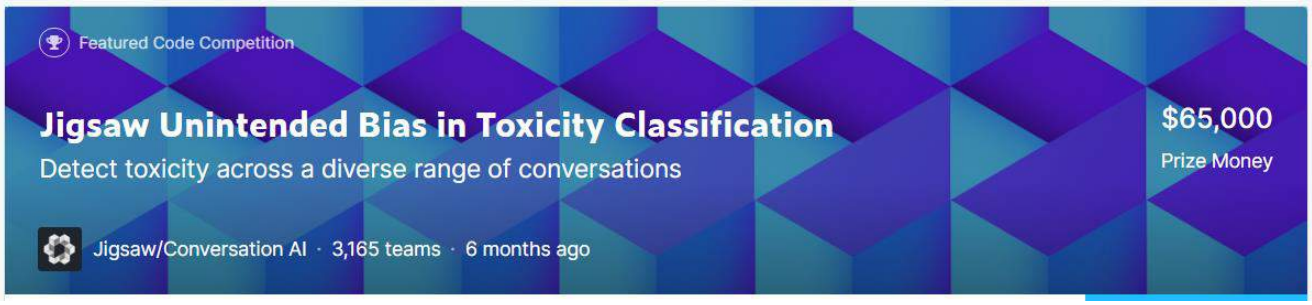
Jigsaw Unintended Bias in Toxicity Classification (Kernels Only)

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

参赛队伍：3165，参赛人数：4397

比赛类型：Featured，比赛数据：文本

Data Competition in 2019



- 比赛背景

Can you help detect toxic comments — and minimize unintended model bias? That's your challenge in this competition.

The Conversation AI team, a research initiative founded by Jigsaw and Google (both part of Alphabet), builds technology to protect voices in conversation. A main area of focus is machine learning models that can identify toxicity in online conversations, where toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion.

Last year, in the Toxic Comment Classification Challenge, you built multi-headed models to recognize toxicity and several subtypes of toxicity. This year's competition is a related challenge: building toxicity models that operate fairly across a diverse range of conversations.

Here's the background: When the Conversation AI team first built toxicity models, they found that the models incorrectly learned to associate the names of frequently attacked identities with toxicity. Models predicted a high likelihood of toxicity for comments containing those identities (e.g. "gay"), even when those comments were not actually toxic (such as "I am a gay woman"). This happens because training data was pulled from available sources where unfortunately, certain identities are overwhelmingly referred to in offensive ways. Training a model from data with these imbalances risks simply mirroring those biases back to users.

In this competition, you're challenged to build a model that recognizes toxicity and minimizes this type of unintended bias with respect to mentions of identities. You'll be using a dataset labeled for identity mentions and optimizing a metric designed to measure unintended bias. Develop strategies to reduce unintended bias in machine learning models, and you'll help the Conversation AI team, and the entire industry, build models that work well for a wide range of conversations.

- 评价指标：

This competition will use a newly developed metric that combines several submetrics to balance overall performance with various aspects of unintended bias.

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

Data Competition in 2019

- 比赛数据

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

- 时间轴：Mar 30, 2019 – July 19, 2019
- 比赛结果

https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/101263

2nd place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/100661

3rd place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/97471

4th place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/101927

9th place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/100530

10th place solution：https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/101630

Two Sigma: Using News to Predict Stock Movements (Kernels Only)

https://www.kaggle.com/c/two-sigma-financial-news/overview

参赛队伍：2927

比赛类型：Featured，比赛数据：结构化

- 比赛背景

Can we use the content of news analytics to predict stock price performance? The ubiquity of data today enables investors at any scale to make better investment decisions. The challenge is ingesting and interpreting the data to determine which data is useful, finding the signal in this sea of information. Two Sigma is passionate about this challenge and is excited to share it with the Kaggle community.

Data Competition in 2019

As a scientifically driven investment manager, Two Sigma has been applying technology and data science to financial forecasts for over 17 years. Their pioneering advances in big data, AI, and machine learning have pushed the investment industry forward. Now, they're eager to engage with Kagglers in this continuing pursuit of innovation.

By analyzing news data to predict stock prices, Kagglers have a unique opportunity to advance the state of research in understanding the predictive power of the news. This power, if harnessed, could help predict financial outcomes and generate significant economic impact all over the world.

- 评价指标：

https://www.kaggle.com/c/two-sigma-financial-news/overview/evaluation

- 比赛数据

https://www.kaggle.com/c/two-sigma-financial-news/data

- 时间轴：Sep 28, 2018 – Aug 6, 2019
- 比赛结果

https://www.kaggle.com/c/two-sigma-financial-news/leaderboard

- 赛后分享

Northeastern SMILE Lab - Recognizing Faces in the Wild

https://www.kaggle.com/c/recognizing-faces-in-the-wild/overview/evaluation

参赛队伍：528，参赛人数：583

比赛类型：Playground，比赛数据：图像

- 比赛背景

In this competition, you will predict if two people share a kinship relationship or not based on their facial images. The data is provided by Families In the Wild (FIW), the largest and most comprehensive image database for automatic kinship recognition.

FIW's dataset is obtained from publicly available images from celebrities. For more information about their labeling process, please visit their database page.

- 评价指标：

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target. Not all pairs will be scored.

Data Competition in 2019

- 比赛数据

https://www.kaggle.com/c/recognizing-faces-in-the-wild/data

- 时间轴：May 14, 2019 – Aug 8, 2019
- 比赛结果

https://www.kaggle.com/c/recognizing-faces-in-the-wild/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/recognizing-faces-in-the-wild/discussion/103670

4th place solution：https://www.kaggle.com/c/recognizing-faces-in-the-wild/discussion/104288

5th place solution：https://www.kaggle.com/c/recognizing-faces-in-the-wild/discussion/103543

6th place solution：https://www.kaggle.com/c/recognizing-faces-in-the-wild/discussion/103457

Generative Dog Images (Kernels Only)

https://www.kaggle.com/c/generative-dog-images

参赛队伍：927

比赛类型：Research，比赛数据：图像

- 比赛背景

A generative adversarial network (GAN) is a class of machine learning system invented by Ian Goodfellow in 2014. Two neural networks compete with each other in a game. Given a training set, this technique learns to generate new data with the same statistics as the training set.

In this competition, you'll be training generative models to create images of dogs. Only this time… there's no ground truth data for you to predict. Here, you'll submit the images and be scored based on how well those images are classified as dogs from pre-trained neural networks. Take these images, for example. Can you tell which are real vs. generated?

Trick question; they are all generated!

Why dogs? We chose dogs because, well, who doesn't love looking at photos of adorable pups? Moreover, dogs can be classified into many sub-categories (breed, color, size), making them ideal candidates for image generation.

Generative methods (in particular, GANs) are currently used in various places on Kaggle for data augmentation. Their potential is vast; they can learn to mimic any distribution of data across any domain: photographs, drawings, music, and prose. If successful, not only will you help advance the state of the art in generative image creation, but you'll enable us to create more experiments across a variety of domains in the future.

- 评价指标：

Submissions are evaluated on MiFID (Memorization-informed Fréchet Inception Distance), which is a modification from Fréchet Inception Distance (FID).

- 比赛数据

https://www.kaggle.com/c/generative-dog-images/data

- 时间轴：Jun 29, 2019 – Aug 29, 2019
- 比赛结果

https://www.kaggle.com/c/generative-dog-images/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/generative-dog-images/discussion/106324

3rd place solution：https://www.kaggle.com/c/generative-dog-images/discussion/106514

Predicting Molecular Properties

https://www.kaggle.com/c/champs-scalar-coupling

参赛队伍：2749，参赛人数：3308

比赛类型：Featured，比赛数据：图像

- 比赛背景

Think you can use your data science smarts to make big predictions at a molecular level?

This challenge aims to predict interactions between atoms. Imaging technologies like MRI enable us to see and understand the molecular composition of tissues. Nuclear Magnetic Resonance (NMR) is a closely related technology which uses the same principles to understand the structure and dynamics of proteins and molecules.

Researchers around the world conduct NMR experiments to further understanding of the structure and dynamics of molecules, across areas like environmental science, pharmaceutical science, and materials science.

This competition is hosted by members of the CHemistry and Mathematics in Phase Space (CHAMPS) at the University of Bristol, Cardiff University, Imperial College and the University of Leeds. Winning teams will have an opportunity to partner with this multi-university research program on an academic publication

In this competition, you will develop an algorithm that can predict the magnetic interaction between two atoms in a molecule (i.e., the scalar coupling constant).

Once the competition finishes, CHAMPS would like to invite the top teams to present their work, discuss the details of their models, and work with them to write a joint research publication which discusses an open-source implementation of the solution.

- 评价指标：

Submissions are evaluated on the Log of the Mean Absolute Error.

- 比赛数据

https://www.kaggle.com/c/champs-scalar-coupling/data

- 时间轴：May 30, 2019 – Aug 29, 2019
- 比赛结果

https://www.kaggle.com/c/champs-scalar-coupling/leaderboard

Data Competition in 2019

- 赛后分享

1st place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106575

3rd place solution： https://www.kaggle.com/c/champs-scalar-coupling/discussion/106572

8th place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106347

9th place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106649

10th place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106271

12th place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106275

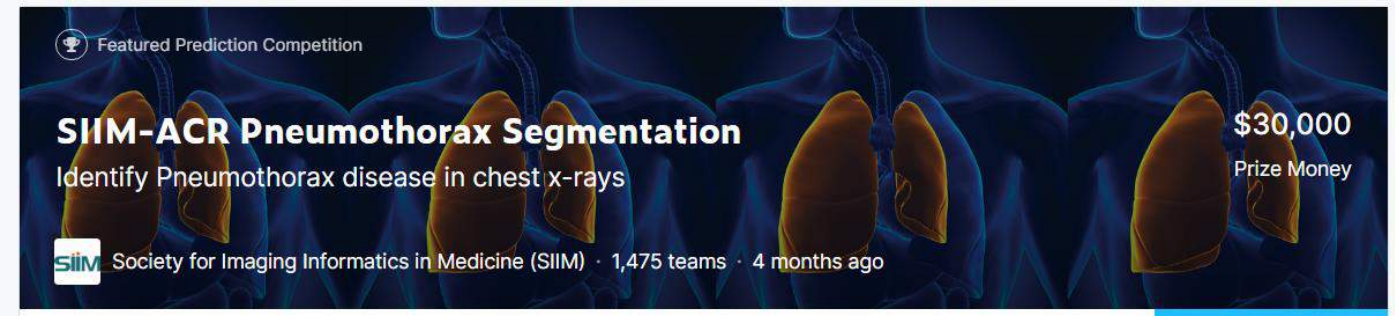
13th place solution：https://www.kaggle.com/c/champs-scalar-coupling/discussion/106377

SIIM-ACR Pneumothorax Segmentation

https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation

参赛队伍：1475，参赛人数：1983

比赛类型：Featured，比赛数据：图像



- 比赛背景

Imagine suddenly gasping for air, helplessly breathless for no apparent reason. Could it be a collapsed lung? In the future, your entry in this competition could predict the answer.

Pneumothorax can be caused by a blunt chest injury, damage from underlying lung disease, or most horrifying—it may occur for no obvious reason at all. On some occasions, a collapsed lung can be a life-threatening event.

Pneumothorax is usually diagnosed by a radiologist on a chest x-ray, and can sometimes be very difficult to confirm. An accurate AI algorithm to detect pneumothorax would be useful in a lot of clinical scenarios. AI could be used to triage chest radiographs for priority interpretation, or to provide a more confident diagnosis for non-radiologists.

Data Competition in 2019

The Society for Imaging Informatics in Medicine (SIIM) is the leading healthcare organization for those interested in the current and future use of informatics in medical imaging. Their mission is to advance medical imaging informatics across the enterprise through education, research, and innovation in a multi-disciplinary community. Today, they need your help.

In this competition, you'll develop a model to classify (and if present, segment) pneumothorax from a set of chest radiographic images. If successful, you could aid in the early recognition of pneumothoraces and save lives.

- 评价指标：

This competition is evaluated on the mean Dice coefficient.

- 比赛数据

https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/data

- 时间轴：Jun 25, 2019 – Sep 5, 2019
- 比赛结果

https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/107824

2nd place solution：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/108009

3rd place solution：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/107981

4th place solution：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/108397

5th place solution：https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/discussion/107603

Ciphertext Challenge III

https://www.kaggle.com/c/ciphertext-challenge-iii

参赛队伍：103，参赛人数：110

比赛类型：Playground，比赛数据：文本

- 比赛背景

In this new decryption competition's dataset, we've gone from perfectly respectable sources of electronic horror to a time before computers—heck, before calculus was called "calculus"! Shakespeare's plays are encrypted, and we time travelers must un-encrypt them so people can do innovative stage productions with intricate makeup, costumes, and possibly—possibly!—Leonardo DiCaprio. Think about it, folks: Leo.*

As in previous ciphertext challenges, simple classic ciphers have been used to encrypt this dataset, along with a slightly less simple surprise that expands our definition of "classic" into the modern age. The mission is the same: to correctly match each piece of ciphertext with its corresponding piece of plaintext. Daunting! Meta-puzzles and difficulty await!

Swag prizes go to the first three teams to crack all four ciphers OR to the top three teams on the leaderboard (in case the ciphers are not all cracked). Additionally, swag prizes will be awarded to the best competition-related kernels, in both visualization and cryptanalysis, based on upvotes. Last, the coveted "Phil Prize"—for the team that correctly deduces the form AND key of the final cipher—is up for grabs again.

- 评价指标：

Submissions are evaluated on Accuracy between the predicted plaintext index and the actual index.

- 比赛数据

https://www.kaggle.com/c/ciphertext-challenge-iii/data

- 时间轴：Aug 9, 2019 – Sep 6, 2019
- 比赛结果

https://www.kaggle.com/c/ciphertext-challenge-iii/leaderboard

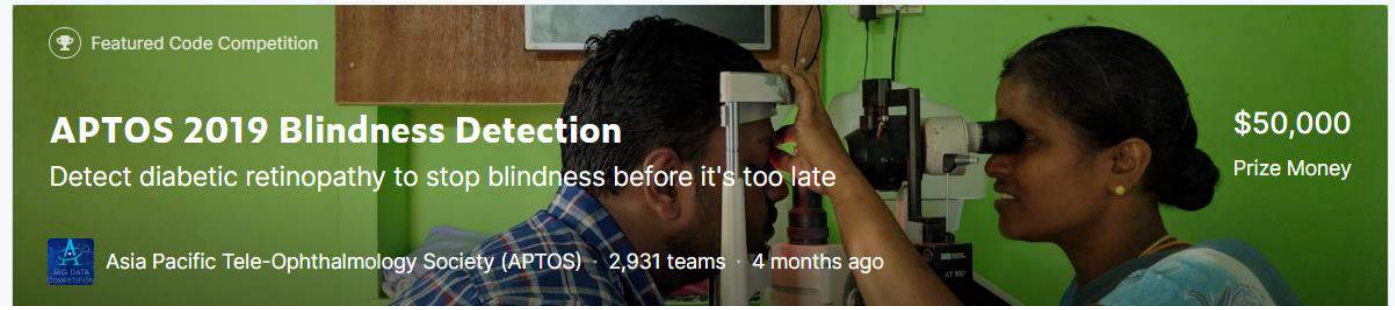
- 赛后分享

APTOS 2019 Blindness Detection (Kernels Only)

https://www.kaggle.com/c/aptos2019-blindness-detection

参赛队伍：2931，参赛人数：3509

比赛类型：Featured，比赛数据：图像



- 比赛背景

Millions of people suffer from diabetic retinopathy, the leading cause of blindness among working aged adults. Aravind Eye Hospital in India hopes to detect and prevent this disease among people living in rural areas where medical screening is difficult to conduct. Successful entries in this competition will improve the hospital's ability to identify potential patients. Further, the solutions will be spread to other Ophthalmologists through the 4th Asia Pacific Tele-Ophthalmology Society (APTOS) Symposium

Currently, Aravind technicians travel to these rural areas to capture images and then rely on highly trained doctors to review the images and provide diagnosis. Their goal is to scale their efforts through technology; to gain the ability to automatically screen images for disease and provide information on how severe the condition may be.

In this synchronous Kernels-only competition, you'll build a machine learning model to speed up disease detection. You'll work with thousands of images collected in rural areas to help identify diabetic retinopathy automatically. If successful, you will not only help to prevent lifelong blindness, but these models may be used to detect other sorts of diseases in the future, like glaucoma and macular degeneration.

- 评价指标：

Submissions are scored based on the quadratic weighted kappa, which measures the agreement between two ratings.

- 比赛数据

https://www.kaggle.com/c/aptos2019-blindness-detection/data

Data Competition in 2019

- 时间轴：Jun 28, 2019 – Sep 8, 2019
- 比赛结果

https://www.kaggle.com/c/aptos2019-blindness-detection/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/108065

4th place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/107926

7th place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/108058

8th place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/108030

11th place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/107958

13th place solution：https://www.kaggle.com/c/aptos2019-blindness-detection/discussion/107944

Recursion Cellular Image Classification

https://www.kaggle.com/c/recursion-cellular-image-classification

参赛队伍：866，参赛人数：1067

比赛类型：Featured，比赛数据：图像

- 比赛背景

The cost of some drugs and medical treatments has risen so high in recent years that many patients are having to go without. You can help with a classification project that could make researchers more efficient.

One of the more surprising reasons behind the cost is how long it takes to bring new treatments to market. Despite improvements in technology and science, research and development continues to lag. In fact, finding new treatments takes, on average, more than 10 years and costs hundreds of millions of dollars.

Recursion Pharmaceuticals, creators of the industry's largest dataset of biological images, generated entirely in-house, believes AI has the potential to dramatically improve and expedite the drug discovery process. More specifically, your efforts could help them understand how drugs interact with human cells.

This competition will have you disentangling experimental noise from real biological signals. Your entry will classify images of cells under one of 1,108 different genetic perturbations. You can help eliminate the noise introduced by technical execution and environmental variation between experiments.

If successful, you could dramatically improve the industry's ability to model cellular images according to their relevant biology. In turn, applying AI could greatly decrease the cost of treatments, and ensure these treatments get to patients faster.

- 评价指标：

Submissions are evaluated on Multiclass Accuracy, which is simply the average number of observations with the correct label.

- 比赛数据

https://www.kaggle.com/c/recursion-cellular-image-classification/data

- 时间轴：Jun 28, 2019 – Sep 27, 2019
- 比赛结果

https://www.kaggle.com/c/recursion-cellular-image-classification/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110543

3rd place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110340

4th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110337

5th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110362

6th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110391

7th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110335

8th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110434

9th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110366

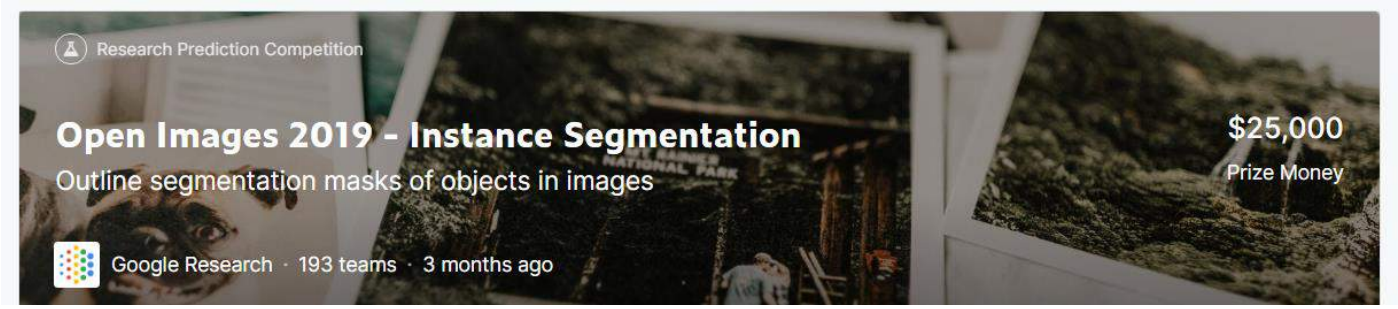
10th place solution：https://www.kaggle.com/c/recursion-cellular-image-classification/discussion/110375

Open Images 2019 - Instance Segmentation

https://www.kaggle.com/c/open-images-2019-instance-segmentation

参赛队伍：193，参赛人数：232

比赛类型：Research，比赛数据：图像



- 比赛背景

Computer vision has advanced considerably but is still challenged in matching the precision of human perception.

Open Images is a collaborative release of ~9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. This uniquely large and diverse dataset is designed to spur state of the art advances in analyzing and understanding images.

Google AI hopes that having a single dataset with unified annotations for image classification, object detection, visual relationship detection, and instance segmentation will stimulate progress towards genuine scene understanding.

In this track of the Challenge, you are asked to provide segmentation masks of objects.This track's training set represents 2.1M segmentation masks for object instances in 300 categories; with a validation set containing an additional 23k masks. The train set masks were produced by our state-of-the-art interactive segmentation process, where professional human annotators iteratively correct the output of a segmentation neural network. The validation and test set masks have been annotated manually with a strong focus on quality.

Data Competition in 2019

- 评价指标：

Submissions are evaluated by computing mean Average Precision, with the mean taken over the 300 segmentable classes of the challenge.

- 比赛数据

https://www.kaggle.com/c/open-images-2019-instance-segmentation/data

- 时间轴：Jun 12, 2019 – Oct 2, 2019
- 比赛结果

https://www.kaggle.com/c/open-images-2019-instance-segmentation/leaderboard

- 赛后分享

7th place solution：https://www.kaggle.com/c/open-images-2019-instance-segmentation/discussion/110983

11 place solution：https://www.kaggle.com/c/open-images-2019-instance-segmentation/discussion/111351

Open Images 2019 - Visual Relationship

https://www.kaggle.com/c/open-images-2019-visual-relationship

参赛队伍：201，参赛人数：247

比赛类型：Research，比赛数据：图像

Data Competition in 2019

- 比赛背景

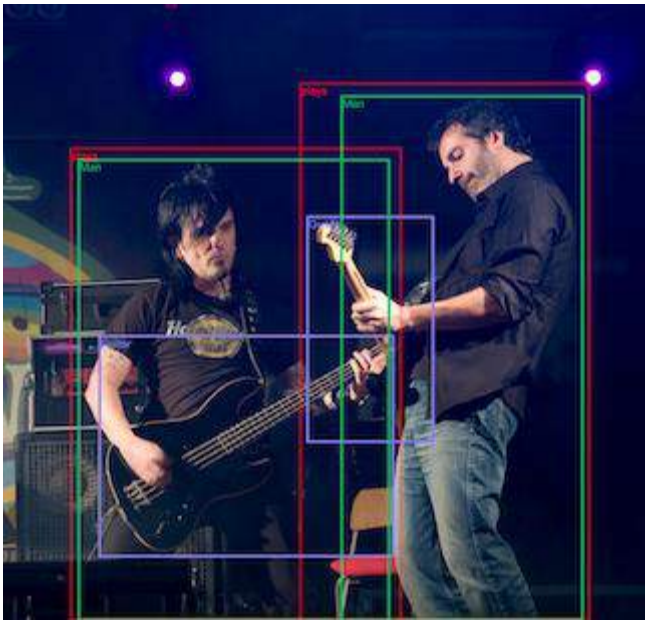
Computer vision has advanced considerably but is still challenged in matching the precision of human perception.

Open Images is a collaborative release of ~9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. This uniquely large and diverse dataset is designed to spur state of the art advances in analyzing and understanding images.

Google AI hopes that having a single dataset with unified annotations for image classification, object detection, visual relationship detection, and instance segmentation will stimulate progress towards genuine scene understanding.

In this track of the Challenge, you are asked to detect pairs of objects and the relationships that connect them.

The training set contains 329 relationship triplets with 375k training samples. These include both human-object relationships (e.g. "woman playing guitar", "man holding microphone"), object-object relationships (e.g. "beer on table", "dog inside car"), and also considers object-attribute relationships (e.g."handbag is made of leather" and "bench is wooden").



- 评价指标：

Submissions are evaluated by computing the weighted mean of three metrics: mean Average Precision (mAP) on relationships detection, Recall@N (where N=50), mean Average Precision on phrase detection (mean in mean Average Precision is taken over per-relationship APs).

- 比赛数据

Data Competition in 2019

https://www.kaggle.com/c/open-images-2019-visual-relationship/data

- 时间轴：Jun 4, 2019 – Oct 2, 2019
- 比赛结果

https://www.kaggle.com/c/open-images-2019-visual-relationship/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/open-images-2019-visual-relationship/discussion/111361

5th place solution：https://www.kaggle.com/c/open-images-2019-visual-relationship/discussion/111219

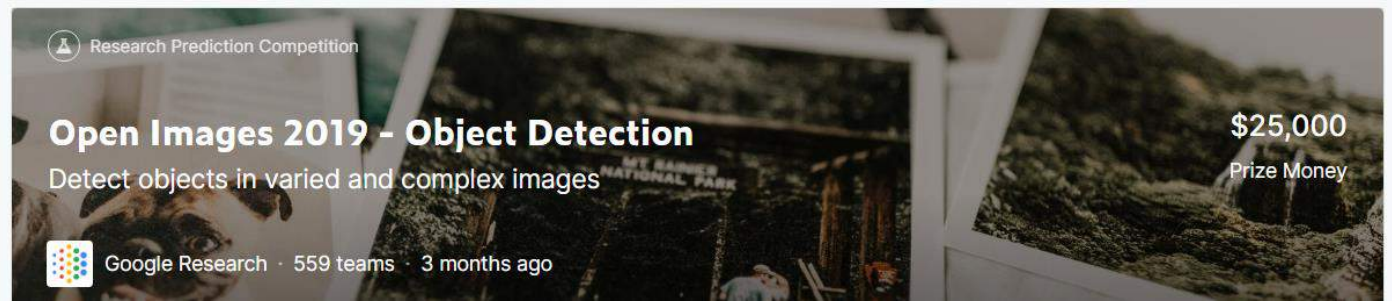
13th place solution：https://www.kaggle.com/c/open-images-2019-visual-relationship/discussion/110935

Open Images 2019 - Object Detection

https://www.kaggle.com/c/open-images-2019-object-detection

参赛队伍：559，参赛人数：698

比赛类型：Research，比赛数据：图像



- 比赛背景

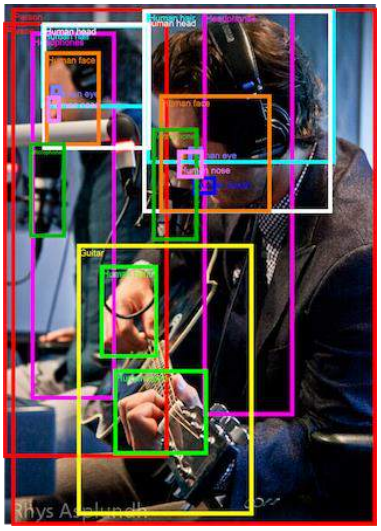
Computer vision has advanced considerably but is still challenged in matching the precision of human perception.

Open Images is a collaborative release of ~9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships. This uniquely large and diverse dataset is designed to spur state of the art advances in analyzing and understanding images.

Google AI hopes that having a single dataset with unified annotations for image classification, object detection, visual relationship detection, and instance segmentation will stimulate progress towards genuine scene understanding.

Data Competition in 2019

In this track of the Challenge, you are asked to predict a tight bounding box around object instances. The training set contains 12.2M bounding-boxes across 500 categories on 1.7M images. The boxes have been largely manually drawn by professional annotators to ensure accuracy and consistency. The images are very diverse and often contain complex scenes with several objects (7 per image on average).



- 评价指标：

Submissions are evaluated by computing mean Average Precision (mAP), modified to take into account the annotation process of Open Images dataset (mean is taken over per-class APs). The metric is described on the Open Images Challenge website.

- 比赛数据

https://www.kaggle.com/c/open-images-2019-object-detection/data

- 时间轴：Jun 4, 2019 – Oct 2, 2019
- 比赛结果

https://www.kaggle.com/c/open-images-2019-object-detection/leaderboard

- 赛后分享

3rd place solution： https://www.kaggle.com/c/open-images-2019-object-detection/discussion/1214141

6th place solution：https://www.kaggle.com/c/open-images-2019-object-detection/discussion/110953

10th place solution：https://www.kaggle.com/c/open-images-2019-object-detection/discussion/111266

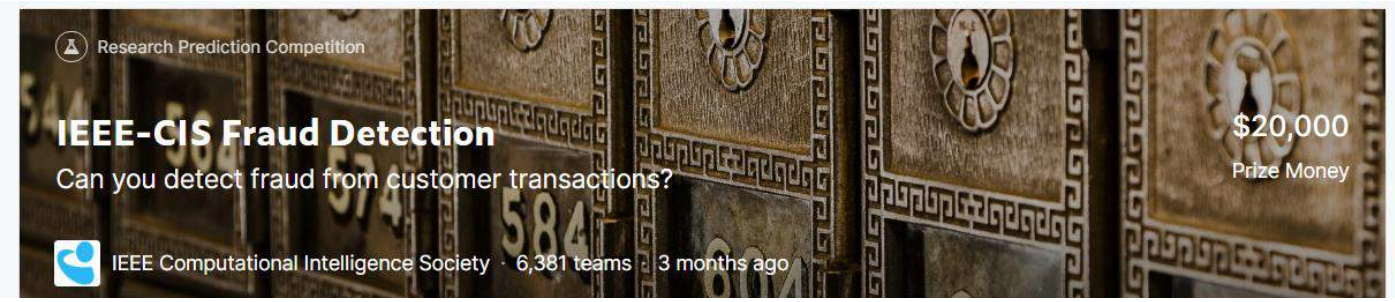
IEEE-CIS Fraud Detection

https://www.kaggle.com/c/ieee-fraud-detection

Data Competition in 2019

参赛队伍：6381，参赛人数：7416

比赛类型：Research，比赛数据：结构化



- 比赛背景

Imagine standing at the check-out counter at the grocery store with a long line behind you and the cashier not-so-quietly announces that your card has been declined. In this moment, you probably aren't thinking about the data science that determined your fate.

Embarrassed, and certain you have the funds to cover everything needed for an epic nacho party for 50 of your closest friends, you try your card again. Same result. As you step aside and allow the cashier to tend to the next customer, you receive a text message from your bank. "Press 1 if you really tried to spend $500 on cheddar cheese."

While perhaps cumbersome (and often embarrassing) in the moment, this fraud prevention system is actually saving consumers millions of dollars per year. Researchers from the IEEE Computational Intelligence Society (IEEE-CIS) want to improve this figure, while also improving the customer experience. With higher accuracy fraud detection, you can get on with your chips without the hassle.

IEEE-CIS works across a variety of AI and machine learning areas, including deep neural networks, fuzzy systems, evolutionary computation, and swarm intelligence. Today they're partnering with the world's leading payment service company, Vesta Corporation, seeking the best solutions for fraud prevention industry, and now you are invited to join the challenge.

In this competition, you'll benchmark machine learning models on a challenging large-scale dataset. The data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features. You also have the opportunity to create new features to improve your results.

If successful, you'll improve the efficacy of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue. And of course, you will save party people just like you the hassle of false positives.

- 评价指标：

Data Competition in 2019

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/ieee-fraud-detection/data

- 时间轴：Jun 16, 2019 – Oct 4, 2019
- 比赛结果

https://www.kaggle.com/c/ieee-fraud-detection/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/ieee-fraud-detection/discussion/111308

2nd place solution：https://www.kaggle.com/c/ieee-fraud-detection/discussion/111321

5th place solution：https://www.kaggle.com/c/ieee-fraud-detection/discussion/111735

The 3rd YouTube-8M Video Understanding Challenge

https://www.kaggle.com/c/youtube8m-2019

参赛队伍：283，参赛人数：341

比赛类型：Research，比赛数据：视频

- 比赛背景

Imagine being able to search for the moment in any video where an adorable kitten sneezes, even though the uploader didn't title or describe the video with such descriptive metadata. Now, apply that same concept to videos that cover important or special events like a baby's first steps or a game-winning goal -- and now we have the ability to quickly find and share special video moments. This technology is called temporal concept localization within video and Google Research can use your help to advance the state of the art in this area.

In most web searches, video retrieval and ranking is performed by matching query terms to metadata and other video-level signals. However, we know that videos can contain an array of topics that aren't always characterized by the uploader, and many of these miss localizations to brief but important moments within the video. Temporal localization can enable applications such as improved video search (including search within video), video summarization and highlight extraction, action moment detection, improved video content safety, and many others.

Data Competition in 2019

In previous years, participants worked on advancements in video-level annotations, building both unconstrained and constrained models. In this third challenge based on the YouTube 8M dataset, Kagglers will localize video-level labels to the precise time in the video where the label actually appears, and do this at an unprecedented scale. To put it another way: at what point in the video does the cat sneeze?

If successful, your new machine learning models will significantly improve video understanding for all, by not only identifying the topics relevant to a video, but also pinpointing where in the video they appear.

- 评价指标：

Submissions are evaluated according to the Mean Average Precision @ K (MAP@K).

- 比赛数据

https://www.kaggle.com/c/youtube8m-2019/data

- 时间轴：Jun 27, 2019 – Oct 12, 2019

- 比赛结果

https://www.kaggle.com/c/youtube8m-2019/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/112869

2nd place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/113663

3rd place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/112929

5th place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/112296

6th place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/112403

7th place solution：https://www.kaggle.com/c/youtube8m-2019/discussion/112349

Kuzushiji Recognition

https://www.kaggle.com/c/kuzushiji-recognition

参赛队伍：293，参赛人数：338

比赛类型：Playground，比赛数据：图像

- 比赛背景

Data Competition in 2019

Imagine the history contained in a thousand years of books. What stories are in those books? What knowledge can we learn from the world before our time? What was the weather like 500 years ago? What happened when Mt. Fuji erupted? How can one fold 100 cranes using only one piece of paper? The answers to these questions are in those books.

Japan has millions of books and over a billion historical documents such as personal letters or diaries preserved nationwide. Most of them cannot be read by the majority of Japanese people living today because they were written in “Kuzushiji”.

Even though Kuzushiji, a cursive writing style, had been used in Japan for over a thousand years, there are very few fluent readers of Kuzushiji today (only 0.01% of modern Japanese natives). Due to the lack of available human resources, there has been a great deal of interest in using Machine Learning to automatically recognize these historical texts and transcribe them into modern Japanese characters. Nevertheless, several challenges in Kuzushiji recognition have made the performance of existing systems extremely poor. (More information in About Kuzushiji)

This is where you come in. The hosts need help from machine learning experts to transcribe Kuzushiji into contemporary Japanese characters. With your help, Center for Open Data in the Humanities (CODH) will be able to develop better algorithms for Kuzushiji recognition. The model is not only a great contribution to the machine learning community, but also a great help for making millions of documents more accessible and leading to new discoveries in Japanese history and culture.

- 评价指标：

Submissions are evaluated on a modified version of the F1 Score.

- 比赛数据

https://www.kaggle.com/c/kuzushiji-recognition/data

- 时间轴：Jul 19, 2019 – Oct 15, 2019
- 比赛结果

https://www.kaggle.com/c/kuzushiji-recognition/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/kuzushiji-recognition/discussion/112788

2nd place solution：https://www.kaggle.com/c/kuzushiji-recognition/discussion/112712

3rd place solution：https://www.kaggle.com/c/kuzushiji-recognition/discussion/113049

8th place solution：https://www.kaggle.com/c/kuzushiji-recognition/discussion/113419

9th place solution：https://www.kaggle.com/c/kuzushiji-recognition/discussion/112807

Severstal: Steel Defect Detection

https://www.kaggle.com/c/severstal-steel-defect-detection

参赛队伍：2431，参赛人数：2870

比赛类型：Featured，比赛数据：图像

- 比赛背景

Steel is one of the most important building materials of modern times. Steel buildings are resistant to natural and man-made wear which has made the material ubiquitous around the world. To help make production of steel more efficient, this competition will help identify defects.

Severstal is leading the charge in efficient steel mining and production. They believe the future of metallurgy requires development across the economic, ecological, and social aspects of the industry—and they take corporate responsibility seriously. The company recently created the country's largest industrial data lake, with petabytes of data that were previously discarded. Severstal is now looking to machine learning to improve automation, increase efficiency, and maintain high quality in their production.

The production process of flat sheet steel is especially delicate. From heating and rolling, to drying and cutting, several machines touch flat steel by the time it's ready to ship. Today, Severstal uses images from high frequency cameras to power a defect detection algorithm.

In this competition, you'll help engineers improve the algorithm by localizing and classifying surface defects on a steel sheet.

If successful, you'll help keep manufacturing standards for steel high and enable Severstal to continue their innovation, leading to a stronger, more efficient world all around us.

- 评价指标：

This competition is evaluated on the mean Dice coefficient.

Data Competition in 2019

- 比赛数据

https://www.kaggle.com/c/severstal-steel-defect-detection/data

- 时间轴：Jul 26, 2019 – Oct 25, 2019
- 比赛结果

https://www.kaggle.com/c/severstal-steel-defect-detection/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/114254

3rd place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/117377

4th place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/114716

5th place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/117208

7th place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/114215

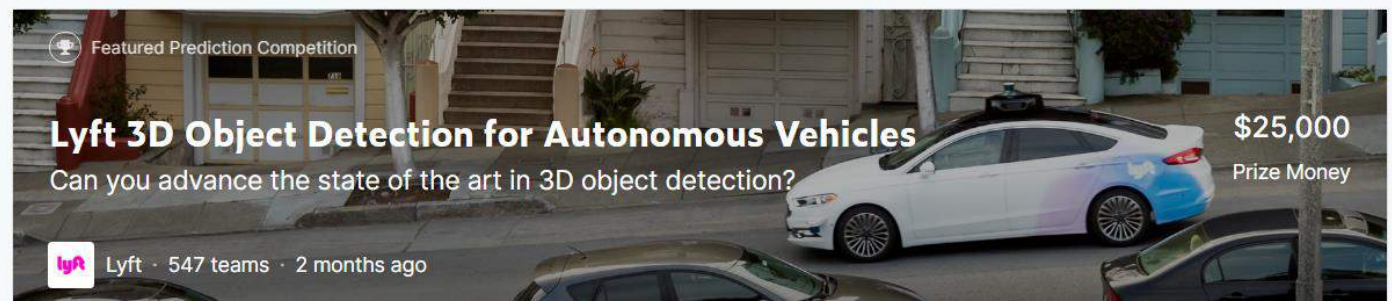
9th place solution：https://www.kaggle.com/c/severstal-steel-defect-detection/discussion/114297

Lyft 3D Object Detection for Autonomous Vehicles

https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles

参赛队伍：547，参赛人数：660

比赛类型：Featured，比赛数据：图像



- 比赛背景

Self-driving technology presents a rare opportunity to improve the quality of life in many of our communities. Avoidable collisions, single-occupant commuters, and vehicle emissions are choking cities, while infrastructure strains under rapid urban growth. Autonomous vehicles are expected to redefine transportation and unlock a myriad of societal, environmental, and economic benefits. You can apply your data analysis skills in this competition to advance the state of self-driving technology.

Lyft, whose mission is to improve people's lives with the world's best transportation, is investing in the future of self-driving vehicles. Level 5, their self-driving division, is working on a fleet of autonomous vehicles, and currently has a team of 450+ across Palo Alto, London, and Munich working to build a leading self-driving system (they're hiring!). Their goal is to democratize access to self-driving technology for hundreds of millions of Lyft passengers.

From a technical standpoint, however, the bar to unlock technical research and development on higher-level autonomy functions like perception, prediction, and planning is extremely high. This implies technical R&D on self-driving cars has traditionally been inaccessible to the broader research community.

This dataset aims to democratize access to such data, and foster innovation in higher-level autonomy functions for everyone, everywhere. By conducting a competition, we hope to encourage the research community to focus on hard problems in this space—namely, 3D object detection over semantic maps.

In this competition, you will build and optimize algorithms based on a large-scale dataset. This dataset features the raw sensor camera inputs as perceived by a fleet of multiple, high-end, autonomous vehicles in a restricted geographic area.

If successful, you'll make a significant contribution towards stimulating further development in autonomous vehicles and empowering communities around the world.

- 评价指标：

This competition is evaluated on the mean average precision at different intersection over union (IoU) thresholds.

- 比赛数据

https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles/data

- 时间轴：Sep 13, 2019 – Nov 13, 2019
- 比赛结果

https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles/discussion/122820

2nd place solution：https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles/discussion/123004

2rd place solution：https://www.kaggle.com/c/3d-object-detection-for-autonomous-vehicles/discussion/117269

RSNA Intracranial Hemorrhage Detection

https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection

参赛队伍：1345，参赛人数：2550

比赛类型：Featured，比赛数据：图像

- 比赛背景

ntracranial hemorrhage, bleeding that occurs inside the cranium, is a serious health problem requiring rapid and often intensive medical treatment. For example, intracranial hemorrhages account for approximately 10% of strokes in the U.S., where stroke is the fifth-leading cause of death. Identifying the location and type of any hemorrhage present is a critical step in treating the patient.

Diagnosis requires an urgent procedure. When a patient shows acute neurological symptoms such as severe headache or loss of consciousness, highly trained specialists review medical images of the patient's cranium to look for the presence, location and type of hemorrhage. The process is complicated and often time consuming.

In this competition, your challenge is to build an algorithm to detect acute intracranial hemorrhage and its subtypes.

You'll develop your solution using a rich image dataset provided by the Radiological Society of North America (RSNA®) in collaboration with members of the American Society of Neuroradiology and MD.ai.

If successful, you'll help the medical community identify the presence, location and type of hemorrhage in order to quickly and effectively treat affected patients.

Challenge participants may be invited to present their AI models and methodologies during an award ceremony at the RSNA Annual Meeting which will be held in Chicago, Illinois, USA, from December 1-6, 2019.

Data Competition in 2019

- 评价指标：

Submissions are evaluated using a weighted multi-label logarithmic loss.

- 比赛数据

https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/data

- 时间轴：Sep 18, 2019 – Nov 14, 2019
- 比赛结果

https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/discussion/117210

2nd place solution：https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/discussion/117228

4th place solution：https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/discussion/118249

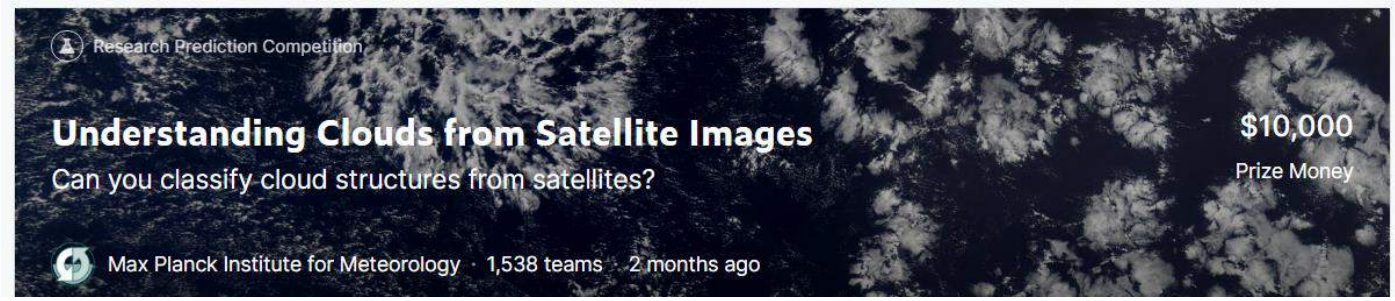
5th place solution：https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection/discussion/117232

Understanding Clouds from Satellite Images

https://www.kaggle.com/c/understanding_cloud_organization

参赛队伍：1538，参赛人数：1859

比赛类型：Research，比赛数据：图像



- 比赛背景

Climate change has been at the top of our minds and on the forefront of important political decision-making for many years. We hope you can use this competition's dataset to help demystify an important climatic variable. Scientists, like those at Max Planck Institute for Meteorology, are leading the charge with new research on the world's ever-changing atmosphere and they need your help to better understand the clouds.

Shallow clouds play a huge role in determining the Earth's climate. They're also difficult to understand and to represent in climate models. By classifying different types of cloud organization, researchers at Max Planck hope to improve our physical understanding of these clouds, which in turn will help us build better climate models.

There are many ways in which clouds can organize, but the boundaries between different forms of organization are murky. This makes it challenging to build traditional rule-based algorithms to separate cloud features. The human eye, however, is really good at detecting features—such as clouds that resemble flowers.

In this challenge, you will build a model to classify cloud organization patterns from satellite images. If successful, you'll help scientists to better understand how clouds will shape our future climate. This research will guide the development of next-generation models which could reduce uncertainties in climate projections.

- 评价指标：

This competition is evaluated on the mean Dice coefficient.

- 比赛数据

https://www.kaggle.com/c/understanding_cloud_organization/data

- 时间轴：Aug 16, 2019 – Nov 19, 2019
- 比赛结果

https://www.kaggle.com/c/understanding_cloud_organization/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/understanding_cloud_organization/discussion/118080

2nd place solution：https://www.kaggle.com/c/understanding_cloud_organization/discussion/118255

3rd place solution：https://www.kaggle.com/c/understanding_cloud_organization/discussion/118065

6th place solution：https://www.kaggle.com/c/understanding_cloud_organization/discussion/118017

10th place solution：https://www.kaggle.com/c/understanding_cloud_organization/discussion/118262

Categorical Feature Encoding Challenge

https://www.kaggle.com/c/cat-in-the-dat

Data Competition in 2019

参赛队伍：1342，参赛人数：1391

比赛类型：Playground，比赛数据：结构化

- 比赛背景

A common task in machine learning pipelines is encoding categorical variables for a given algorithm in a format that allows as much useful signal as possible to be captured.

Because this is such a common task and important skill to master, we've put together a dataset that contains only categorical features, and includes:

✓ binary features

✓ low- and high-cardinality nominal features

✓ low- and high-cardinality ordinal features

✓ (potentially) cyclical features

This Playground competition will give you the opportunity to try different encoding schemes for different algorithms to compare how they perform. We encourage you to share what you find with the community.

- 评价指标：

Submissions are evaluated on area under the ROC curve between the predicted probability and the observed target.

- 比赛数据

https://www.kaggle.com/c/cat-in-the-dat/data

- 时间轴：Aug 23, 2019 – Dec 10, 2019
- 比赛结果

https://www.kaggle.com/c/cat-in-the-dat/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/cat-in-the-dat/discussion/121356

2nd place solution：https://www.kaggle.com/c/cat-in-the-dat/discussion/121063

3rd place solution：https://www.kaggle.com/c/cat-in-the-dat/discussion/122649

4th place solution：https://www.kaggle.com/c/cat-in-the-dat/discussion/121584

BigQuery-Geotab Intersection Congestion

https://www.kaggle.com/c/bigquery-geotab-intersection-congestion

参赛队伍：436，参赛人数：491

比赛类型：Playground，比赛数据：结构化

- 比赛背景

We've all been there: Stuck at a traffic light, only to be given mere seconds to pass through an intersection, behind a parade of other commuters. Imagine if you could help city planners and governments anticipate traffic hot spots ahead of time and reduce the stop-and-go stress of millions of commuters like you.

Geotab provides a wide variety of aggregate datasets gathered from commercial vehicle telematics devices. Harnessing the insights from this data has the power to improve safety, optimize operations, and identify opportunities for infrastructure challenges.

The dataset for this competition includes aggregate stopped vehicle information and intersection wait times. Your task is to predict congestion, based on an aggregate measure of stopping distance and waiting times, at intersections in 4 major US cities: Atlanta, Boston, Chicago & Philadelphia.

This competition is being hosted in partnership with BigQuery, a data warehouse for manipulating, joining, and querying large scale tabular datasets. BigQuery also offers BigQuery ML, an easy way for users to create and run machine learning models to generate predictions through a SQL query interface.

Kaggle recently released a BigQuery integration within our kernels notebook environment, and this starter kernel gives you a great starting point for how to use BQ & BQML. You're encouraged to use your data savvy, resourcefulness & intuition to find and join in additional external datasets that will increase your models' predictive power.

- 评价指标：

Submissions are scored on the root mean squared error.

Data Competition in 2019

- 比赛数据

https://www.kaggle.com/c/bigquery-geotab-intersection-congestion/data

- 时间轴：Sep 13, 2019 – Dec 13, 2019
- 比赛结果

https://www.kaggle.com/c/bigquery-geotab-intersection-congestion/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/bigquery-geotab-intersection-congestion/discussion/124567

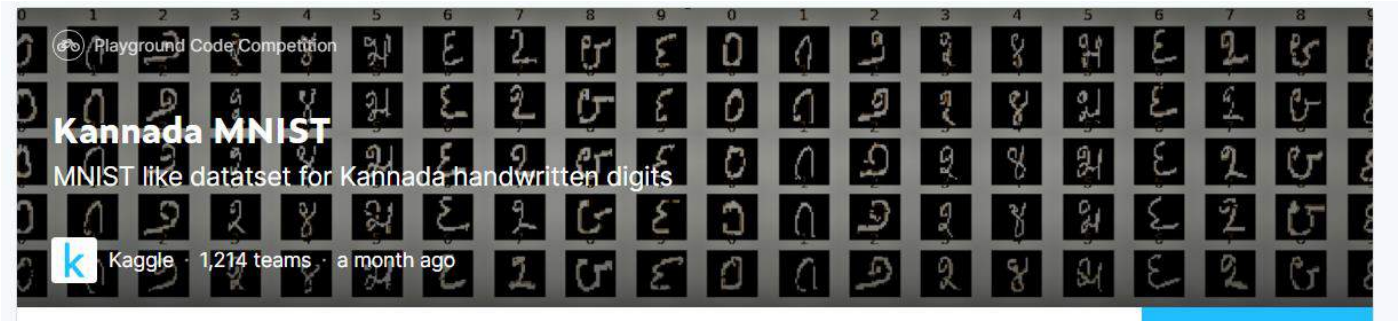
2nd place solution：https://www.kaggle.com/c/bigquery-geotab-intersection-congestion/discussion/122314

Kannada MNIST

https://www.kaggle.com/c/Kannada-MNIST

参赛队伍：1214，参赛人数：1344

比赛类型：Playground，比赛数据：图像



- 比赛背景

The goal of this competition is to provide a simple extension to the classic MNIST competition we're all familiar with. Instead of using Arabic numerals, it uses a recently-released dataset of Kannada digits.

Kannada is a language spoken predominantly by people of Karnataka in southwestern India. The language has roughly 45 million native speakers and is written using the Kannada script. Wikipedia

Data Competition in 2019

This competition uses the same format as the MNIST competition in terms of how the data is structured, but it's different in that it is a synchronous re-run Kernels competition. You write your code in a Kaggle Notebook, and when you submit the results, your code is scored on both the public test set, as well as a private (unseen) test set.

- 评价指标：

This competition is evaluated on the categorization accuracy of your predictions (the percentage of images you get correct).

- 比赛数据

https://www.kaggle.com/c/Kannada-MNIST/data

- 时间轴：Sep 18, 2019 – Dec 18, 2019
- 比赛结果

https://www.kaggle.com/c/Kannada-MNIST/leaderboard

- 赛后分享

2nd place solution：https://www.kaggle.com/c/Kannada-MNIST/discussion/122230

3rd place solution：https://www.kaggle.com/c/Kannada-MNIST/discussion/122167

4th place solution：https://www.kaggle.com/c/Kannada-MNIST/discussion/122430

5th place solution：https://www.kaggle.com/c/Kannada-MNIST/discussion/122160

ASHRAE - Great Energy Predictor III

https://www.kaggle.com/c/ashrae-energy-prediction

参赛队伍：3614

比赛类型：Featured，比赛数据：结构化

- 比赛背景

Thankfully, significant investments are being made to improve building efficiencies to reduce costs and emissions. The question is, are the improvements working? That's where you come in. Under pay-for-performance financing, the building owner makes payments based on the difference between their real energy consumption and what they would have used without any retrofits. The latter values have to come from a model. Current methods of estimation are fragmented and do not scale well. Some assume a specific meter type or don't work with different building types.

In this competition, you'll develop accurate models of metered building energy usage in the following areas: chilled water, electric, hot water, and steam meters. The data comes from over 1,000 buildings over a three-year timeframe. With better estimates of these energy-saving investments, large scale investors and financial institutions will be more inclined to invest in this area to enable progress in building efficiencies.

- 评价指标：

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.

- 比赛数据

https://www.kaggle.com/c/ashrae-energy-prediction/data

- 时间轴：Oct 16, 2019 – Dec 20, 2019
- 比赛结果

https://www.kaggle.com/c/ashrae-energy-prediction/leaderboard

- 赛后分享

4th place solution：https://www.kaggle.com/c/ashrae-energy-prediction/discussion/124788

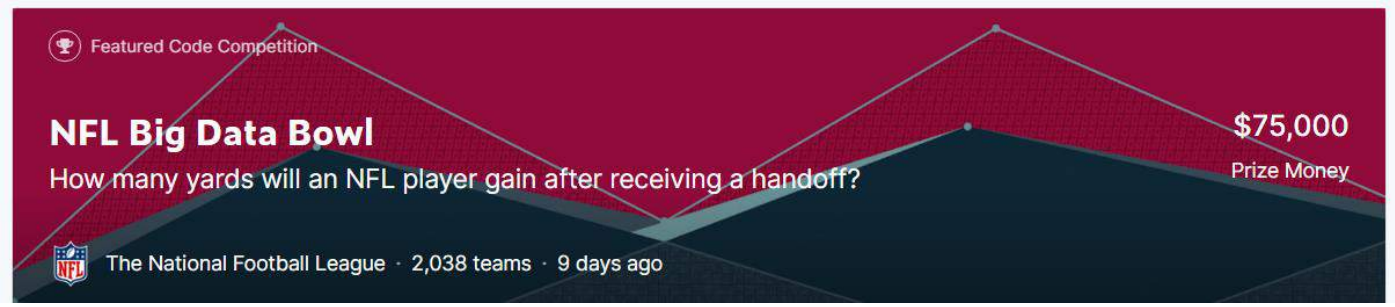
9th place solution：https://www.kaggle.com/c/ashrae-energy-prediction/discussion/123525

NFL 1st and Future - Analytics

https://www.kaggle.com/c/nfl-playing-surface-analytics

参赛队伍：2038，参赛人数：3101

比赛类型：Featured，比赛数据：结构化



- 比赛背景

Thankfully, significant investments are being made to improve building efficiencies to reduce costs and emissions. The question is, are the improvements working? That's where you come in. Under pay-for-performance financing, the building owner makes payments based on the difference between their real energy consumption and what they would have used without any retrofits. The latter values have to come from a model. Current methods of estimation are fragmented and do not scale well. Some assume a specific meter type or don't work with different building types.

In this competition, you'll develop accurate models of metered building energy usage in the following areas: chilled water, electric, hot water, and steam meters. The data comes from over 1,000 buildings over a three-year timeframe. With better estimates of these energy-saving investments, large scale investors and financial institutions will be more inclined to invest in this area to enable progress in building efficiencies.

- 评价指标：

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.

- 比赛数据

https://www.kaggle.com/c/nfl-big-data-bowl-2020/data

- 时间轴：Oct 16, 2019 – Dec 20, 2019
- 比赛结果

https://www.kaggle.com/c/nfl-big-data-bowl-2020/leaderboard

- 赛后分享

1st place solution：https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/119400

4th place solution：https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/121397

7th place solution：https://www.kaggle.com/c/nfl-big-data-bowl-2020/discussion/124999

1.2 天池

津南数字制造算法挑战赛【赛场一】

https://tianchi.aliyun.com/competition/entrance/231695/introduction

参赛队伍：2682，比赛类型：结构化

- 比赛背景

异烟酸用作医药中间体，主要用于制抗结核病药物异烟肼，也用于合成酰胺、酰肼、酯类等衍生物。烟酰胺生产过程包含水解脱色、结晶甩滤等过程。每个步骤会受到温度、时间、压强等各方面因素的影响，造成异烟酸收率的不稳定。为保证产品质量和提高生产效率，需要调整和优化生产过程中的参数。然而，根据传统经验的人工调整工艺参数费时费力。近年来，人工智能在工艺参数优化以及视频检测等领域取得了突飞猛进的成果。AI 技术的发展有望助力原料药制造企业实现工艺生产革新，规范生产操作过程，从而达到提高产品的收率的目标。

本次大赛要求选手以异烟酸生产过程中的各参数，包括各主要步骤的时间、温度、压强等参数为基础，设计精确智能的优秀算法，提升异烟酸的收率。阿里云将为参赛选手提供机器资源，复赛团队可申请使用。

- 评价指标

选手提交结果与实际检测到的收率结果进行对比，以均方误差为评价指标，结果越小越好，均方误差计算公式如下：

$$f = \frac{1}{2m}\sum_{i=1}^{m}(y'(i) - y(i))^2$$

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231695/information
- 比赛时间：2019 年 1 月 2 日 – 2019 年 1 月 31 日

Data Competition in 2019

- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231695/rankingList
- 赛后分享

冠军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.6.54587b9delQbXD&postId=54381

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.12.54587b9delQbXD&postId=54342

季军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.9.54587b9delQbXD&postId=54439

津南数字制造算法挑战赛【赛场二】

https://tianchi.aliyun.com/competition/entrance/231703/introduction

参赛队伍：2157，比赛类型：计算机视觉、图像、目标检测

- 比赛背景

包裹 X 光限制品监测作为日常包裹物流行业及安防行业的重要环节，承担着防止易燃易爆等危险品进入货运渠道，管理刀具等特殊货运物品，监测毒品等国家重点违禁品偷运等工作。随着线上购物的普及和快速发展，线上物流包裹数量已经远超人工可以处理的范围，给物流包裹监管带来了巨大挑战。

本次比赛着眼于此，邀请选手针对给出的限制品种类，利用 X 光图像及标注数据，研究开发高效的计算机视觉算法，监测图像是否包含危险品及其大致位置。通过自动化监测包裹携带品算法，降低漏检风险及误报率，提升危险品管理效率。本次比赛是一次未来物流及安防行业的有益尝试，对行业未来发展有着不可估量的价值。

- 评价指标

初赛：评测方式采用计算 box mAP 的方式，对 IoU = 0.5:0.05:0.95，分别计算 mAP，再做平均得到最后的 mAP

Data Competition in 2019

复赛及决赛：5 类 IoU 的平均值

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231703/information
- 比赛时间：2019 年 3 月 1 日 – 2019 年 5 月 15 日
- 比赛结果

https://tianchi.aliyun.com/competition/entrance/231703/rankingList

- 赛后分享

冠军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.9.8dc78064n7T6nb&postId=57411

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.12.8dc78064n7T6nb&postId=57849

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.18.8dc78064n7T6nb&postId=57797

全球城市计算 AI 挑战赛

https://tianchi.aliyun.com/competition/entrance/231708/introduction

参赛队伍：2319，比赛类型：结构化

- 比赛背景

2019 年，杭州市公安局联合阿里云智能启动首届全球城市计算 AI 挑战赛，本次挑战赛的题目选定为“地铁乘客流量预测”。地铁目前是城市交通出行的主要工具之一，地铁站突发人流量的增加极容易引起拥塞，引发大客流对冲，造成安全隐患。因此，地铁运营部门和公安机关亟需通过流量预测技术提前部署相应的安保策略，保障市民安全出行。

参赛者将根据主办方提供的地铁人流相关数据，挖掘隐藏在背后的出行规律，准确预测各个地铁站点未来流量的变化。主办方希望通过这次挑战赛，用大数据和人工智能等技术助力未来城市安全出行。

- 评价指标

评估指标用以评判选手对未来一天以 10 分钟为单位各时段各地铁站的出站和入站人次的总量预测是否准确，因此采用平均绝对误差（Mean Absolute Error，MAE）分别对入站人数和出站人数预测结果进行评估，最后再对两者取平均，得到最终评分。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231708/information
- 比赛时间：2019 年 3 月 20 日 – 2019 年 4 月 11 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231708/rankingList
- 赛后分享：

第一名：https://tianchi.aliyun.com/notebook-ai/detail?postId=54550

第三名：https://tianchi.aliyun.com/notebook-ai/detail?postId=56649

第五名：https://tianchi.aliyun.com/forum/postDetail?postId=54653

第七名：https://tianchi.aliyun.com/notebook-ai/detail?postId=54574

第八名：https://tianchi.aliyun.com/notebook-ai/detail?postId=54557

第 14 名：https://tianchi.aliyun.com/forum/postDetail?postId=54599

第 20 名：https://tianchi.aliyun.com/forum/postDetail?postId=55137

2019 Future Food Challenge

https://tianchi.aliyun.com/competition/entrance/231705/introduction

参赛队伍：661，比赛类型：自然语言处理

- 比赛背景

Home to a quarter of the world's population, the South-East Asia has the second highest burden of foodborne diseases per population. In this region, foodborne diseases are responsible for 150 Mio illnesses and 175,000 deaths annually [Source: WHO]. Data analytics has the potential to become an important tool for both producers and consumers to reduce the number of foodborne illnesses.

In order to protect consumer health and lower the risk of foodborne incidents Bühler has recently launched the freemium service safefood.ai to create additional intelligence on food-safety related events such as product recalls, foodborne outbreaks or fraudulent activity.

The service continuously monitors thousands of data sources, including official databases, social media, news and others, and alerts users in close to real-time about any evidence related to foodborne outbreaks, product recalls or food fraud. One of the value propositions for the service is to detect food safety related events as early as possible and identify related news items to give an estimate of the scale of the event.

In this challenge we will provide an extract of news items which have been collected in the past weeks. In the training set you will find examples of what type of events we are looking for. For the test set we ask you to auto-discover new events, provide a description of the event, as well as the first news item that provides information on the event.

- 评价指标

The teams will be ranked according to their OVERALL SCORE, the overall score will be given based on three separate scores of their submissions.

File 1 Score: assess how well can the teams distinguish between relevant and other news items by the recall/ precision/ accuracy of the File 1.

File 2 Score: assess how well can the teams distinguish between the individual events by comparing the File 2 with our standard answer.

Data Competition in 2019

File 3 Score: assess the quality of the report (File 3) by the completeness/ insights and innovation shows in the report.

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231705/information
- 比赛时间：2019 年 4 月 1 日 – 2019 年 6 月 4 日
- 比赛结果

https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.3.25a85e37ZZGwZV&postId=57248

- 赛后分享：

第五名：https://unclegem.cn/2019/08/27/2019-Future-Food-Challenge%E7%AB%9E%E8%B5%9B%E6%80%BB%E7%BB%93/

IJCAI-19 阿里巴巴人工智能对抗算法竞赛

https://tianchi.aliyun.com/competition/entrance/231701/introduction

参赛队伍：2519，比赛类型：计算机视觉、图像分类、对抗样本

- 比赛背景

近年来深度学习技术的不断突破，极大促进了人工智能行业的发展，而人工智能模型本身的安全问题，也日益受到 AI 从业人员的关注。2014 年 Christian Szegedy 等人首次提出针对图像的对抗样本这一概念，并用实验结果展示了深度学习模型在安全方面的局限性。我们可以通过对原始样本有针对性的加入微小扰动来构造对抗样本，该扰动不易被人眼所察觉，但会导致 AI 模型识别错误，这种

攻击被称为“对抗攻击“。该成果引发了学术界和工业界的广泛关注，成为目前深度学习领域最热的研究课题之一，新的对抗攻击方法不断涌现，应用课题从图像分类扩展到目标检测。对抗攻击技术也引发了业界对于 AI 模型安全的担忧，研究人员开展了针对对抗攻击的防御技术研究，也提出了若干种提升模型安全性能的方法，但迄今为止仍然无法完全防御来自对抗样本的攻击。

IJCAI-19 阿里巴巴人工智能对抗算法竞赛的目的是对 AI 模型的安全性进行探索。这个比赛主要针对图像分类任务，包括模型攻击与模型防御。参赛选手既可以作为攻击方，对图片进行轻微扰动生成对抗样本，使模型识别错误；也可以作为防御方，通过构建一个更加鲁棒的模型，准确识别对抗样本。

本次比赛，首次采用电商场景的图片识别任务进行攻防对抗。总共会公开 110,000 左右的商品图片，来自 110 个商品类目，每个类目大概 1000 个图片。选手可以使用这些数据训练更加鲁棒的识别模型或者生成更具攻击性的样本。

本次比赛包括以下三个任务：

1. 无目标攻击: 生成对抗样本，使模型识别错误。
2. 目标攻击: 生成对抗样本，使模型识别到指定的错误类别。
3. 模型防御: 构建能够正确识别对抗样本的模型。

- 评价指标：https://tianchi.aliyun.com/competition/entrance/231701/information
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231701/information
- 比赛时间：2019 年 3 月 4 日 – 2019 年 6 月 10 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231701/rankingList
- 赛后分享

阿里巴巴优酷视频增强和超分辨率挑战赛

https://tianchi.aliyun.com/competition/entrance/231711/introduction

参赛队伍：1514，比赛类型：计算机视觉、视频增强和超分

- 比赛背景

视频增强和超分是计算机视觉领域的核心算法之一，目的是恢复降质视频本身的内容，提高视频的清晰度。该技术在工业界有着重要的实用意义，对于早期胶片视频的质量和清晰度的提升有着重大的意义。

本竞赛不仅提供一个平台，让大家展示最前沿的视频增强和超分算法，而且给出了业界最大、最具广泛性的数据集，包括不同内容品类，不同噪声模型、不同难度等。该数据集的生成模型完全是模拟实际业务中的噪声模式，研究人员可以真正的在实际场景中打磨算法，推动视频增强和超分算法在实际问题中的应用，促进工业界和学术界的深度合作。

比赛要求通过训练样本对视频增强和超分模型进行建模，对测试集中的低分辨率视频样本预测高分辨率视频。其中，高分辨率视频来自优酷高清媒资库，优酷拥有视频的知识产权。低分辨率视频的生成模型是模拟实际业务中的噪声模式。

因此，解决此问题对视频产业有重要的贡献。为了更好的研究该问题，优酷将建立业界最大、最具有广泛性的视频超分和增强数据集，数据集将包括 10000+视频对，包括不同内容品类，不同难度、不同业务场景下的噪声模型等等。第一批 1000 个视频数据集供本次比赛使用；比赛结束后，公开约 2000 个视频；剩余 7000 个视频也将逐步公开。

- 评价指标

评估程序最终 VMAF 结果为完整视频所有帧 VMAF 结果的平均值；最终 PSNR 的结果为完整视频和抽帧视频中所有帧的平均值。PSNR 和 VMAF 得分进行加权得到竞赛得分。

score = PSNR 指标得分×80% + VMAF 指标得分×20%

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231711/information
- 比赛时间：2019 年 5 月 5 日 – 2019 年 8 月 7 日

Data Competition in 2019

- 比赛结果： https://tianchi.aliyun.com/competition/entrance/231711/rankingList
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/84749080

全球数据智能大赛【赛场一】

https://tianchi.aliyun.com/competition/entrance/231724/introduction

参赛队伍：1635，比赛类型：计算机视觉、目标检测

- 比赛背景

全球数据智能大赛(2019)由广西壮族自治区人民政府主办，面向全球人工智能优秀团队征集多领域的算法智力成果和解决方案，集结众智，探索数字广西的蓝图，共领数据智能未来发展。大赛共设“数字人体”和“科技扶农”两个赛场，聚焦医疗和农业真实痛点。

中国是肺部疾病大国，肺炎、肺癌的发病率在全球水平较高。肺部疾病除了会引起与肺部有关的疾病之外，还会引发多种并发症，如：气管炎、心脏疾病、淋巴系统疾病等。作为肺部疾病最有效的无创检测技术，胸部 CT 影像以其分层薄、高清、低噪声等优点，被广泛用于肺部疾病筛查和辅助诊断。

人工阅片一方面耗时耗力，存在漏检、错检的误差。而海量的影像数据每天都被生产出来，机器阅片帮助人工做病灶位置粗筛、疾病辅助诊已经成为潮流。在人工智能技术加持下，机器阅片具有速度快、准确率高、高并发等优势。以肺结节为例，据不完全统计，目前拿出肺结节产品的企业达到数十家，在核心医院影像科普遍有 3 家以上的肺结节系统。在肺结节单影像特征的临床实践中，AI 技术已经取得了较好的效果，下一步的重点，是对肺部多种疾病进行智能综合诊治。

赛场一“数字人体”挑战赛以肺部 CT 多病种智能诊断为课题，开放高质量 CT 标注数据，要求选手提出并综合运用目标检测、深度学习等人工智能算法，识别肺结节、索条（条索状影）、动脉硬化

或钙化、淋巴结钙化等多个影像特征，避免同一部位单影像特征的反复筛查，提高检测的速度和精度，辅助医生进行诊断。

- 评价指标

根据提供的每种病灶的检测概率，计算一个 FROC 曲线，Sensitivity 在 1/8, 1/4, 1/2, 1, 2, 4 和 8 一共 7 个不同的误报情况下的平均值作为其中一种病灶的得分，最后的得分是四种病灶得分的平均值。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231724/information
- 比赛时间：2019 年 6 月 25 日 – 2019 年 8 月 28 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231724/rankingList
- 赛后分享

冠军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.18.3389c6b0fxC624&postId=73774

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.3.3389c6b0fkxJ4i&postId=74041

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.6.3389c6b0fkxJ4i&postId=74162

2019 年县域农业大脑 AI 挑战赛

https://tianchi.aliyun.com/competition/entrance/231717/introduction

参赛队伍：1520，比赛类型：计算机视觉、图像分类

- 比赛背景

农作物的资产盘点与精准产量预测是实现农业精细化管理的核心环节。当前，我国正处于传统农业向现代农业的加速转型期，伴随着农业的转型升级，政府宏观决策、社会各界对农业数据的需求不断增加，现有农业统计信息的时效性与质量，已不足以为市场各主体的有效决策提供科学依据。在

农作物资产盘点方面，传统的人工实地调查的方式速度慢、劳动强度大，数据采集质量受主观因素影响大，统计数据有较大的滞后性，亟待探索研究更高效准确度更高的农业调查统计技术。在产量预测方面，及时准确地获取区域作物单产及其空间分布信息，对作物进行精准的产能预测，对于农业生产安全预警、农产品贸易流通，以及农业产业结构优化具有重要意义。

本次大赛，我们选择了具有独特的地理环境、气候条件以及人文特色的贵州省兴仁市作为研究区域，聚焦当地的特色优势产业和支柱产业——薏仁米产业， 以薏仁米作物识别以及产量预测为比赛命题，要求选手开发算法模型，通过无人机航拍的地面影像，探索作物分类的精准算法，识别薏仁米、玉米、烤烟、人造建筑四大类型，提升作物识别的准确度，降低对人工实地勘察的依赖，提升农业资产盘点效率，并结合产量标注数据预测当年的薏仁米产量，提升农业精准管理能力。

- 评价指标：赛题的评估指标为 mIoU，为所有计算所有类别 IoU 后取平均的结果。
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231717/information
- 比赛时间：2019 年 6 月 20 日 – 2019 年 8 月 30 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231717/rankingList
- 赛后分享

冠军：
https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.6.25ea4054jfvidy&postId=78945

亚军：
https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.3.25ea4054LuhpRe&postId=79094

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.6.3389c6b0fkxJ4i&postId=74162

首届中文 NL2SQL 挑战赛

https://tianchi.aliyun.com/competition/entrance/231716/introduction

参赛队伍：1458，比赛类型：自然语言处理

- 比赛背景

数据库中存储了海量的高价值数据，用户可以通过执行 SQL 与结构化数据直接进行交互，也可以通过设计好的交互界面进行交互。但 SQL 的使用难度限制了非技术用户，交互界面的设计也限制了使用的界限。通过自然语言直接与结构化数据进行交互，可以充分利用结构化数据的价值，为用户带来体验和效率的提升。

首届中文 NL2SQL 挑战赛，使用金融以及通用领域的表格数据作为数据源，提供在此基础上标注的自然语言与 SQL 语句的匹配对，希望选手可以利用数据训练出可以准确转换自然语言到 SQL 的模型。

- 评价指标

Logic Form Accuracy: 预测完全正确的 SQL 语句。其中，列的顺序并不影响准确率的计算。

Execution Accuracy: 预测的 SQL 的执行结果与真实 SQL 的执行结果一致。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231716/information
- 比赛时间：2019 年 6 月 24 日 – 2019 年 9 月 9 日
- 比赛结果

奖项	获奖团队	团队成员
冠军	不上90不改名字	张啸宇、赛斌、王苏宏
亚军	BugCreater	戴威、戴泽辉、陈华杰
季军	Model S	陈曦、辜乘风、黄伊
优胜奖	老哥们不放假吗	赵猛、任雪峰
优胜奖	大佬带我飞	戴志港

- 赛后分享

冠军：https://tianchi.aliyun.com/forum/postDetail?postId=78781

亚军：https://tianchi.aliyun.com/forum/postDetail?postId=78855

季军： https://tianchi.aliyun.com/forum/postDetail?postId=78747

第二届阿里巴巴大数据智能云上编程大赛-智联招聘人岗智能匹配

https://tianchi.aliyun.com/competition/entrance/231728/introduction

参赛队伍：1261，比赛类型：数据挖掘、推荐排序

- 比赛背景

阿里云计算平台深耕大数据以及人工智能领域，经过多年锤炼，推出了 MaxCompute、Dataworks、PAI、EMR 等多款大数据相关领域云产品。智联招聘作为国内大型的综合性招聘平台，二十多年行业深耕，为海量的求职者创造就业机会，帮助企业找到心仪的人才。本次比赛将由智联提供相关比赛数据和评估标准，选手需要使用阿里云计算平台的相关产品完成赛题。

本次大赛要求参赛者根据智联招聘抽样的经过脱敏的求职者标签数据、职位信息、及部分求职者行为信息、用人单位反馈信息，训练排序模型，对求职者的职位候选集进行排序，尽可能使得双端都满意的职位（求职者满意以及用人单位满意）优先推荐。本次比赛里，假定对于曝光给求职者的职位候选集里，假如求职者感兴趣会产生浏览职位行为，浏览职位后，如果求职者满意会产生主动投递行为。用人单位收到求职者主动投递的简历后会给出是否满意的反馈信号。

- 评价指标

测试数据由 n 组曝光职位数据集合组成，每组数据包含一个求职者以及一序列曝光候选职位。参赛者需要对每组职位进行预测并排序给出排序后的职位序列。对 n 组排序后的职位序列，比赛采用以下计算方式作为评估指标。通过计算所有 n 组排序后的职位序列里，求职者投递(delivered，0.3)职位的 MAP 值以及用人单位中意(satisfied，0.7)职位的 MAP（Mean Average Precision），由最终的加权评价值作为模型的评价指标，分数越高表示预测效果越好。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231728/information
- 比赛时间：2019 年 7 月 24 日 – 2019 年 9 月 21 日
- 比赛结果

奖项	获奖团队	团队成员
冠军	azu	黄益德
亚军	北方的郎	陈宇
季军	江离	花志祥、郭鹏博
优胜奖	OTTO	刘世欢、陈欣、王艳
优胜奖	alaskaw	翁灿栋

- 赛后分享

冠军：https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.15.3a324adaWWSm4S&postId=78122

亚军：https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.12.3a324adaWWSm4S&postId=78439

亚军：https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.18.3a324adaWWSm4S&postId=78158

安泰杯 —— 跨境电商智能算法大赛

https://tianchi.aliyun.com/competition/entrance/231718/introduction

参赛队伍：1960，比赛类型：推荐系统

- 比赛背景

今天许多中国互联网公司都在响应习近平主席一带一路的号召积极开拓海外市场。在我们开拓海外市场时往往会遭遇到用户习惯与国内不同造成国内的优秀策略难以在海外奏效等问题。即使成功开拓了某一国的市场，当需要进一步向更多国家扩张时，也会遇到不同国家的用户心智不统一的问题。

AliExpress 是中国最大出口 B2C 电商平台，2010 年平台成立至今已过 8 年，高速发展，日趋成熟。我们覆盖全球 230 个国家和地区，支持世界 18 种语言站点，22 个行业囊括日常消费类目，商品备受海外消费者欢迎；海外装机量超过 6 亿，入围全球应用榜单 TOP 10；目前的主要交易市场为俄、美、西、巴、法等国。

对于 AliExpress 来说，目前某些国家的用户群体比较成熟。这些成熟国家的用户在 AliExpress 尽享买买买之乐的同时，为我们沉淀了大量的该国用户的行为数据。这些沉淀下来的用户数据被我们挖掘利用后形成我们的推荐算法，用来更好的服务于该国用户。

但是还有一些待成熟国家的用户在 AliExpress 上的行为比较稀疏，对于这些国家用户的推荐算法如果单纯不加区分的使用全网用户的行为数据，可能会忽略这些国家用户的一些独特的心智；而如果只使用这些国家的用户的行为数据，由于数据过于稀疏，不具备统计意义，会难以训练出正确的模型。于是怎样利用已成熟国家的稠密用户数据和待成熟国家的稀疏用户数据训练出对于待成熟国家用户的正确模型对于我们更好的服务待成熟国家用户具有非常重要的意义。

本次比赛给出若干日内来自成熟国家的部分用户的行为数据，以及来自待成熟国家的 A 部分用户的行为数据，以及待成熟国家的 B 部分用户的行为数据去除每个用户的最后一条购买数据，让参赛人预测 B 部分用户的最后一条行为数据。。

- 评价指标

MRR(Mean Reciprocal Rank)：首先对选手提交的表格中的每个用户计算用户得分

$$score(buyer) = \sum_{k=1}^{30} \frac{s(buyer, k)}{k}$$

其中, 如果选手对该 buyer 的预测结果 predict k 命中该 buyer 的最后一条购买数据则 s(buyer，k)=1； 否则 s(buyer，k)=0。 而选手得分为所有这些 score(buyer)的平均值。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231718/information
- 比赛时间：2019 年 7 月 16 日 – 2019 年 9 月 16 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231718/rankingList
- 赛后分享

冠军： https://zhuanlan.zhihu.com/p/100827940

全球数据智能大赛【赛场二】

https://tianchi.aliyun.com/competition/entrance/231753/introduction

参赛队伍：719，比赛类型：数据挖掘、时间序列

- 比赛背景

全球数据智能大赛(2019)由广西壮族自治区人民政府主办，面向全球人工智能优秀团队征集多领域的算法智力成果和解决方案，集结众智，探索数字广西的蓝图，共领数据智能未来发展。大赛共设“数字人体”和“科技扶农”两个赛场，聚焦医疗和农业真实痛点。

广西是农业大省，由于地块相对破碎，种植结构复杂，农业统计调查工作量大。目前，广西已经实现了地面、卫星、雷达的气象数据的综合、立体监测。对农业产业而言，通过预测天气与作物生长的关系，为三农提供服务，对产业收割期意义重大。

近年来，人工智能在气象预测及农产品生长预测等领域取得了突飞猛进的成果。AI 技术的发展有望助力农产品的种植和生产，从而达到提升产量、防御灾害的目标。。

本次比赛给出若干日内来自成熟国家的部分用户的行为数据，以及来自待成熟国家的 A 部分用户的行为数据，以及待成熟国家的 B 部分用户的行为数据去除每个用户的最后一条购买数据，让参赛人预测 B 部分用户的最后一条行为数据。

算法赛要求选手以“科技扶农”水稻种植预测为课题，依据广西 81 个县早稻、晚稻作物 4 年产量的相关历史数据，结合 4 年气象降雨、温度、光照、温差等气象数据，一方面探索广西各地区气象环境局部特征，构建未来气象预测系统，另一方面挖掘气象和水稻产量的关系，构建因地制宜精准产量预测模型。

- 评价指标

选手提交结果与实际检测到的收率结果进行对比，以均方误差为评价指标，结果越小越好，均方误差计算公式如下：

$$f = \frac{1}{2m}\sum_{i=1}^{m}(y_i - y^*)^2$$

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231753/information
- 比赛时间

2019 年 9 月 4 日 – 2019 年 10 月 15 日

- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231753/rankingList
- 赛后分享

亚军： https://zhuanlan.zhihu.com/p/88895768

CIKM 2019 EComm AI：用户行为预测

https://tianchi.aliyun.com/competition/entrance/231719/introduction

参赛队伍：1623，比赛类型：推荐系统

- 比赛背景

在电商场景中，推荐系统作为电商核心功能之一，对用户体验的提升有重要作用。预测用户的兴趣，为其做出合理的推荐是工业界与学术界长久以来研究的课题。

经典方法包括基于内容的推荐、协同过滤等，一定程度上完成了推荐系统的任务。近年来，随着图神经网络研究的兴起，基于深度学习的 GNN(Graph Neural Network)在推荐领域也逐渐称为研究热点。

电商场景中， 用户，商品，以及两者之间的行为可以用一张二部图来表示。预测用户未来的行为，转化为预测二部图中用户-商品边的概率，有更好的可解释性、可推理性。

- 评价指标

召回@50: 对于特定的用户 i， 基于选手提交的 50 个预测商品字段 item_ids， 系统会根据用户 i 真实点击的商品计算分数。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231719/information
- 比赛时间：2019 年 6 月 12 日 – 2019 年 11 月 6 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231719/rankingList
- 赛后分享

冠军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.9.1faa194dSpScNG&postId=81672

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.3.1faa194dSpScNG&postId=81149

季军：https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.12.1faa194dSpScNG&postId=81501

CIKM 2019 EComm AI：超大规模推荐之用户兴趣高效检索

https://tianchi.aliyun.com/competition/entrance/231721/introduction

参赛队伍：1003，比赛类型：推荐系统

- 比赛背景

近年来，不管是学术界还是工业界，都在如何提升推荐系统中的用户兴趣预估精度方向上进行了很多的工作。然而，大规模的推荐系统由于受到了响应时间的约束，必须在模型的精度和复杂度上做出权衡。一个典型的大规模电商推荐系统，往往需要在很短的时间内，从千万级别的商品库 C 中为用户挑选出其最可能感兴趣的 k 个商品。对于这样规模的问题而言，使用兴趣预估模型来逐一地预测每个用户-商品对的兴趣度，再挑选出最靠前的 k 个商品的模式，由于计算效率太低而难以在实际系统中应用。

本次竞赛将聚焦在解决大规模推荐中用户兴趣检索的问题上，即，如何在避免穷举计算的情况下，高效地从全量商品库中精确地检索到用户最感兴趣的 k 个商品。

- 评价指标：使用 Recall@50 和 Novel-Recall@50 作为评价指标。
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231721/information
- 比赛时间：2019 年 6 月 12 日 – 2019 年 11 月 6 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231721/rankingList
- 赛后分享

冠军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.12.762d50596QySng&postId=81152

亚军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.3.762d50596QySng&postId=81487

季军：https://tianchi.aliyun.com/notebook-ai/detail?spm=5176.12586969.1002.6.762d50596QySng&postId=81648

2019 广东工业智造创新大赛【赛场一】

https://tianchi.aliyun.com/competition/entrance/231748/introduction

参赛队伍：2714，比赛类型：计算机视觉、目标检测

- 比赛背景

在布匹的实际生产过程中，由于各方面因素的影响，会产生污渍、破洞、毛粒等瑕疵，为保证产品质量，需要对布匹进行瑕疵检测。布匹疵点检验是纺织行业生产和质量管理的重要环节，目前人工检测易受主观因素影响，缺乏一致性；并且检测人员在强光下长时间工作对视力影响极大。由于布匹疵点种类繁多、形态变化多样、观察识别难道大，导致布匹疵点智能检测是困扰行业多年的技术瓶颈。

近年来，人工智能和计算机视觉等技术突飞猛进，在工业质检场景中也取得了不错的成果。纺织行业迫切希望借助最先进的技术，实现布匹疵点智能检测。革新质检流程，自动完成质检任务，降低对大量人工的依赖，减少漏检发生率，提高产品的质量。

本赛场聚焦布匹疵点智能检测，要求选手研究开发高效可靠的计算机视觉算法，提升布匹疵点检验的准确度，降低对大量人工的依赖，提升布样疵点质检的效果和效率。要求算法既要检测布匹是否包含疵点，又要给出疵点具体的位置和类别，既考察疵点检出能力、也考察疵点定位和分类能力。

- 评价指标：赛题分数计算方式:0.2ACC+0.8mAP

ACC：是有瑕疵或无瑕疵的分类指标，考察瑕疵检出能力。其中提交结果 name 字段中出现过的测试图片均认为有瑕疵，未出现的测试图片认为是无瑕疵。

mAP：参照 PASCALVOC 的评估标准计算瑕疵的 mAP 值。参考链接：https://github.com/rafaelpadilla/Object-Detection-Metrics。具体逻辑见 evaluator 文件。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231748/information
- 比赛时间：2019 年 8 月 19 日 – 2019 年 11 月 16 日

- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231748/rankingList
- 赛后分享

第一名：https://tianchi.aliyun.com/forum/postDetail?postId=83698

第二名：https://tianchi.aliyun.com/forum/postDetail?postId=83587

第三名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83602

第四名：https://tianchi.aliyun.com/forum/postDetail?postId=83838

第五名:https://tianchi.aliyun.com/forum/postDetail?postId=83791

第六名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83776

2019 广东工业智造创新大赛【赛场二】

https://tianchi.aliyun.com/competition/entrance/231749/introduction

参赛队伍：832，比赛类型：优化算法

- 比赛背景

本赛场聚焦面料剪裁利用率优化，要求选手研究开发高效可靠的算法，在较短时间范围内计算获得高质量可执行的排版结果，减少切割中形成的边角废料，提升面料切割利用率，减少计划时间、提高工作效率和避免人工计算的失误，提升价值降低成本。

在规则面料的情况下，满足零件旋转角度、零件最小间距、最小边距的约束，解决以下两类问题：

初赛赛题：基于所给零件，进行面料排版加工，耗料长度最短，面料利用率最高；

复赛赛题：在问题一的基础上，避开瑕疵区域面料加工，耗料长度最短，面料利用率最高。

- 评价指标

初赛（A 榜）:总分=（0.5*批次 1 面料利用率+0.5*批次 2 面料利用率）*100 总分=（0.5∗批次 1 面料利用率+0.5∗批次 2 面料利用率）∗100

复赛（B 榜）:总分=（0.5*批次 1 面料利用率+0.5*批次 2 面料利用率）*100 总分=（0.5∗批次 1 面料利用率+0.5∗批次 2 面料利用率）∗100

复赛（C 榜）:总分= 权重参数 1*面料利用率-权重参数 2*计算时间分值总分=权重参数 1∗面料利用率−权重参数 2∗计算时间分值

决赛总分=0.3*B 榜成绩+0.4*C 榜成绩+0.3*现场答辩决赛总分=0.3∗B 榜成绩+0.4∗C 榜成绩+0.3∗现场答辩

面料利用率=一个批次包含的零件总面积/消耗的面料总面积（消耗面料长度*面料宽度）；解释：用于衡量布匹原材料的利用情况，即使用长度越短、耗料越少的面料满足全部订单的生产，则切割利用率越高。

计算时间分值=f(一个批次排版的平均计算时间)；解释：计算时间越长，对应的计算时间分值呈阶梯的方式上升。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231749/information
- 比赛时间：2019 年 8 月 19 日 – 2019 年 11 月 16 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231749/rankingList
- 赛后分享

第一名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83654

第二名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83625

第三名：https://tianchi.aliyun.com/forum/postDetail?postId=83617

第四名：https://tianchi.aliyun.com/forum/postDetail?postId=83646

第五名: https://tianchi.aliyun.com/notebook-ai/detail?postId=83864

第七名：https://tianchi.aliyun.com/forum/postDetail?postId=83590

Apache Flink 极客挑战赛——垃圾图片分类

https://tianchi.aliyun.com/competition/entrance/231743/introduction

参赛队伍：2684，比赛类型：计算机视觉、图片分类

- 比赛背景

随着人工智能、移动互联网和物联网的兴起，大数据越变越大，也带来无限想象力和商业应用价值。深度学习是近十年来人工智能领域取得的最重要的突破之一。它在语音识别、自然语言处理、图像与视频分析等诸多领域都取得了巨大成功。

本次竞赛将聚焦在结合大数据计算引擎 Flink 和深度学习的计算平台 Intel Analytics Zoo 应用在图片识别场景，进行垃圾图片的分类。达摩院机器智能技术视觉实验室为本次竞赛提供垃圾图片数据集产出。

- 评价指标

系统给定 600 张图片，运行用户提供程序进行预测，根据预测结果进行打分。分数计算方法（500 毫秒内识别准确图片数/总图片数[600]）*100。

- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231743/information
- 比赛时间：2019 年 8 月 20 日 – 2019 年 11 月 12 日
- 比赛结果

Data Competition in 2019

奖项	获奖团队	团队成员
冠军	湖人总冠军	詹建文、李磊
亚军	SpacePro	刘冲、孙雪、管建宇
季军	SkyPeace	胡石生
优胜奖	小野兽队	孙甜、李欣、陈旭园
优胜奖	贪吃的小香猪-148	杨川、雷婷

- 赛后分享

第一名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83259

第二名：https://tianchi.aliyun.com/forum/postDetail?postId=83611

第三名：https://tianchi.aliyun.com/notebook-ai/detailpostId=83701

第四名：https://tianchi.aliyun.com/notebook-ai/detail?postId=83464

第五名：https://tianchi.aliyun.com/forum/postDetail?postId=83414

"合肥高新杯"心电人机智能大赛

https://tianchi.aliyun.com/competition/entrance/231754/introduction

参赛队伍：2353，比赛类型：表格数据

- 比赛背景

心电图是医院心脏疾病常用辅助诊断指标。心电图由于其价格低、无创的特性被广泛用于心脏疾病的预筛查以及体检中，每天的检测量巨大。目前，多导联的心电图设备已经广泛用于临床当中，部分设备已经具有自动分析诊断功能，但自动分析对于多心电异常事件的判别还不够精确，需要医生做进一步修改。

近年来，人工智能在心电图预测领域有了应用。AI 技术、深度学习的发展有望助力心电图波形、心电异常事件的预测，从而达到提升预测精度的目标。

本次大赛要求选手以心电图异常事件预测为赛题方向，依据心电图机 8 导联的数据，以及病患年龄、性别等因素，用统计学、机器学习、深度学习等方式探索挖掘心电波形与心电异常事件之间的关系，构建精准预测模型。

- 评价指标：F1
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231754/information
- 比赛时间：2019 年 9 月 5 日 – 2019 年 11 月 19 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231754/rankingList
- 赛后分享

第一名：https://tianchi.aliyun.com/notebook-ai/detail?postId=85958

第二名：https://tianchi.aliyun.com/notebook-ai/detail?postId=85999

第四名：https://tianchi.aliyun.com/forum/postDetail?postId=85990

第六名：https://tianchi.aliyun.com/notebook-ai/detail?postId=86093

安全 AI 挑战者计划第二期 - ImageNet 图像分类对抗攻击

https://tianchi.aliyun.com/competition/entrance/231761/introduction

参赛队伍：1000，比赛类型：计算机视觉、图片分类、对抗样本

- 比赛背景

图像分类（image classification）作为人工智能领域最典型的任务之一，是计算机视觉系统的基础，许多计算机视觉任务（如物体检测、图像分割等）都是在图像分类模型的基础上进行微调从而得到的。然而，基于机器学习的分类器对于对抗样本是高度敏感的。对抗样本是攻击者对原始输入样本添加微小的扰动，使得分类器发生预测错误的样本。攻击者通过对抗样本欺骗图像分类模型，会在很多实际系统中带来巨大的安全隐患。

此次比赛针对图像分类模型进行对抗攻击。为了模拟更加困难且真实的攻击场景，我们挑选了 3 个性能良好的防御模型作为被攻击模型。选手需要在不知道模型详细信息的前提下，构造有目标或者无目标的对抗样本。其中有目标的对抗样本需要被模型识别为特定类别；无目标的对抗样本需要被错分为任意类别，但是有目标和无目标攻击的难度不一样，所以设置的得分也不一样。选手在线下对测试样本进行修改，然后提交至线上做攻击测试。

- 评价指标：https://tianchi.aliyun.com/competition/entrance/231761/information
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231761/information
- 比赛时间：2019 年 12 月 6 日 – 2020 年 1 月 20 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231761/rankingList
- 赛后分享

1st：https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.9.3a678e84c52JnN&postId=87325

2nd: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12282029.0.0.1a3533669qfa3x&postId=87190

3rd: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.3.3a678e84c52JnN&postId=87278

4th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.15.3a678e84c52JnN&postId=87552

5th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.27.3a678e84c52JnN&postId=87389

6th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.18.3a678e84c52JnN&postId=87366

7th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.24.3a678e84c52JnN&postId=87411

8th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.6.3a678e84c52JnN&postId=87513

9th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.21.3a678e84c52JnN&postId=87436

10th: https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.30.3a678e84c52JnN&postId=87308

“数字人体”视觉挑战赛——宫颈癌风险智能诊断

https://tianchi.aliyun.com/competition/entrance/231757/introduction

参赛队伍：2358，比赛类型：计算机视觉、目标检测

- 比赛背景

大赛旨在通过提供大规模经过专业医师标注的宫颈癌液基薄层细胞检测数据，选手能够提出并综合运用目标检测、深度学习等方法对宫颈癌细胞学异常鳞状上皮细胞进行定位以及对宫颈癌细胞学图片分类，提高模型检测的速度和精度，辅助医生进行诊断。

- 评价指标：https://tianchi.aliyun.com/competition/entrance/231757/information
- 比赛数据：https://tianchi.aliyun.com/competition/entrance/231757/information
- 比赛时间：2019 年 10 月 24 日 – 2020 年 1 月 10 日
- 比赛结果：https://tianchi.aliyun.com/competition/entrance/231757/rankingList
- 赛后分享

baseline 方案（基于 RetinaNet）：
https://tianchi.aliyun.com/forum/postDetail?spm=5176.12586969.1002.6.76de2a3cqysMBD&postId=80827

1.3 DataFountain

智能盘点—钢筋数量 AI 识别

https://www.datafountain.cn/competitions/332

参赛队伍：1801，比赛类型：计算机视觉、物体检测

- 比赛背景

在工地现场，对于进场的钢筋车，验收人员需要对车上的钢筋进行现场人工点根，确认数量后钢筋车才能完成进场卸货。目前现场采用人工计数的方式，这种方式过程繁琐、消耗人力且速度很慢（一般一车钢筋需要半小时，一次进场盘点需数个小时）。针对上述问题，希望通过手机拍照->目标检测计数->人工修改少量误检的方式，智能、高效的完成此任务。

- 评价指标：F1-Score
- 比赛数据：https://www.datafountain.cn/competitions/332/datasets
- 时间轴：2019 年 1 月 10 日 – 2019 年 3 月 16 日
- 比赛结果：https://www.datafountain.cn/competitions/332/ranking?isRedance=0
- 赛后分享

海上风场 SCADA 数据缺失智能修复

https://www.datafountain.cn/competitions/333

参赛队伍：887，比赛类型：数据挖掘

- 比赛背景

由于风电场（尤其海上风电场）地处偏远，人工维护困难，远程数据监控系统（SCADA）能够远程获取风机运行状态数据，是风电场健康运行的保障。但是 SCADA 系统往往受到传感器失效、网络

阻塞等各种因素的影响，导致数据的缺失。我们希望通过大数据分析，利用已知数据对缺失的部分数据进行估计，尽量挽回由于数据缺失带来的损失。

我们抽取某一海上风电场实际 SCADA 数据，并人为地去除其中的部分数据，包括但不限于删去某个时间段的全部数据、某台机组在某段时间的数据、某台机组在某段时间的部分字段信息等等，参赛者需要利用剩余数据对删去的数据进行恢复，最终以恢复的准确度为评价基准。

- 评价指标

最终结果的评分采用如下计算公式：

$$F = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i} f_{i,j}(x_{i,j}, \hat{x}_{i,j})$$

- 比赛数据：https://www.datafountain.cn/competitions/333/datasets
- 时间轴：019 年 1 月 10 日 – 2019 年 3 月 16 日
- 比赛结果：https://www.datafountain.cn/competitions/333/ranking?isRedance=0
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/66410871

文化传承—汉字书法多场景识别

https://www.datafountain.cn/competitions/334

参赛队伍：1203，比赛类型：计算机视觉、预测分析

- 比赛背景

书法是汉字的书写艺术，是中华民族对人类审美的伟大贡献。从古至今，有大量照亮书法艺术星空的经典之作，是中华文明历经漫长岁月留下的艺术精华。这些书法作品现在仍以各种形式呈现给世

人：博物馆里的字画作品、旅游景点里的碑刻、建筑上的题词、对联、牌匾、甚至寻常家居里也会悬挂带有书法艺术的字画。在全球化、电子化的今天，书法的外部环境有了非常微妙的变化，对于年轻一代，古代书法字体越来越难以识别，一些由这些书法文字承载的传统文化无法顺利传承。 所以利用先进的技术，实时、准确、自动地识别出这些书法文字，对于记录整理书法艺术和传播书法背后的中国文化有着重要的社会价值。

利用人工智能技术，现在的中文识别已经有了很大的突破和极高的准确率。但是对于各种真实场景下（比如国画背景、建筑背景）的非楷书书法识别效果还不是很好。本次大赛希望征集到优秀的有创意的 AI 算法，解决这个问题。

- 评价指标：F1-score
- 比赛数据：https://www.datafountain.cn/competitions/334/datasets
- 时间轴：2019 年 1 月 22 日 – 2019 年 3 月 24 日
- 比赛结果：https://www.datafountain.cn/competitions/334/ranking?isRedance=0
- 赛后分享

大数据医疗—肝癌影像 AI 诊断

https://www.datafountain.cn/competitions/335

参赛队伍：1538，比赛类型：计算机视觉、医学图像处理

- 比赛背景

肝癌是病死率最高的恶性肿瘤之一，我国是肝癌第一大国，每年约有 38.3 万人死于肝癌，占全球肝癌死亡病例数的 51%，近年来肝癌的发病率还在逐渐增高。世卫组织预计，如不采取紧急行动提高诊疗可行性，2015 至 2030 年间中国将有约 1000 万人因肝硬化和肝癌死亡。CT 具有较高的分辨率，对肝癌的定位和定性诊断价值肯定，已成为常规检查项目，是一种安全、创伤较小的检查方法，诊断符合率可达到 90%以上，对肝癌的诊断及其程度的判断有重要的临床意义。

本赛题基于肝部腹腔强化 CT 断层扫描数据，以及相应的诊断结果。希望参赛者利用数据建模技术，构建基于医学影像的肝癌辅助诊断模型，利用人工智能手段对腹部 CT 影像进行诊断，判断患者良恶性，以帮助医生更加高效地对肝癌患者进行筛查。

- 评价指标：F1-score
- 比赛数据：https://www.datafountain.cn/competitions/335/datasets
- 时间轴：2019 年 1 月 22 日 – 2019 年 3 月 24 日
- 比赛结果：https://www.datafountain.cn/competitions/335/ranking?isRedance=0
- 赛后分享

混凝土泵车砼活塞故障预警

https://www.datafountain.cn/competitions/336

参赛队伍：964，比赛类型：机器学习、预测性维护

- 比赛背景

对生产设备的维护，传统的做法主要有两类，一种是等故障发生后再维修，但这会导致非计划性的停产，经济损失大；第二种是以固定计划进行维护，但维修成本高，停机时间长。预测性维护，则通过分析故障历史数据和实时监测数据，对设备关键部件的剩余寿命或故障进行提前预测预警，并据此进行维护维修，从而减少设备非计划停机时间、降低维护成本。

砼活塞是混凝土泵车的关键部件，也是消耗性部件，活塞故障将导致泵车无法正常工作，同时可能导致整个工地其他配套设备无法正常施工，从而带来相当大的经济损失。活塞寿命与设备的具体工况等密切相关，通过物联网将泵车的实时工况数据等上传至工业互联网云平台，基于积累的数据建立合适的模型，有望对砼活塞在未来一定工作任务期间内可能出现的故障做出有效的预测预警，从而提醒作业人员在施工前进行必要的维护，避免因计划外停机而带来的经济损失。

本赛题由中科云谷科技有限公司提供某类混凝土泵车砼活塞故障有关的数据，包括工作时间、发动机转速、油温、压力等多类工况数据，以及对应情况下，在未来完成给定工作量（混凝土泵送方量）的过程中，活塞是否故障的标识信息。希望参赛者利用大数据分析、机器学习、深度学习等方法，提取合适的特征、建立合适的故障预测模型，再根据测试数据预测该活塞在未来给定工作量内（泵送方量），是否会发生故障。

- 评价指标：Macro-F1-Score
- 比赛数据：https://www.datafountain.cn/competitions/336/datasets
- 时间轴：2019 年 1 月 22 日 – 2019 年 3 月 24 日
- 比赛结果：https://www.datafountain.cn/competitions/336/ranking?isRedance=0
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/66324559

消费者人群画像—信用智能评分

https://www.datafountain.cn/competitions/337

参赛队伍：2522，比赛类型：机器学习、数据挖掘

- 比赛背景

随着社会信用体系建设的深入推进，社会信用标准建设飞速发展，相关的标准相继发布，包括信用服务标准、信用数据采集和服务标准、信用修复标准、城市信用标准、行业信用标准等在内的多层次标准体系亟待出台，社会信用标准体系有望快速推进。社会各行业信用服务机构深度参与广告、政务、涉金融、共享单车、旅游、重大投资项目、教育、环保以及社会信用体系建设，社会信用体系建设是个系统工程，通讯运营商作为社会企业中不可缺少的部分 同样需要打造企业信用评分体系，助推整个社会的信用体系升级。同时国家也鼓励推进第三方信用服务机构与政府数据交换，以增强政府公共信用信息中心的核心竞争力。

传统的信用评分主要以客户消费能力等少数的维度来衡量，难以全面、客观、及时的反映客户的信用。中国移动作为通信运营商拥有海量、广泛、高质量、高时效的数据，如何基于丰富的大数据对客户进行智能评分是中国移动和新大陆科技集团目前攻关的难题。运营商信用智能评分体系的建立不仅能完善社会信用体系，同时也中国移动内部提供了丰富的应用价值，包括全球通客户服务品质的提升、客户欠费额度的信用控制、根据信用等级享受各类业务优惠等，希望通过本次建模比赛，征集优秀的模型体系，准确评估用户信用分值。

中国移动福建公司提供 2018 年 x 月份的样本数据（脱敏），包括客户的各类通信支出、欠费情况、出行情况、消费场所、社交、个人兴趣等丰富的多维度数据，参赛者通过分析建模，运用机器学习和深度学习算法，准确评估用户消费信用分值。

- 评价指标：MAE

平均绝对差值是用来衡量模型预测结果对标准结果的接近程度一种衡量方法。计算方法如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|pred_i - y_i|$$

其中 pred_i 为预测样本，y_i 为真实样本。MAE 的值越小，说明预测数据与真实数据越接近。最终结果为:

$$Score = \frac{1}{1+MAE}$$

- 比赛数据：https://www.datafountain.cn/competitions/337/datasets
- 时间轴：2019 年 1 月 22 日 – 2019 年 3 月 24 日
- 比赛结果：https://www.datafountain.cn/competitions/337/ranking?isRedance=0
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/65987372

基于虚拟仿真环境下的自动驾驶交通标志识别

https://www.datafountain.cn/competitions/339

参赛队伍：1250，比赛类型：计算机视觉、图像识别

- 比赛背景

随着汽车产业变革的推进，自动驾驶已经成为行业新方向。如今，无论是科技巨头还是汽车厂商都在加紧布局自动驾驶，如何保障研发优势、降低投入成本，从而加快实现自动驾驶汽车商业化成为了主要焦点。作为典型的自主式智能系统，自动驾驶是集人工智能、机器学习、控制理论和电子技术等多种技术学科交叉的产物。

虚拟仿真测试作为一种新兴测试方法，可快速提供实车路测难以企及的测试里程并模拟任意场景，凭借“低成本、高效率、高安全性”成为验证自动驾驶技术的关键环节，根据各传感器采集到的数据信息作出精准分析和智能决策，从而提高自动驾驶汽车行驶安全性，成为自动驾驶发展过程中不可或缺的技术支撑手段。天津卡达克数据有限公司在此背景下，积极应对产业变革，依托数据资源及相关产业背景，研发智能网联汽车仿真云平台，助推自动驾驶技术快速落地。

自动驾驶系统的环境感知能力是决定仿真结果准确性的重要因素之一，天津卡达克数据有限公司发布此赛题的目的旨在推动仿真环境下环境感知算法的科研水平。本题目以虚拟仿真环境下依托视频传感器数据进行交通标志检测与识别为例，希望在全球范围内发掘和培养自动驾驶算法技术人才。

在此任务中，我们将提供给参赛者一系列基于虚拟仿真环境下的自动驾驶视频图像，其中交通标志牌将作出标注。要求参赛者识别测试数据中随机出现的交通标志牌，并按照出现顺序反馈对应的识别结果。该虚拟仿真环境下伴随行人、非机动车等干扰因素，并具备多样性天气条件（包含光照）。

- 评价指标

本次竞赛根据提交结果计算 Score 作为评分标准；

（1）图像检测：

图像检测需要使用矩形框将目标检测物体选中，根据用户检测结果和目标框之间重叠比率大于 0.90，视为合格候选，预测的实例 A 和真实实例 B 之间的 IoU (Intersection over Union)的计算公式为：

$$IoU(A,B) = \frac{A \bigcap B}{A \bigcup B}$$

（2）图像识别：

判断图像内容是否匹配是根据图片中汽车道路标志牌名称与候选名称是否一致。

计算作品的召回率和正确率：

$$R（召回率） = \frac{检测正确的目标数量}{检测正确的目标数量 + 漏检的目标数量}$$

$$P（准确率） = \frac{检测正确的目标数量}{检测正确的目标数量 + 检测错误的目标数量}$$

计算 Score：

$$Score = \frac{2PR}{P + R}$$

- 比赛数据：https://www.datafountain.cn/competitions/339/datasets
- 时间轴：2019 年 4 月 19 日 – 2019 年 5 月 31 日
- 比赛结果：https://www.datafountain.cn/competitions/339/ranking?isRedance=0
- 赛后分享

基于 OCR 的身份证要素提取

https://www.datafountain.cn/competitions/346

参赛队伍：1684，比赛类型：计算机视觉、OCR

- 比赛背景

光学字符识别（OCR）技术在商业银行的影像数据解析中有着广泛应用,其中一个重要领域就是身份证影像识别。

身份证影像文件包含姓名、地址等多项个人基本信息，信息准确度和权威性高，在商业银行中被广泛应用于身份认证、信息采集等领域。

然而，商业银行的影像数据来源渠道复杂，时间跨度很大，质量层次不齐，目前市面上的身份证识别模型尚不能满足银行质量参差的影像识别需求。因此，一个具备强抗噪声干扰能力的 OCR 模型有着极高的商业价值。

以下列举两个实际应用中的挑战：

1. 图像质量参差：黑白复印件与彩色照片混杂，影像清晰度不尽相同，使得寻找具有普适性的图像处理手段和模型成为困难。

2. 文字重叠：商业银行为保护客户信息时常在保存影像件时叠加水印，尤其是深色的文字水印，例如“仅供 xxx 使用，复印无效”，这些水印与身份证上的文字重叠，给文字识别带来困难。

- 评价指标

本次比赛采用的评价方法为准确率(accuracy)

$$P（准确率）= \frac{检测正确的目标数量}{所有待检测位置的目标数量}$$

- 比赛数据：https://www.datafountain.cn/competitions/346/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/346/ranking?isRedance=0
- 赛后分享

冠军：https://discussion.datafountain.cn/questions/2260

线上 1-5 名：https://discussion.datafountain.cn/questions/2232

云计算时代的大数据查询分析优化

https://www.datafountain.cn/competitions/347

参赛队伍：561，比赛类型：性能优化、数据库

- 比赛背景

大数据时代，各行各业的数据不断爆炸性增长，数据分析需求和复杂度不断增加，如何提升数据分析的性能在学术界和工业界都受到很大的重视。在查询分析系统中，IO 和执行是两个最大的性能瓶颈，随着新硬件近几年的蓬勃发展，使用新硬件来解决这两个问题逐渐引起重视。使用 CPU 的新指令或 GPU、FPGA 等新硬件对数据进行处理并在执行层加速 SQL 的执行。本赛题旨在基于海量数据场景下，借助于 CPU 新指令或新硬件，提升复杂计算的效率。

- 评价指标

评测时将综合考虑输出的正确性，计算速度、成本等因素，原则上选手结果正确才可以得分。在结果正确的前提下，按计算耗费时间和消耗资源来综合排名打分。

- 比赛数据：https://www.datafountain.cn/competitions/347/datasets

Data Competition in 2019

- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/347/ranking?isRedance=0
- 赛后分享

线上 1-6 名：https://discussion.datafountain.cn/questions/2237

多人种人脸识别

https://www.datafountain.cn/competitions/348

参赛队伍：897，比赛类型：计算机视觉、人脸识别

- 比赛背景

人脸识别已经在生活中快速的普及开来， 但是人脸识别技术在实际应用中遇到的一个广为人知的问题是它在不同人种的性能有差异。 如何快速的提升人脸识别系统在不同人种的性能， 是一个实用的人脸识别算法应该考虑的问题。

本次比赛目标是提高人脸识别模型在不同人种上面的性能。以人脸 1：1 比对为场景， 参赛队需要同时优化人脸识别模型在不同人种上的性能，提高在低误识率情况下不同人种的通过率。

- 评价指标

根据提交结果画出不同人种的 ROC 曲线， 然后按照 FAR=1e-2, 1e-3 和 1e-4 时候对应的不同人种 TAR 的平均值进行结果排名。

- 比赛数据：https://www.datafountain.cn/competitions/348/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/348/ranking?isRedance=0
- 赛后分享

Data Competition in 2019

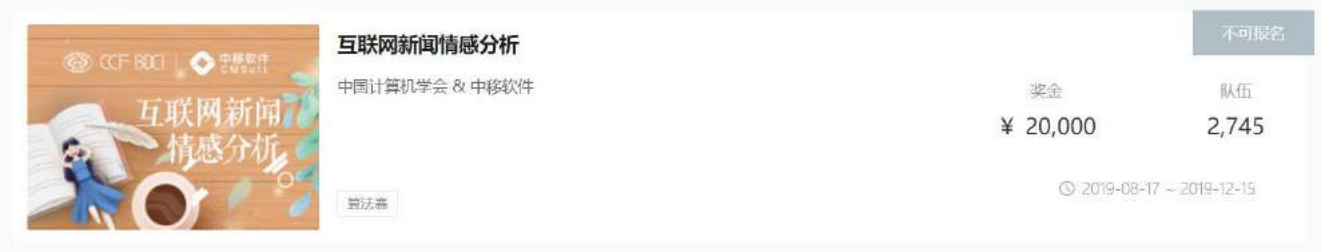
冠军：https://discussion.datafountain.cn/questions/2244/answers/23377

线上 1-5 名：https://discussion.datafountain.cn/questions/2225

互联网新闻情感分析

https://www.datafountain.cn/competitions/350

参赛队伍：2745，比赛类型：自然语言处理、情感分析



- 比赛背景

随着各种社交平台的兴起，网络上用户的生成内容越来越多，产生大量的文本信息，如新闻、微博、博客等，面对如此庞大且富有情绪表达的文本信息，完全可以考虑通过探索他们潜在的价值为人们服务。因此近年来情绪分析受到计算机语言学领域研究者们的密切关注，成为一项进本的热点研究任务。

本赛题目标为在庞大的数据集中精准的区分文本的情感极性，情感分为正中负三类。面对浩如烟海的新闻信息，精确识别蕴藏在其中的情感倾向，对舆情有效监控、预警及疏导，对舆情生态系统的良性发展有着重要的意义。

- 评价指标：Macro-F1 值
- 比赛数据：https://www.datafountain.cn/competitions/350/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/350/ranking?isRedance=0
- 赛后分享

线上 1-5 名：https://discussion.datafountain.cn/questions/2231

离散制造过程中典型工件的质量符合率预测

https://www.datafountain.cn/competitions/351

参赛队伍：2264，比赛类型：数据挖掘、分类预测

- 比赛背景

在高端制造领域，随着数字化转型的深入推进，越来越多的数据可以被用来分析和学习，进而实现制造过程中重要决策和控制环节的智能化，例如生产质量管理。从数据驱动的方法来看，生产质量管理通常需要完成质量影响因素挖掘及质量预测、质量控制优化等环节，本赛题将关注于第一个环节，基于对潜在的相关参数及历史生产数据的分析，完成质量相关因素的确认和最终质量符合率的预测。在实际生产中，该环节的结果将是后续控制优化的重要依据。

由于在实际生产中，同一组工艺参数设定下生产的工件会出现多种质检结果，所以我们针对各组工艺参数定义其质检标准符合率，即为该组工艺参数生产的工件的质检结果分别符合优、良、合格与不合格四类指标的比率。相比预测各个工件的质检结果，预测该质检标准符合率会更具有实际意义。本赛题要求参赛者对给定的工艺参数组合所生产工件的质检标准符合率进行预测。

- 评价指标：本次竞赛初赛评价指标使用 MAE 系数。

平均绝对差值是用来衡量模型预测结果对标准结果的接近程度一种衡量方法。计算方法如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|pred_i - y_i|$$

其中 pred_i 为预测样本，y_i 为真实样本。MAE 的值越小，说明预测数据与真实数据越接近。最终结果为:

$$Score = \frac{1}{1 + 10 * MAE}$$

- 比赛数据：https://www.datafountain.cn/competitions/351/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日

Data Competition in 2019

- 比赛结果：https://www.datafountain.cn/competitions/351/ranking?isRedance=0
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/100520412

线上 1-5 名：https://discussion.datafountain.cn/questions/2234

乘用车细分市场销量预测

https://www.datafountain.cn/competitions/352

参赛队伍：2999，比赛类型：数据挖掘、时间序列

- 比赛背景

近几年来，国内汽车市场由增量市场逐步进入存量市场阶段，2018 年整体市场销量首次同比下降。在市场整体趋势逐步改变的环境下，消费者购车决策的过程也正在从线下向线上转移，我们希望能在销量数据自身趋势规律的基础上，找到消费者在互联网上的行为数据与销量之间的相关性，为汽车行业带来更准确有效的销量趋势预测。

本赛题需要参赛队伍根据给出的 60 款车型在 22 个细分市场（省份）的销量连续 24 个月（从 2016 年 1 月至 2018 年 12 月）的销量数据，建立销量预测模型；

基于该模型预测同一款车型和相同细分市场在接下来一个季度连续 4 个月份的销量；除销量数据外，还提供同时期的用户互联网行为统计数据，包括：各细分市场每个车型名称的互联网搜索量数据；主流汽车垂直媒体用户活跃数据等。参赛队伍可同时使用这些非销量数据用于建模。

除了模型的准确性外，参赛队伍需对本赛题任务有系统性的思考和设计，在决赛阶段，参赛队伍对于所提交的模型的适应性、可扩展性、代码的工程性等方面也会影响参赛队伍的最终名次。

- 评价指标

初赛复赛阶段的在线评分采用 NRMSE（归一化均方根误差）的均值作为评估指标。首先单独计算每个车型在每个细分市场（省份）的 NRMSE，再计算所有 NRMSE 的均值。

$$NRMSE_k = \frac{RMSE_k}{\overline{y_k}} = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_{ki} - \widehat{y_{ki}})^2}{n_k}}}{\overline{y_k}}$$

$$Score = 1 - \frac{\sum NRMSE_k}{m}$$

- 比赛数据：https://www.datafountain.cn/competitions/352/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/352/ranking?isRedance=0
- 赛后分享

冠军：https://zhuanlan.zhihu.com/p/98926322

季军：https://zhuanlan.zhihu.com/p/98611487

线上 1-5 名：https://discussion.datafountain.cn/questions/2220

金融信息负面及主体判定

https://www.datafountain.cn/competitions/353

参赛队伍：1477，比赛类型：自然语言处理、情感识别

- 比赛背景

随着互联网的飞速进步和全球金融的高速发展，金融信息呈现爆炸式增长。如何从海量的金融文本中快速准确地挖掘出关键信息，成为了投资者和决策者重点考虑的问题之一。本赛题关注金融文本中的信息主体的挖掘和面向主体的负面消息检测，在风控和舆情分析等领域有很大现实意义。

该任务分为两个子任务：

给定一条金融文本和文本中出现的金融实体列表，

1. 负面信息判定：判定该文本是否包含金融实体的负面信息。如果该文本不包含负面信息，或者包含负面信息但负面信息未涉及到金融实体，则负面信息判定结果为 0。
2. 负面主体判定：如果任务 1 中包含金融实体的负面信息，继续判断负面信息的主体对象是实体列表中的哪些实体。

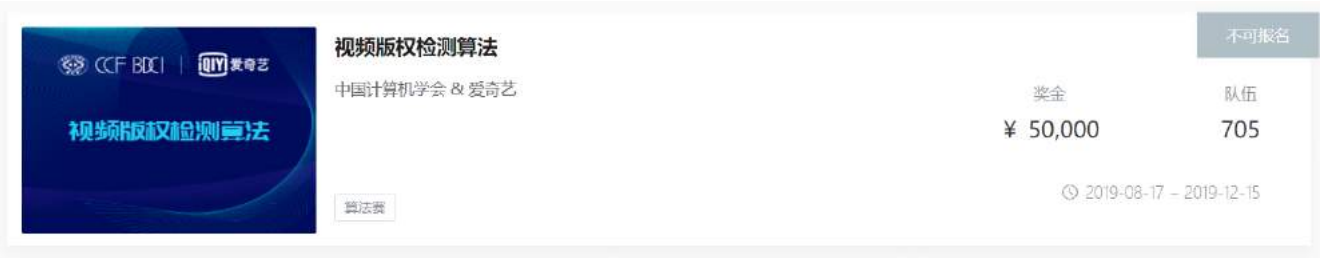
- 评价指标：https://www.datafountain.cn/competitions/353/datasets
- 比赛数据：https://www.datafountain.cn/competitions/353/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/353/ranking?isRedance=0
- 赛后分享

线上 1-5 名：https://discussion.datafountain.cn/questions/2233

视频版权检测算法

https://www.datafountain.cn/competitions/354

参赛队伍：705，比赛类型：计算机视觉、目标识别



- 比赛背景

随着移动互联网的发展和智能手机的普及，短视频已经成为重要的信息传播媒介，与此同时也带来了大量针对版权长视频的侵权行为。为了保护视频制作公司及原创者权益，需要通过自动化方式进行针对短视频的侵权行为检测。当前的侵权行为出现多样化及规模化特点，侵权视频多经过复合变换，要求算法模型中图像特征具有一定鲁棒性，并且有较快执行速度和并发能力。

Data Competition in 2019

本次竞赛将考察经过复合变换后的短视频关联到对应长视频的算法效果，其中不仅要找到短视频的原始长视频，还要计算出对应的时间段。过程中可能包括视频解码抽帧、视频或图像特征及指纹、视频相似检索等相关算法及技术方案。除了考察视频特征的鲁棒性外，也需要算法模型有一定的实时及并发能力。

- 评价指标

本模型依据提交的结果文件，采用 F1-score 进行评价。执行时间及特征索引大小将在复赛进行考察，初赛不进行相应限制和评分。

- 比赛数据：https://www.datafountain.cn/competitions/354/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/354/ranking?isRedance=0
- 赛后分享

线上 1-5 名：https://discussion.datafountain.cn/questions/2236

“技术需求”与“技术成果”项目之间关联度计算模型

https://www.datafountain.cn/competitions/359

参赛队伍：862，比赛类型：自然语言处理、关系挖掘

- 比赛背景

技术需求库和技术成果库的数据来源有两种：（1）会员单位发布；（2）非会员单位官方网站采集。每月新增数据量约 3000 个项目。根据项目信息的文本含义，为供需双方提供关联度较高的对应信息（需求——成果智能匹配服务），是平台的一项功能需求。

通过大赛发现好的方法、算法或模型，并提供用于验证的程序源代码，可应用于平台模拟人工，实现“需求——成果智能匹配服务”。

技术需求与技术成果之间的关联度分为四个层级：强相关、较强相关、弱相关、无相关。

人工判断技术需求和技术成果关联度的方法是：从事技术转移工作的专职工作人员，阅读技术需求文本和技术成果文本，根据个人经验予以标注。

- 评价指标

本次竞赛初赛评价指标使用 MAE 系数。平均绝对差值是用来衡量模型预测结果对标准结果的接近程度一种衡量方法。最终 MAE 加一后取倒数。

- 比赛数据：https://www.datafountain.cn/competitions/359/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/359/ranking?isRedance=0
- 赛后分享

冠军：https://discussion.datafountain.cn/questions/2247/answers/23373

互联网金融新实体发现

https://www.datafountain.cn/competitions/361

参赛队伍：2384，比赛类型：自然语言处理、命名实体识别

- 比赛背景

随着互联网的飞速进步和全球金融的高速发展，金融信息呈现爆炸式增长。投资者和决策者在面对浩瀚的互联网金融信息时常常苦于如何高效的获取需要关注的内容。针对这一问题，金融实体识别方案的建立将极大提高金融信息获取效率，从而更好的为金融领域相关机构和个人提供信息支撑。

从提供的金融文本中识别出现的未知金融实体，包括金融平台名、企业名、项目名称及产品名称。持有金融牌照的银行、证券、保险、基金等机构、知名的互联网企业如腾讯、淘宝、京东等和训练集中出现的实体认为是已知实体。

- 评价指标

考察未知实体的识别，对参赛队提交的未知实体识别结果（unknownEntities）基于未知实体集合进行 Micro-Averaging 评测并给出评测分数 MicroF。实体的全称和缩写视为不同实体。

- 比赛数据：https://www.datafountain.cn/competitions/361/datasets
- 时间轴：2019 年 8 月 17 日 – 2019 年 11 月 18 日
- 比赛结果：https://www.datafountain.cn/competitions/361/ranking?isRedance=0
- 赛后分享

线上 1-5 名：https://discussion.datafountain.cn/questions/2226

人工识云赛道-识云竞答

https://www.datafountain.cn/competitions/355

参赛队伍：310，比赛类型：计算机视觉、图像识别

- 比赛背景

“观云识天”人机对抗大赛旨在贯彻落实习近平总书记对人工智能工作重要指示精神、《国务院关于印发新一代人工智能发展规划的通知》文件精神及 2019 年全国气象局长会刘雅鸣局长工作报告中对智能观测所提出的具体要求，努力打造气象部门首届面向全国的高影响力气象+人工智能赛事，激发社会各界发挥创意，将人工智能技术落地到气象观测、预报及服务领域中去。并以此为契机，形成气象+人工智能常态化赛事，搭建交流合作平台。

作为首届人工智能技术与气象业务深度融合的高规格赛事，大赛集中展现新时代气象观测领域新技术、新应用、新模式，主要突出专业化、智能化、应用化三大特点，以众智、众包的创新方式，汇聚政产学研用多方智慧，面向全球技术人才开放报名，助力气象领域专业化人才发展及行业生态良性建设。

在“卦天气象”APP 中报名并学习人工识别云状知识，参与竞赛答题。

- 评价指标：https://www.datafountain.cn/competitions/355/datasets
- 比赛数据：https://www.datafountain.cn/competitions/355/datasets
- 时间轴：2019 年 9 月 05 日 – 2019 年 11 月 10 日
- 比赛结果： https://www.datafountain.cn/competitions/355/ranking?isRedance=0
- 赛后分享

机器图像算法赛道-天气识别

https://www.datafountain.cn/competitions/356

参赛队伍：1054，比赛类型：计算机视觉、图像识别

- 比赛背景

根据大赛组织方提供的图片数据训练算法，能够区分降雨、降雪、冰雹、露、霜、雾（霾）、雾凇、雨凇、电线积冰 9 种天气现象。

- 评价指标：初赛期间作品采用宏平均 F1-score 进行评价。
- 比赛数据：https://www.datafountain.cn/competitions/356/datasets
- 时间轴：2019 年 9 月 05 日 – 2019 年 11 月 10 日
- 比赛结果：https://www.datafountain.cn/competitions/356/ranking?isRedance=0
- 赛后分享

机器图像算法赛道-云状识别

https://www.datafountain.cn/competitions/357

Data Competition in 2019

参赛队伍：1318，比赛类型：计算机视觉、图像识别

- 比赛背景：根据大赛组织方提供的图片数据训练算法，识别 3 族 10 属 29 类云状。
- 评价指标：初赛期间作品采用宏平均 F1-score 进行评价。
- 比赛数据：https://www.datafountain.cn/competitions/357/datasets
- 时间轴：2019 年 9 月 05 日 – 2019 年 11 月 10 日
- 比赛结果：https://www.datafountain.cn/competitions/357/ranking?isRedance=0
- 赛后分享

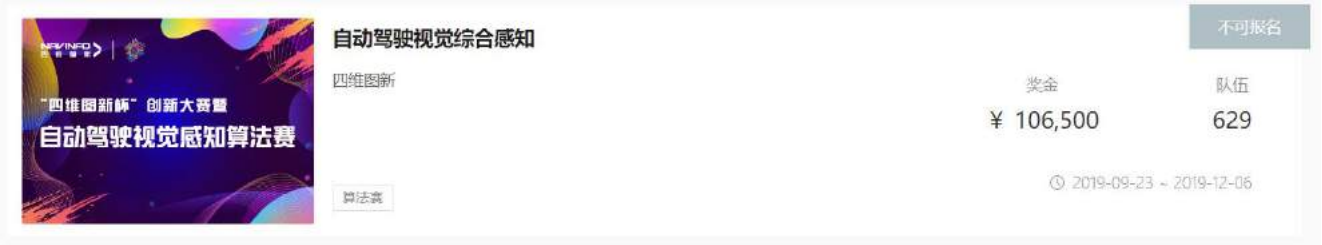
CV_七少：https://discussion.datafountain.cn/questions/2152/answers/23219

我有翅膀：https://discussion.datafountain.cn/questions/2153/answers/23231

自动驾驶视觉综合感知

https://www.datafountain.cn/competitions/366

参赛队伍：629，比赛类型：计算机视觉、语义分割、物体检测



- 比赛背景

自动驾驶这颗科技明珠是当前科研和产业界共同面对的难题，环境感知技术是自动驾驶众多关键技术之一，能够适应各种场景的成熟感知技术也是自动驾驶能够大规模落地的必经之路。 本次赛题旨在推动日新月异的计算机视觉和机器学习算法领域接轨自动驾驶面临的实际问题。为此我们选取了

基于视觉图像的目标检测和场景分割的综合任务，并开放了一个具备精细标注的大规模数据集，希望参赛者为自动驾驶开发出新颖独特的融合算法和框架。

此次开放的数据集来源为测绘级高精度双目，我们希望行业内首次开放的 4K 级高精度数据可以让自动驾驶相关的应用都有所受益，包括但不限于 2D/3D 场景解析、定位、迁移学习和驾驶仿真。针对本次赛题开放了近一万帧针对不同场景下自动驾驶视觉任务的训练数据，包含移动物体、信号灯、交通标牌的检测以及车道线、可行驶区域的分割精细标注数据。要求参赛者提供的决赛方案可以在 NVIDIA 1080 GPU 或相同算力设备上运行达到 15fps（初赛不考虑速度，如果决赛方案创新性很强，也可适当放宽实时性要求）。

- 评价指标

我们使用 MAP 和 MeanIOU 来分别评估目标检测和实例分割的精度，并将两个结果的分数累加。在目标检测任务中，我们使用 MAP 作为检测评估标准。如果参赛者预测的结果和真实目标框之间的重叠比率，即交并比（IOU）大于 0.75，视为被匹配上，如果一个真实目标框匹配了多个预测的结果，则匹配上了的并具备最大置信度的预测实例被选为正样例，剩下的匹配上的预测实例被归为负样例。预测实例如果没有被匹配到任何真实实例，则被归为负样例。预测的检测框 A 和真实目标框 B 之间的 IOU 的计算公式为：

$$IOU\,(A,\ B) = \frac{A \bigcap B}{A \bigcup B}$$

我们使用 MeanIOU 作为实例分割的评估标准，我们对参赛者提交的每一个 test 预测结果分别计算预测的车道线、可行驶区域和灯杆灯柱和真实实例之间的交并比（IOU），并计算平均值。最终的 SCORE = (MAP+MIOU) / 2

- 比赛数据：https://www.datafountain.cn/competitions/366/datasets
- 时间轴：2019 年 9 月 23 日 – 2019 年 12 月 06 日
- 比赛结果：https://www.datafountain.cn/competitions/366/ranking?isRedance=0
- 赛后分享

汽车论坛消费者用车体验内容的判别与标注

https://www.datafountain.cn/competitions/365

参赛队伍：837，比赛类型：自然语言处理、判别与标注

- 比赛背景

国内某知名汽车厂商的商务车业务部门，除了从传统渠道如 4S 店等获得汽车质量与用车反馈信息，还需要从互联网上的网友发布内容获得自有与竞品车型的汽车消费者关于用车体验与质量反馈信息。海量信息受该部门委托，针对互联网中各大汽车论坛中相关车型论坛的网友发帖内容进行数据分析，根据帖子内容，首先判断是否为真实消费者发布,其次判断是否为用车体验与质量反馈内容，如果是用车体验与质量反馈内容，则进一步标注出汽车的关于用车的哪个方面，以及具体哪个零部件的哪类质量问题。通过这一系列内容判别与标注处理，业务部门可以通过互联网渠道反馈，掌握自家各车型与竞品车型的消费者用车评价分类分布情况，进而指导产品设计与生产部门进行设计改进与品控工作，以及给市场部门提供宣传基础数据素材。希望实现算法自动准确的判别与标注，能够节省人工标注成本，提升数据分析处理效率，提升消费者用车评价反馈到厂家的及时性。

本题目所涉及数据均为汽车论坛网站中，指定车型论坛版面下的帖子内容。需要算法分析处理的任务：

任务：对帖子内容进行真实汽车消费者发布的用车体验与质量问题内容，以及其它内容的分类判别

对于真实汽车消费者发布的内容，界定标准为：需要发帖人描述的现象为自己的亲身感受的用车体验与质量反馈问题，其中质量问题部分排除非操作不当、非事故造成的质量问题。

该部分难点为需要从内容整体语义，区分是否为真实用车人所发布帖子，尤其需要从描述内容的主语部分入手判别是否为真实车主。其次在判别是否为真实用车体验与质量故障方面，语义处理难度也比较大，需要算法设计时重点关注。

- 评价指标：评分算法采用 Macro-F1-Score 的计算方式。
- 比赛数据：https://www.datafountain.cn/competitions/365/datasets
- 时间轴：2019 年 8 月 31 日 – 2019 年 11 月 24 日

Data Competition in 2019

- 比赛结果：https://www.datafountain.cn/competitions/365/ranking?isRedance=0
- 赛后分享

Baseline：https://zhuanlan.zhihu.com/p/93359772

无人集群空地协同攻防对抗挑战赛

https://www.datafountain.cn/competitions/365

参赛队伍：403，比赛类型：创意创新

- 比赛背景

无人集群具备通过个体间的有效协作涌现出高于个体的群体智能，完成复杂环境下作战任务的能力，是军事装备体系智能化的驱动引擎，也将颠覆未来的作战形态。如何基于群体智能技术进行分布式协同感知及识别、协同认知与决策、协同制导与控制，是提高无人集群分布式协同作战效能的关键所在，也是新一代人工智能的核心研究领域。

基于虚拟的异构无人集群攻防对抗仿真平台，攻防双方在生成的地图上拥有各自的阵地，参赛队开发群体智能协同算法，采用一定数量的无人机和无人车相互配合，协同搜索、识别和摧毁对方阵地内的静止和移动目标，同时协同拦截对方的无人机和无人车，保护己方的指挥所不被摧毁。

其中，无人机具备协同探测和多机协同摧毁对方无人机的能力；无人车具协同探测和摧毁地面目标的能力；两者之间可以互相通信和协同。该任务综合考量协同决策控制能力。

- 评价指标：https://www.datafountain.cn/competitions/373/datasets
- 比赛数据：https://www.datafountain.cn/competitions/373/datasets
- 时间轴

2019 年 10 月 17 日 – 2019 年 11 月 25 日

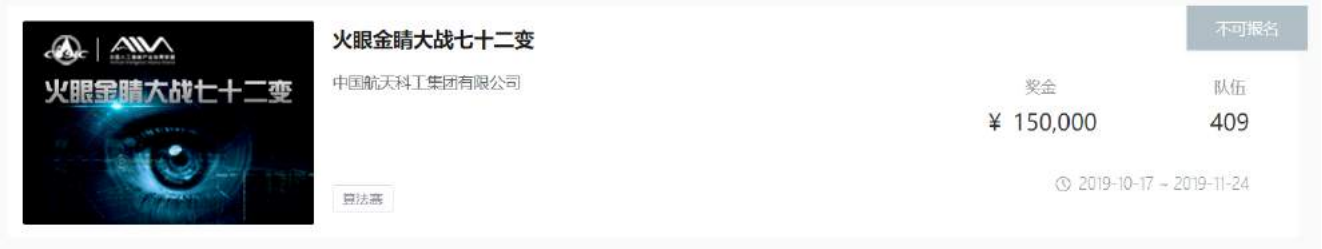
- 比赛结果：https://www.datafountain.cn/competitions/373/ranking?isRedance=0

- 赛后分享

火眼金睛大战七十二变

https://www.datafountain.cn/competitions/372

参赛队伍：409，比赛类型：计算机视觉、对抗样本



- 比赛背景

当今人工智能技术不断引领计算机视觉向前发展，但伴随着对人工智能安全性考虑的不断深入，智能识别也不完全是火眼金睛。未来，如何在对抗攻击条件下解决智能识别模型安全问题，是人工智能技术能够普及的关键。

针对图像分类问题，主办方提供共用数据集，供“火眼金睛“（守）方和”七十二变“（攻）方使用。参赛选手可自行选定阵营。围绕着智能识别的攻与守，攻方通过对图像进行添加噪声等轻微扰动，生成“七十二变”的对抗样本，使人工智能模型无法准确分类识别；守方构建更加鲁棒的识别模型，练就“火眼金睛“，以准确识别对抗样本。

- 评价指标：https://www.datafountain.cn/competitions/372/datasets
- 比赛数据：https://www.datafountain.cn/competitions/372/datasets
- 时间轴：2019 年 10 月 17 日 – 2019 年 11 月 24 日
- 比赛结果：https://www.datafountain.cn/competitions/372/ranking?isRedance=0
- 赛后分享

1.4 和鲸

2018 GAMMA 智能营销科技大赛

https://www.kesci.com/home/competition/5c063451b61b67001076bd0b

Data Competition in 2019

参赛人数：队伍数 252 人数 389

比赛类型：团队创新赛，数据类型：结构化

2018 GAMMA智能营销科技大赛 ¥ 120,000

本次智能营销科技大赛由平安金融壹账通联合科赛举办，致力于发掘 AI 时代最有创意和趣味的AI 营销产品方案。大赛围绕将已有的 AI 模型应用到生活中的具体营销场景，让技术真正落地。不论你是 AI 技术人才还是营销产品人才，都可以来一决高下！

参赛人数 389 参赛团队 252 2018/12/06 - 2019/02/22

- 比赛背景：

「2018 GAMMA 智能营销科技大赛」由「金融壹账通」联合科赛举办，金融壹账通作为平安集团旗下的创新式综合金融理财平台，专注于为中小金融机构科技赋能，是一家全球领先的金融科技服务公司。

公司旨在以世界顶级的金融科技组合为中小金融机构提供领先的经营模块解决方案，构建庞大的金融科技服务生态圈。 基于人工智能、大数据、区块链、云平台以及金融应用等五大核心科技，结合平安深耕多年并经实践验证的专业技术，金融壹账通打造了智能销售方案、智能风控方案、智能产品方案、智能服务方案和智能运营方案五大利器，帮助金融机构全面提升获客、风控、产品、客服、运营五大能力，实现经营管理水平与收入的快速升级。

- 主办方: 金融壹账通
- 评价指标：创新、趣味、市场和商业性等；
- 比赛数据：
- 时间轴：2018/12/06 - 2019/02/22
- 比赛结果：

  https://www.kesci.com/home/competition/5c063451b61b67001076bd0b/content/12
- 赛后分享：

“默克”杯逆合成反应预测大赛

https://www.kesci.com/home/competition/5c35b0aa4ea711002cafcaa6

参赛人数：队伍数 370 人数 710

“默克”杯逆合成反应预测大赛 ￥50,000 + ￥20,000 GPU资源

近年来，AI技术不断参与化学、制药领域，但主要是被用来预测反应物。这次我们反过来，用反应物预测生成物，不仅是复杂程度、挑战性提高了，它的意义也上升了。而在未来，AI将成为科学家的定制工具，对于提高药物研发的速度和效率、降低成本等，都有着巨大的益处。 这次，我们把海量的化学反应方程式交给计算机去学习，让计算机进行逆合成反应预测。

参赛人数 710 参赛团队 370 2019/01/21 - 2019/04/12

- 比赛背景：

近年来，AI 技术不断参与化学、制药领域，但主要是被用来预测反应物。然而逆合成分析作为有机化学的基石，在新药研发中也占有十分重要的地位。

传统化学研究中，化学家们完成逆合成反应预测耗时耗力：需先从目标产物的分子式开始分析，再利用 Scifinder 搜索相似的结构和文献报道过的合成路径，确认需要哪些试剂、怎样的反应序列，甚至要依据直觉制定几十个化学反应，通过这些反应逐步生成目标产物，才能开始实验。这个过程往往会浪费化学家几天甚至更长时间。

本次大赛由默克集团（Merck KGaA）旗下默克生命科学主办，和鲸科技（前身科赛）协办，将海量的化学反应方程式交给计算机去学习，让计算机进行逆合成反应预测。逆合成反应预测，不仅提高了预测的复杂程度和挑战性，也强调了它在新药研发中的意义。而在未来，AI 将成为科学家的定制工具，对于提高药物研发的速度和效率、降低成本等，都有着巨大的益处。

- 主办方: 默克集团
- 评价指标：`F1_score`和`exact_match_score`两个指标加权，作为最后的评测指标
- 比赛数据：https://www.kesci.com/home/competition/5c35b0aa4ea711002cafcaa6/content/4
- 时间轴：2018/12/06 - 2019/02/22
- 比赛结果：

  https://www.kesci.com/home/competition/5c35b0aa4ea711002cafcaa6/content/9
- 赛后分享：

2019 中国高校计算机大赛——大数据挑战赛

https://www.kesci.com/home/competition/5cc51043f71088002c5b8840

参赛人数：队伍数 888 人数 2092

【正式赛】2019中国高校计算机大赛——大数据挑战赛 ¥ 300,000

中国高校计算机大赛是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，面向高校学生的高水平计算机类系列竞赛，其中大数据挑战赛被列入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。 2019大数据挑战赛（以下简称"大赛"）是在中国高校计算机大赛主办单位的指导下，由清华大学、南开大学与字节跳动公司联合主办，由亚马逊AWS提供资源支持以及科赛提供竞赛平台支持，并以企业真实场景和实际数据为基础的高端算法竞赛。大赛面向全球高校在校生开放，旨在提升高校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校教师参与指导。

参赛人数 2092　参赛团队 888　2019/05/26 - 2019/08/11

- 比赛背景：

中国高校计算机大赛是由教育部高等学校计算机类专业教学指导委员会、教育部高等学校软件工程专业教学指导委员会、教育部高等学校大学计算机课程教学指导委员会和全国高等学校计算机教育研究会联合主办，面向高校学生的高水平计算机类系列竞赛，其中大数据挑战赛被列入全国普通高校学科竞赛排行榜，获得社会各界的高度关注和广泛好评。

2019 大数据挑战赛（以下简称“大赛”）是在中国高校计算机大赛主办单位的指导下，由清华大学、南开大学与字节跳动公司联合主办，亚马逊 AWS 提供资源支持以及科赛提供竞赛平台支持，并以企业真实场景和实际数据为基础的高端算法竞赛。大赛面向全球高校在校生开放，旨在提升高校学生对数据分析与处理的算法研究与技术应用能力，探索大数据的核心科学与技术问题，尝试创新大数据技术，推动大数据的产学研用，本次大赛鼓励高校教师参与指导。

- 主办方: 清华大学、南开大学与字节跳动公司
- 评价指标：Mean AUC
- 比赛数据：https://www.kesci.com/home/competition/5cc51043f71088002c5b8840/content/1
- 时间轴：2019/05/26 - 2019/08/11
- 比赛结果：

  https://www.kesci.com/home/competition/5cc51043f71088002c5b8840/leaderboard

- 赛后分享：

第二名：https://www.kesci.com/home/competition/forum/5d6e5a00ecd45c001dcb6f56

第三名:https://www.kesci.com/home/competition/forum/5d664430326504000f1dfa04

Data Competition in 2019

第四名:https://www.kesci.com/home/competition/forum/5d6627be28c55d000f24a36a

第五名：https://www.kesci.com/home/competition/forum/5d674531326504000f1dfa6c

贵在联通——“联创黔线”杯大数据应用创新大赛

https://www.kesci.com/home/competition/5be92233954d6e001063649a/content

参赛人数：队伍数 205 人数 359

贵在联通——“联创黔线”杯大数据应用创新大赛 ¥ 100,000

本次大赛由中国联通贵阳大数据创新创业中心主办，期待通过本次赛事发掘一批具有创新精神和数据解决方案能力的人才，以及具备市场发展前景和实用性价值的大数据产品。中国联通贵阳大数据创新创业中心成立于“中国数谷”——贵阳，是由贵阳市人民政府与联通创新创业投资有限公司共同发起成立的以大数据为核心的创新创业孵化和产业投资机构。

参赛人数 359 参赛团队 205 2019/06/11 - 2019/08/07

- 比赛背景：

中国联通贵阳大数据创新创业中心成立于“中国数谷”——贵阳，是由贵阳市人民政府与联通创新创业投资有限公司共同发起成立的以大数据为核心的创新创业孵化和产业投资机构。我中心在贵州省省委、省政府的政策引导下，抢占大数据发展先机，通过贵州联通数据资源推动大数据应用发展，并与省气象局进行深度沟通与合作，推进联通大数据与气象大数据相结合，全面实施大数据发展战略行动，加快推动数据资源共享和开放，助力产业转型升级和社会治理创新，实现大数据产业加速拓展。

我们期待通过本次赛事发掘一批具有创新精神和数据解决方案能力的人才，以及具备市场发展前景和实用性价值的大数据产品。对于通过 2019 “联创黔线杯”大数据应用创新大赛发掘的大数据领域优质项目和团队，双创中心将结合中国联通在大数据领域的资源优势，对极具价值的项目进行孵化和投资；同时本赛事作为省科技厅主办的第八届中国创新创业大赛（贵州赛区）赛事系列活动之一，“联创黔线杯”大赛中的优质项目和团队也将获得该赛事组委会认可，我们也会将 “联创黔线杯”大赛中发现的贵州省内大数据领域优质项目和团队，推荐参加第八届中国创新创业大赛（贵州赛区）赛事。

- 主办方: 中国联通贵阳大数据创新创业中心
- 评价指标：AUC

Data Competition in 2019

- 比赛数据：https://www.kesci.com/home/competition/5be92233954d6e001063649a/content/1
- 时间轴：2019/06/11 - 2019/08/07
- 比赛结果：

  https://www.kesci.com/home/competition/5be92233954d6e001063649a/content/10
- 赛后分享：

第 10 名：https://www.kesci.com/home/competition/forum/5d7785c3ecd45c001dcb731e

第 13 名：https://www.kesci.com/home/competition/forum/5d776548080a0f001df6dc99

第 15 名：https://www.kesci.com/home/competition/forum/5d679188080a0f001df6d61a

莱斯杯：全国第二届“军事智能机器阅读”挑战赛

https://www.kesci.com/home/competition/5d142d8cbb14e6002c04e14a

参赛人数：队伍数 625 人数 850

- 比赛背景：

由军委装备发展部和中国电子科技集团有限公司指导，军委装备发展部中国电科联合基金资助，中国电科 X+AI 系列挑战赛之"莱斯杯” 全国第二届“军事智能•机器阅读”挑战赛即将正式开赛，该赛事将于 2019 年 7 月 2 日开启正式线上报名，诚邀国内学术界和工业界的研究者与开发者积极参与，共同挑战机器阅读能力极限，争夺军事智能王者桂冠。

“莱斯杯”是军事领域顶尖的人工智能和大数据技术竞赛平台，旨在面向实际需求，汇聚产学研各界精英，共同推进智能算法在军事垂直领域的落地应用。2018 年，“莱斯杯”全国首届“军事智能•机器阅读”挑战赛成功举办。大赛围绕机器阅读理解这一“AI 的终极任务”，引起了学术界和科技界的广泛关注，近千人报名参赛，冠军选手成绩已达国际领先水平，比赛部分成果已在项目中实际转化，推进了自然语言处理技术在军事领域的应用和进步。

本届“莱斯杯”秉承“算法顶天、应用立地”的理念，在赛题设置上面向用户关键信息需求，以文字情报整编业务面临的实际痛点问题为切入点，旨在由机器筛选、整编出多篇文字报中用户所关心的活动时间、地点、频率、性能参数等中心内容。相比上一届“莱斯杯”，赛题在更加贴合实际应用同时，也将带来 NLP 算法的全新挑战，属当下最前沿热点研究领域之一。一旦成功应用，将颠覆以往情报整编工作以人工分析提炼为主的模式，逐步实现由机器替代繁琐、重复性情报整编业务流程。

- 主办方: 中电莱斯信息系统有限公司
- 评价指标：Rouge-L
- 比赛数据：https://www.kesci.com/home/competition/5d142d8cbb14e6002c04e14a/content/5
- 时间轴：2019/06/11 - 2019/08/07
- 比赛结果：
- https://www.kesci.com/home/competition/5d142d8cbb14e6002c04e14a/leaderboard
- 赛后分享：

第一名：https://www.kesci.com/home/project/5dbbec9f080dc300371eda5d

第二名：https://www.kesci.com/home/project/5dbd09b3080dc300371f056f

第四名：https://www.bilibili.com/video/av74006222?from=search&seid=14864289781361104410M

第六名：https://github.com/caishiqing/joint-mrc

首届“全国人工智能大赛”（行人重识别 Person ReID 赛项）

https://www.kesci.com/home/competition/5d90401cd8fc4f002da8e7be

参赛人数：队伍数 2652 人数 1935

- 比赛背景：

2019 年 8 月，深圳市人民政府决定专门设立人工智能领域权威赛事——全国人工智能大赛（以下简称大赛）。大赛将立足国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展。大赛由深圳市人民政府主办，深圳市科创委、鹏城实验室及科技部指导成立的新一代人工智能产业技术创新战略联盟（AITISA，以下简称“联盟”）共同承办，腾讯科技等协办。

大赛以“AI 赋能视界”为主题，总奖金达到了 536 万。分为“AI+4K HDR”和“行人重识别”两个赛项，每个赛项的总奖金池金额 268 万，其中一等奖高达 100 万，无疑是今年国内总奖金和一等奖金额最高的人工智能赛事。。

- 主办方: 深圳市人民政府
- 评价指标：rank1 与 mAP@200
- 比赛数据：https://www.kesci.com/home/competition/5d90401cd8fc4f002da8e7be/content/2
- 时间轴：2019/10/17 -
- 比赛结果：

  https://www.kesci.com/home/competition/5d90401cd8fc4f002da8e7be/leaderboard
- 赛后分享：

首届“全国人工智能大赛”（AI+4K HDR 赛项）

https://www.kesci.com/home/competition/5d84728ab1468c002ca1825a

参赛人数：队伍数 1602 人数 1082

- 比赛背景：

2019 年 8 月，深圳市人民政府决定专门设立人工智能领域权威赛事——全国人工智能大赛（以下简称大赛）。大赛将立足国际视野，营造人工智能创新创造氛围，促进产业、学术、资本、人才等创新要素融合发展。大赛由深圳市人民政府主办，深圳市科创委、鹏城实验室及科技部指导成立的新一代人工智能产业技术创新战略联盟（AITISA，以下简称“联盟”）共同承办，腾讯科技等协办。

大赛以“AI 赋能视界”为主题，总奖金达到了 536 万。分为“AI+4K HDR”和“行人重识别”两个赛项，每个赛项的总奖金池金额 268 万，其中一等奖高达 100 万，无疑是今年国内总奖金和一等奖金额最高的人工智能赛事。。

- 主办方: 深圳市人民政府
- 评价指标：25 * PSNR 项 + 25 * SSIM 项 + 50 * VMAF 项
- 比赛数据：https://www.kesci.com/home/competition/5d84728ab1468c002ca1825a/content/2
- 时间轴：2019/10/17 -
- 比赛结果：

  https://www.kesci.com/home/competition/5d84728ab1468c002ca1825a/leaderboard

- 赛后分享：

1.5 DataCastle

地球物候的深度学习预测

https://www.pkbigdata.com/common/cmpt/%E5%9C%B0%E7%90%83%E7%89%A9%E5%80%99%E7%9A%84%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0%E9%A2%84%E6%B5%8B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：708 支队伍 • 777 名参赛者

比赛类型：团队算法赛，数据类型：图像

- 比赛背景：

经过数十亿年的进化，地球的生物圈适应了这颗行星的春夏秋冬。各地植物的生长节律与季节同步，春华秋实，这被称为植被的物候。然而，每一年的季节变换都不完全一样。有些年份的气温和雨水和其他年份都很不一样。这会给这颗行星的农林牧业带来一些波动，让植被的生长节奏变动：比如说干热带来的山火肆虐，比如说春天提前导致的生长季变长，比如说降水增大带来的洪涝，等等。如果我们能够提前预测各地未来的物候，那么就可以趋利避害。

然而，传统上的气候/物候学预测能力相当有限。但是在今天，深度学习革命有可能能够改变这一切，突破传统思维，帮助人类解锁地球物候的节律。希望你，就是那个屠龙的天才少年。

- 主办方：佳格天地科技有限公司
- 评价指标：RMSE
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/%E5%9C%B0%E7%90%83%E7%89%A9%E5%80%99%E7%9A%84%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0%E9%A2%84%E6%B5%8B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/01/21-2019/04/30
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/%E5%9C%B0%E7%90%83%E7%89%A9%E5%80%99%E7%9A%84%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0%E9%A2%84%E6%B5%8B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

第四届“魔镜杯”数据应用大赛

https://www.pkbigdata.com/common/cmpt/%E7%AC%AC%E5%9B%9B%E5%B1%8A%E2%80%9C%E9%AD%94%E9%95%9C%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BA%94%E7%94%A8%E5%A4%A7%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：1276 支队伍 • 2002 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

资金流动性管理迄今仍是金融领域的经典问题。在互联网金融信贷业务中，单个资产标的金额小且复杂多样，对于拥有大量出借资金的金融机构或散户而言，资金管理压力巨大，精准地预测出借资金的流动情况变得尤为重要。本次比赛以互联网金融信贷业务为背景，以《现金流预测》为题，希望选手能够利用我们提供的数据，精准地预测资产组合在未来一段时间内每日的回款金额。

本赛题涵盖了信贷违约预测、现金流预测等金融领域常见问题，同时又是复杂的时序问题和多目标预测问题。希望参赛者利用聪明才智把互联网金融的数据优势转化为行业解决方案。

- 主办方：拍拍贷
- 评价指标：RMSE
- 比赛数据：https://www.pkbigdata.com/common/cmpt/%E7%AC%AC%E5%9B%9B%E5%B1%8A%E2%80%9C%E9%AD%94%E9%95%9C%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BA%94%E7%94%A8%E5%A4%A7%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/06/05-2019/07/21
- 比赛结果：

Data Competition in 2019

https://www.pkbigdata.com/common/cmpt/%E7%AC%AC%E5%9B%9B%E5%B1%8A%E2%80%9C%E9%AD%94%E9%95%9C%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BA%94%E7%94%A8%E5%A4%A7%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html

- 赛后分享：

Rank2:https://github.com/LibraM9/ppd_mirror

AI in RTC-超分辨率图像质量比较挑战赛

https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E5%9B%BE%E5%83%8F%E8%B4%A8%E9%87%8F%E6%AF%94%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：506 支队伍 • 657 名参赛者

比赛类型：团队算法赛，数据类型：图像

- 比赛背景：

实时音视频通话（RTC）已经不是互联网时代的陌生词汇，这一技术在我们常见的视频电话、语音通话、会议系统、远程桌面与控制等场景中都有广泛应用。在音视频通信过程中存在一个绕不开的问题，那就是音视频的质量。例如实时视频，视频发出到观看要经历编码、传输、解码、显示这一过程，由于网络环境质量和硬件设备能力的影响就有可能造成视频低清。将 AI 技术与 RTC 相结合，在解码之后借助插入 AI 算法，通过识别视频内容进行视频超分处理，利用超分辨率来提升实时视频中模糊图像的细节，提高视觉体验。同时 AI 也被用来做算法补偿，提升传输质量。

AI in RTC 创新大赛是由声网 Agora 和 RTC 开发者社区发起的人工智能类技术创新大赛，面向全球开发者发起数据算法及创新应用类挑战。

Data Competition in 2019

2019 年，第一届 AI in RTC 创新大赛的主题是 RTC 技术栈中的人工智能。本届大赛发起两项挑战，“超分辨率算法挑战赛”与“编程挑战赛”，开发者可以自由报名参加。开发者可以在图像超分辨率、RTC 应用开发中，发挥自己的创造力，解决学术、商业中、社会中的实际问题。

- 主办方：声网 Agora 和 RTC 开发者社区
- 评价指标：PI (perceptual index)指标和 RMSE
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E5%9B%BE%E5%83%8F%E8%B4%A8%E9%87%8F%E6%AF%94%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/07/01-2019/09/30
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E5%9B%BE%E5%83%8F%E8%B4%A8%E9%87%8F%E6%AF%94%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

AI in RTC-超分辨率算法性能比较挑战

https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E7%AE%97%E6%B3%95%E6%80%A7%E8%83%BD%E6%AF%94%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：283 支队伍 • 359 名参赛者

比赛类型：团队算法赛，数据类型：图像

- 比赛背景：

实时音视频通话（RTC）已经不是互联网时代的陌生词汇，这一技术在我们常见的视频电话、语音通话、会议系统、远程桌面与控制等场景中都有广泛应用。在音视频通信过程中存在一个绕不开的问题，那就是音视频的质量。例如实时视频，视频发出到观看要经历编码、传输、解码、显示这一过程，由于网络环境质量和硬件设备能力的影响就有可能造成视频低清。将 AI 技术与 RTC 相结合，在解码之后借助插入 AI 算法，通过识别视频内容进行视频超分处理，利用超分辨率来提升实时视频中模糊图像的细节，提高视觉体验。同时 AI 也被用来做算法补偿，提升传输质量。

AI in RTC 创新大赛是由声网 Agora 和 RTC 开发者社区发起的人工智能类技术创新大赛，面向全球开发者发起数据算法及创新应用类挑战。

2019 年，第一届 AI in RTC 创新大赛的主题是 RTC 技术栈中的人工智能。本届大赛发起两项挑战，“超分辨率算法挑战赛”与“编程挑战赛”，开发者可以自由报名参加。开发者可以在图像超分辨率、RTC 应用开发中，发挥自己的创造力，解决学术、商业中、社会中的实际问题。

- 主办方：声网 Agora 和 RTC 开发者社区
- 评价指标：PI (perceptual index)指标和 RMSE
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E7%AE%97%E6%B3%95%E6%80%A7%E8%83%BD%E6%AF%94%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/07/01-2019/09/30
- 比赛结果：
- https://www.pkbigdata.com/common/cmpt/AI%20in%20RTC-%E8%B6%85%E5%88%86%E8%BE%A8%E7%8E%87%E7%AE%97%E6%B3%95%E6%80%A7%E8%83%BD%E6%AF%9

4%E8%BE%83%E6%8C%91%E6%88%98%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html

- 赛后分享：

大地量子 AI 台风路径预测大赛

https://www.pkbigdata.com/common/cmpt/%E5%A4%A7%E5%9C%B0%E9%87%8F%E5%AD%90AI%E5%8F%B0%E9%A3%8E%E8%B7%AF%E5%BE%84%E9%A2%84%E6%B5%8B%E5%A4%A7%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：294 支队伍 • 305 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

台风，是一种热带气旋，2018 年超强台风“山竹”使中国南方地区遭受了数十亿的经济损失和 5 人死亡。随着 AI 和遥感技术的发展，各界能够更有效、更精准地预测台风路径，从而减少台风带来的经济损失和人员伤亡。大地量子关注物理世界发生的自然过程，台风就是其中之一。我们每天都在思考如何通过技术来为社会创造价值，在第一次 Hackathon 时，经过我们多位科学家的严格甄选，将题目选定为“通过 AI 手段去对台风路径进行预测”，目的是为了能够减少全球各地可能因为台风所造成的损失。同样，也希望你可以接受这个挑战，和我们一道，共同解决人类关心的重大问题。

- 主办方：大地量子
- 评价指标：RMSE
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/%E5%A4%A7%E5%9C%B0%E9%87%8F%E5%AD%90AI%E5%8F%B0%E9%A3%8E%E8%B7%AF%E5%BE%84%E9%A2%84%E6%B5%8B%E5%A4%A7%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html

Data Competition in 2019

- 时间轴：2019/08/16-2019/11/04
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/%E5%A4%A7%E5%9C%B0%E9%87%8F%E5%AD%90AI%E5%8F%B0%E9%A3%8E%E8%B7%AF%E5%BE%84%E9%A2%84%E6%B5%8B%E5%A4%A7%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html

- 赛后分享：

国能日新第二届光伏功率预测赛

https://www.pkbigdata.com/common/cmpt/%E5%9B%BD%E8%83%BD%E6%97%A5%E6%96%B0%E7%AC%AC%E4%BA%8C%E5%B1%8A%E5%85%89%E4%BC%8F%E5%8A%9F%E7%8E%87%E9%A2%84%E6%B5%8B%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：611 支队伍 • 669 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

光伏发电具有波动性和间歇性，大规模光伏电站的并网运行对电力系统的安全性和稳定造成较大的影响。对光伏电站输出功率的高精度预测，有助于调度部门统筹安排常规能源和光伏发电的协调配合，及时调整调度计划，合理安排电网运行方式。因此，本题旨在通过利用气象信息、历史数据，通过机器学习、人工智能方法，预测未来电站的发电功率，进一步为光伏发电功率提供准确的预测结果。

- 主办方：国能日新科技股份有限公司
- 评价指标：MAE

Data Competition in 2019

- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/%E5%9B%BD%E8%83%BD%E6%97%A5%E6%96%B0%E7%AC%AC%E4%BA%8C%E5%B1%8A%E5%85%89%E4%BC%8F%E5%8A%9F%E7%8E%87%E9%A2%84%E6%B5%8B%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/08/22-2019/10/23
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/%E5%9B%BD%E8%83%BD%E6%97%A5%E6%96%B0%E7%AC%AC%E4%BA%8C%E5%B1%8A%E5%85%89%E4%BC%8F%E5%8A%9F%E7%8E%87%E9%A2%84%E6%B5%8B%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

Rank25:https://github.com/jsnuwjl/Photovoltaic_power

2019 数据智能算法大赛

https://www.pkbigdata.com/common/cmpt/2019%E6%95%B0%E6%8D%AE%E6%99%BA%E8%83%BD%E7%AE%97%E6%B3%95%E5%A4%A7%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：697 支队伍 • 992 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

随着国家新一代人工智能规划的推进，算法成为人工智能产业落地的核心竞争力。为促进高校学生在人工智能领域的行业实践，激发广大高校学子参与人工智能前沿理论和算法研究的热情，推动大

Data Competition in 2019

数据与人工智能的产学研结合落地，西安交通大学软件学院与深圳市云积分科技有限公司将于 2019 年 8 月-12 月联合举办“2019 数据智能算法大赛”。

如果数据是信息化时代的“石油”，那么智能就是“发动机”。在数据密集性行业特别是电子商务领域，数据智能成为赋能行业效能提升的基础。本次竞赛将秉持“以赛引才、以赛促研、以赛兴业”的基本思路，聚焦人工智能“基础研究”+“产融结合”，探索如何利用大数据与人工智能技术解决电子商务行业的关键问题。

本届算法竞赛将于 2019 年 8 月-12 月举行，竞赛面向全国高等院校在校大学生，包含初赛、复赛和决赛三轮。主办单位本着精益求精的精神，将根据实际情况认真研究制定活动方案，周密组织竞赛的各项事宜，保障竞赛的顺利进行。

- 主办方：西安交通大学软件学院、深圳市云积分科技有限公司
- 评价指标：logloss
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/2019%E6%95%B0%E6%8D%AE%E6%99%BA%E8%83%BD%E7%AE%97%E6%B3%95%E5%A4%A7%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/08/23-2019/12/10
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/2019%E6%95%B0%E6%8D%AE%E6%99%BA%E8%83%BD%E7%AE%97%E6%B3%95%E5%A4%A7%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

2019 年“创青春・交子杯”新网银行高校金融科技挑战赛-分布式算法赛道

https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-%E5%88%86%E5%

B8%83%E5%BC%8F%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：242 支队伍 • 375 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

本赛道旨在鼓励学生运用工程思维、分布式技术解决金融科技等领域中的实际问题，激发学生创新能力，增强算法设计与实现能力。比赛挑战题目为“智能路由”，该项技术可解决在互联网高速发展下，银行与互联网平台、科技服务商、其他金融机构相互融合、互联互通所形成的复杂网络产生的数据交换问题。期待挑战者在数据传输受到网络速度、信息安全等各方面限制条件下，设计出一个高可用低延时的网络动态拓扑结构和对应的路由规则。

- 主办方：四川新网银行
- 评价指标：

1、评分与消息的送达率和延迟有关，送达率权重远高于延迟。

2、选手需要在保证送达率的前提下，尽可能降低延迟。

3、详细公式如下：

$$G=(\frac{\sigma\times 2-\bar{\tau}}{\sigma}\times 3.5+\frac{0.5}{D_\tau+1})\times \eta^{2.5}\times 10$$

其中：η 表示：送达率 τ 表示：延迟 σ 表示：最大延迟 D 表示：方差

- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-%E

Data Competition in 2019

5%88%86%E5%B8%83%E5%BC%8F%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html

- 时间轴：2019/09/18-2019/11/16
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-%E5%88%86%E5%B8%83%E5%BC%8F%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E6%8E%92%E8%A1%8C%E6%A6%9C.html

- 赛后分享：

2019 年“创青春・交子杯”新网银行高校金融科技挑战赛-AI 算法赛道

https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-AI%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：773 支队伍 • 1269 名参赛者

比赛类型：团队算法赛，数据类型：图像

- 比赛背景：

2019 年“创青春・交子杯”新网银行高校金融科技挑战赛-AI 算法赛道，旨在鼓励学生运用前沿的人工智能技术解决金融科技等领域中的实际问题，激发学生创新能力,增强其动手能力，本届比赛也

是西南财经大学第十六届统计软件应用暨统计建模大赛。比赛挑战题目为“唇语识别”，该项技术可应用于金融在线业务的生物识别、噪声环境下辅助语音识别、辅助听障人士交流，体育赛事暴力语言识别、安防取证等多个领域，期待挑战者利用机器学习和人工智能的最新成果，设计出区分能力强，稳定性高的唇语识别模型。

- 主办方：西南财经大学、四川新网银行
- 评价指标：ACC
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-AI%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/09/18-2019/11/16
- 比赛结果：

  https://www.pkbigdata.com/common/cmpt/2019%E5%B9%B4%E2%80%9C%E5%88%9B%E9%9D%92%E6%98%A5%C2%B7%E4%BA%A4%E5%AD%90%E6%9D%AF%E2%80%9D%E6%96%B0%E7%BD%91%E9%93%B6%E8%A1%8C%E9%AB%98%E6%A0%A1%E9%87%91%E8%9E%8D%E7%A7%91%E6%8A%80%E6%8C%91%E6%88%98%E8%B5%9B-AI%E7%AE%97%E6%B3%95%E8%B5%9B%E9%81%93_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

Rank1:https://github.com/TimeChi/Lip_Reading_Competition

Rank6:https://github.com/liuzhejun/XWbank_LipReading

2019 厦门国际银行“数创金融杯”数据建模大赛

https://www.pkbigdata.com/common/cmpt/2019%E5%8E%A6%E9%97%A8%E5%9B%BD%E9%99%85%E9%93%B6%E8%A1%8C%E2%80%9C%E6%95%B0%E5%88%9B%E9%87%91%E8%

Data Competition in 2019

9E%8D%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BB%BA%E6%A8%A1%E5%A4%A7%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

参赛人数：1668 支队伍 • 2089 名参赛者

比赛类型：团队算法赛，数据类型：结构化

- 比赛背景：

信用风险是金融监管机构重点关注的风险，关乎金融系统运行的稳定。在实际业务开展和模型构建过程中，面临着高维稀疏特征以及样本不平衡等各种问题，如何应用机器学习等数据挖掘方法提高信用风险的评估和预测能力，是各家金融机构积极探索的方向。本次竞赛提供实际业务场景中的信贷数据作为建模的对象，希望能借此展现各参赛选手数据挖掘的实战能力。

- 主办方：西南财经大学、四川新网银行
- 评价指标：AUC
- 比赛数据：
  https://www.pkbigdata.com/common/cmpt/2019%E5%8E%A6%E9%97%A8%E5%9B%BD%E9%99%85%E9%93%B6%E8%A1%8C%E2%80%9C%E6%95%B0%E5%88%9B%E9%87%91%E8%9E%8D%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BB%BA%E6%A8%A1%E5%A4%A7%E8%B5%9B_%E8%B5%9B%E4%BD%93%E4%B8%8E%E6%95%B0%E6%8D%AE.html
- 时间轴：2019/10/08-2019/12/20
- 比赛结果：
- https://www.pkbigdata.com/common/cmpt/2019%E5%8E%A6%E9%97%A8%E5%9B%BD%E9%99%85%E9%93%B6%E8%A1%8C%E2%80%9C%E6%95%B0%E5%88%9B%E9%87%91%E8%9E%8D%E6%9D%AF%E2%80%9D%E6%95%B0%E6%8D%AE%E5%BB%BA%E6%A8%A1%E5%A4%A7%E8%B5%9B_%E6%8E%92%E8%A1%8C%E6%A6%9C.html
- 赛后分享：

Data Competition in 2019

Rank6:https://github.com/yanqiangmiffy/Data-Finance-Cup

1.6 biendata

短视频内容理解与推荐竞赛

https://biendata.com/competition/icmechallenge2019/

参赛人数：1057 支队伍 • 2714 名参赛者

比赛类型：团队算法赛，数据类型：结构化+文本+音视频

- 比赛背景：

近年来，机器学习在图像识别、语音识别等领域取得了重大进步，但在视频内容理解领域仍有许多问题需要探索。字节跳动旗下短视频产品抖音很受用户欢迎，短视频的内容理解与推荐技术成为了我们关注的焦点。

一图胜千言，仅一张图片就包含大量信息，难以用几个词来描述，更何况是短视频这种富媒体形态。面对短视频内容理解的难题，字节跳动作为一家拥有海量短视频素材和上亿级用户行为数据的公司，通过视频内容特征和用户行为数据，可以有充足的数据来预测用户对短视频的喜好。

本次竞赛提供多模态的短视频内容特征，包括视觉特征、文本特征和音频特征，同时提供了脱敏后的用户点击、喜爱、关注等交互行为数据。参赛者需要通过一个视频及用户交互行为数据集对用户兴趣进行建模，然后预测该用户在另一视频数据集上的点击行为。

- 主办方: ICME2019 & 字节跳动
- 评价指标：AUC
- 比赛数据：https://biendata.com/competition/icmechallenge2019/data/
- 时间轴：2019-02-11 ~ 2019-04-09
- 比赛结果：

Data Competition in 2019

https://biendata.com/competition/icmechallenge2019/leaderboard/

https://biendata.com/competition/icmechallenge2019/final-leaderboard/

- 赛后分享：

官方 baseline：https://github.com/challenge-ICME2019-Bytedance/Bytedance_ICME_challenge

Top8:https://www.zhuanzhi.ai/document/57000873027e2910fd452f42c90b2104

2019 搜狐校园算法大赛

https://biendata.com/competition/sohu2019/

参赛人数：694 支队伍 • 1615 名参赛者

比赛类型：团队算法赛，数据类型：文本

- 比赛背景：

自然语言是人类智慧的结晶，自然语言处理是人工智能中最为困难的问题之一，而对自然语言处理的研究也是充满魅力和挑战的。

在业界，搜狐深耕互联网资讯传播，多年来始终保持着对自然语言处理技术应用的不断探索，积累了丰硕的成果。在学界，清华大学作为顶尖学府，其计算机系在自然语言处理这个前瞻领域收获了大量理论及实践的重要成就。

作为先行者，搜狐携手清华计算机系共同发起本届内容识别算法大赛，旨在通过提供业务场景、真实数据、专家指导，选拔和培养有志于自然语言处理领域的算法研究、应用探索的青年才俊，共同探索更多可能、开启无限未来。

- 主办方: 搜狐
- 评价指标：F1 和 F2
- 比赛数据：https://biendata.com/competition/sohu2019/data/

- 时间轴：2019-04-08 ~ 2019-05-22
- 比赛结果：

  https://biendata.com/competition/sohu2019/leaderboard/

  https://biendata.com/competition/sohu2019/final-leaderboard/

- 赛后分享：

第十名：https://github.com/lmhgithi/2019-sohu-finals

其他：https://github.com/LLouice/Sohu2019

CCKS 2019 中文短文本的实体链指

https://biendata.com/competition/ccks_2019_el/

参赛人数：349 支队伍 • 818 名参赛者

比赛类型：团队算法赛，数据类型：文本

- 比赛背景：

近年来，随着深度学习的重燃以及海量大数据的支撑，NLP 领域迎来了蓬勃发展，百度拥有全球最大的中文知识图谱，拥有数亿实体、千亿事实，具备丰富的知识标注与关联能力，不仅构建了通用知识图谱，还构建了汉语语言知识图谱、关注点图谱、以及包含业务逻辑在内的行业知识图谱等多维度图谱。我们希望通过开放百度的数据，邀请学界和业界的青年才俊共同推进算法进步，激发更多灵感和火花。

面向中文短文本的实体识别与链指，简称 ERL（Entity Recognition and Linking），是 NLP 领域的基础任务之一，即对于给定的一个中文短文本（如搜索 Query、微博、用户对话内容、文章标题等）识别出其中的实体，并与给定知识库中的对应实体进行关联。ERL 整个过程包括实体识别和实体链指两个子任务。

传统的实体链指任务主要是针对长文档，长文档拥有在写的上下文信息能辅助实体的歧义消解并完成链指。相比之下，针对中文短文本的实体链指存在很大的挑战，主要原因如下：（1）口语化严重，导致实体歧义消解困难；（2）短文本上下文语境不丰富，须对上下文语境进行精准理解；（3）相比英文，中文由于语言自身的特点，在短文本的链指问题上更有挑战。

- 主办方: CCKS 2019 & 百度
- 评价指标：F1
- 比赛数据：https://biendata.com/competition/ccks_2019_el/data/
- 时间轴：2019-04-19 ~ 2019-07-25
- 比赛结果：

  https://biendata.com/competition/ccks_2019_el/leaderboard/

  https://biendata.com/competition/ccks_2019_el/final-leaderboard/
- 赛后分享：

CCKS 2019 中文知识图谱问答

https://biendata.com/competition/ccks_2019_6/

参赛人数：180 支队伍 • 670 名参赛者

比赛类型：团队算法赛，数据类型：文本

- 比赛背景：

本评测任务为基于中文知识图谱的自然语言问答，简称 CKBQA （Chinese Knowledge Base Question Answering）。即输入一句中文问题，问答系统从给定知识库中选择若干实体或属性值作为该问题的答案。问题均为客观事实型，不包含主观因素。理解并回答问题的过程中可能需要进行

实体识别、关系抽取、语义解析等子任务。这些子任务的训练可以使用额外的资源，但是最终的答案必须来自给定的知识库。

本次任务一方面扩充了去年开放领域问答的数据规模，另一方面额外增加了适量规模的金融领域问答数据（占总数据量的四分之一）。我们期望参赛选手的问答系统既能处理开放领域的浅层问题，也能处理具备一定领域知识的较深层问题。在后续的任务中，我们可能会针对其他领域加入更多的问答数据。

- 主办方: 北京大学&恒生电子股份有限公司
- 评价指标：Averaged F1
- 比赛数据：https://biendata.com/competition/ccks_2019_6/data//
- 时间轴：2019-04-20 ~ 2019-07-25
- 比赛结果：

  https://biendata.com/competition/ccks_2019_6/leaderboard/

  https://biendata.com/competition/ccks_2019_6/final-leaderboard/
- 赛后分享：

CCKS 2019 人物关系抽取

https://biendata.com/competition/ccks_2019_ipre/

参赛人数：360 支队伍 • 82 名参赛者，比赛类型：团队算法赛

- 比赛背景：

关系抽取（Relation Extraction）是信息抽取的一个重要子任务，其任务是从文本内容中找出给定实体对之间的语义关系，是智能问答、信息检索等智能应用的重要基础，和知识图谱的构建有着密切的联系。

在本次任务中，我们重点关注人物之间的关系抽取研究，简称 IPRE（Inter-Personal Relationship Extraction）。给定一组人物实体对和包含该实体对的句子，找出给定实体对在已知关系表中的关系。

- 评价指标：

本次任务的评价指标包括精确率（Precision, P）、召回率（Recall, R）和 F1 值（F1-measure, F1），分为 Sent-Track 和 Bag-Track 两个部分，每部分按 F1 值分别排名。只统计预测结果中非 NA 的数目（如果 NA 关系预测错误，也会计入到评价指标中计算）。

- 主办方：苏州大学、狗尾草公司
- 比赛数据：https://biendata.com/competition/ccks_2019_ipre/data/
- 时间轴：2019-04-20 ~ 2019-07-25
- 比赛结果：

https://biendata.com/competition/ccks_2019_ipre/leaderboard/

https://biendata.com/competition/ccks_2019_ipre/final-leaderboard/

- 赛后分享：

CrowdHuman 人体检测大赛

https://biendata.com/competition/crowdhuman/

参赛人数：152 支队伍 • 482 名参赛者，比赛类型：团队算法赛

- 比赛背景：

目标检测是计算机视觉和模式识别领域的基础问题之一，对计算机视觉和模式识别领域具有重要的应用价值。因此，旷视和北京智源人工智能研究院联合推出了两个检测任务的新基准：Objects365 和 CrowdHuman，它们都是在自然场景设计和收集的。Objects365 基准目标用于解决具有 365 个对象类别的大规模检测。而 CrowdHuman 则针对人群中的人体检测问题。

我们希望这两个数据集能够为推进目标检测研究提供多样化、实用性的基准。同时，我们围绕数据集组织了两个评测，以及一个挂靠 CVPR 2019 会议的 Workshop。我们希望这些活动可以成为一个平台，推动目标检测研究的上限。

- 评价指标：Jaccard Index（JI）
- 主办方：北京智源人工智能研究院、旷视
- 比赛数据：https://biendata.com/competition/crowdhuman/data/
- 时间轴：2019-04-29 ~ 2019-06-13
- 比赛结果：https://biendata.com/competition/crowdhuman/final-leaderboard/
- 赛后分享：

Objects365 图片物体检测

https://biendata.com/competition/objects365/

参赛人数：211 支队伍 • 607 名参赛者，比赛类型：团队算法赛

- 比赛背景：

目标检测是计算机视觉和模式识别领域的基础问题之一，对计算机视觉和模式识别领域具有重要的应用价值。因此，旷视和北京智源人工智能研究院联合推出了两个检测任务的新基准：Objects365 和 CrowdHuman，它们都是在自然场景设计和收集的。Objects365 基准目标用于解决具有 365 个对象类别的大规模检测。而 CrowdHuman 则针对人群中的人体检测问题。

我们希望这两个数据集能够为推进目标检测研究提供多样化、实用性的基准。同时，我们围绕数据集组织了两个评测，以及一个挂靠 CVPR 2019 会议的 Workshop。我们希望这些活动可以成为一个平台，推动目标检测研究的上限。

- 评价指标：COCO 数据集评测指标
- 主办方：北京智源人工智能研究院、旷视
- 比赛数据：https://biendata.com/competition/objects365/data/
- 时间轴：2019-04-29 ~ 2019-06-13
- 比赛结果：

  https://biendata.com/competition/objects365/leaderboard/

  https://biendata.com/competition/objects365/final-leaderboard/
- 赛后分享：

CCKS 2019 面向金融领域的事件主体抽取

https://biendata.com/competition/ccks_2019_4/

参赛人数：496 支队伍 • 1228 名参赛者，比赛类型：团队算法赛

- 比赛背景：

“事件识别”是舆情监控领域和金融领域的重要任务之一，“事件”在金融领域是投资分析，资产管理的重要决策参考。“事件识别”的复杂性在于事件类型和事件主体的判断，比如“公司 A 产品出现添加剂，其下属子公司 B 和公司 C 遭到了调查”，对于“产品出现问题”事件类型，该句中事件主体是“公司 A”，而不是“公司 B”或“公司 C”。我们称发生特定事件类型的主体成为事件主体，本任务中事件主体范围限定为：公司和机构。事件类型范围确定为：产品出现问题、高管减持、违法违规…

本次评测任务的主要目标是从真实的新闻语料中，抽取特定事件类型的主体。即给定一段文本 T，和文本所属的事件类型 S，从文本 T 中抽取指定事件类型 S 的事件主体。

- 评价指标：

$$\text{事件主体精确率} = \text{识别事件主体与标注相同}/\text{识别事件主体总数量}$$

$$\text{事件主体召回率} = \text{识别事件主体与标注相同}/\text{标注事件主体总数量}$$

$$\text{事件主体F1值} = \frac{2 * \text{事件主体精确率} * \text{事件主体召回率}}{\text{事件主体精确率} + \text{事件主体召回率}}$$

- 主办方：CCKS & 蚂蚁金服 & 中科院自动化所
- 比赛数据：https://biendata.com/competition/ccks_2019_4/data/
- 时间轴：2019-05-01 ~ 2019-08-02
- 比赛结果：https://biendata.com/competition/ccks_2019_4/final-leaderboard/
- 赛后分享：

SMP 2019 ETST “语通杯”文本溯源技术评测

https://biendata.com/competition/smpetst2019/

参赛人数：62 支队伍 • 205 名参赛者，比赛类型：团队算法赛

- 比赛背景：

全国社会媒体处理大会（SMP）由中国中文信息学会社会媒体处理专委会主办，专注于以社会媒体处理为主题的科学研究与工程开发，为传播社会媒体处理最新的学术研究与技术成果提供广泛的交流平台，旨在构建社会媒体处理领域的产学研生态圈，成为中国乃至世界社会媒体处理的风向标，会议将以社交网络的形式改变传统的学术会议交流体验。第八届全国社会媒体处理大会（SMP

2019）由哈尔滨工业大学（深圳）承办，将于 2019 年 8 月 16—18 日在深圳召开。本次会议的评测单元有隐式情感分析、中文人机对话和文本溯源三个项目。

SMP 2019 文本溯源评测由中国中文信息学会社会媒体处理专业委员会主办，黑龙江工程学院承办。本次技术评测以科研立项或成果创新型审查为应用背景，文本溯源的目标是判断一个文本的内容是否复制或改编于另外一个或者多个文本。文本溯源技术在学术诚信检测、搜索引擎优化等领域有广泛应用。

- 评价指标：PlagDet：精准率（Precision)、召回率（Recall）、粒度（Granularity）
- 主办方：2019 社会媒体处理大会
- 比赛数据：https://biendata.com/competition/smpetst2019/data/
- 时间轴：2019-05-20 ~ 2019-07-08
- 比赛结果：https://biendata.com/competition/smpetst2019/final-leaderboard/
- 赛后分享：

Science of Science 数据黑客松

https://biendata.com/competition/hackathon/

参赛人数：156 支队伍 • 331 名参赛者，比赛类型：团队算法赛

- 比赛背景：

科学研究已经成为现代社会创新的主要动力。大量科研数据的积累也让我们可以理解和预测科研发展，并能用来指导未来的研究。因此，芝加哥大学知识实验室（Knowledge Lab at the University of Chicago），清华大学人工智能研究院知识智能研究中心（Tsinghua Joint Research Center for Knowledge and Intelligence）和芝加哥大学北京中心（The University of Chicago Center in Beijing）联合组织了“科学的科学（Science of Science）”国际会议。本次黑客松任务即挂靠这次

会议。在 4 天的时间里，参赛选手需要通过平台提交预测结果。前三名选手将获得 4 万元奖金，并被邀请到会议上介绍获奖方法。

科学论文虽然代表了人类对自然最前沿的理解，但大众却很难读懂论文。所以论文发表后，科研机构经常会发布研究新闻稿解释论文中的研究。本次比赛的任务是给定一篇科研新闻，找到这篇新闻描述的论文。。

- 评价指标：Mean Average Precision @ 3 (MAP@3)
- 主办方：芝加哥大学知识实验室、清华大学人工智能研究院知识智能研究中心、芝加哥大学北京中心
- 比赛数据：https://biendata.com/competition/hackathon/data/
- 时间轴：2019-06-08 ~ 2019-06-11
- 比赛结果：https://biendata.com/competition/hackathon/final-leaderboard/
- 赛后分享：

成语阅读理解大赛

https://biendata.com/competition/idiom/

参赛人数：181 支队伍 • 628 名参赛者，比赛类型：团队算法赛

- 比赛背景：

成语作为汉语的一大特色用语，其形式的简洁与丰富的表现力使得它广泛应用于日常交流与各种文体中。许多成语的含义并非简单字面意义的拼接或合成，而是可能来源于历史故事或具有隐喻含义等，这导致了成语往往不能“望文生义”。同时，相近词之间的细微差别也经常导致成语被误用，如「侃侃而谈」和「口若悬河」，尽管这两个成语都表示说话又多又长，但前者侧重描述说话者的神情，而后者则用以描述说话者的口才。由此可见，对成语有很好的理解和表示，对于中文领域的

机器阅读理解将有很好的促进意义，并且对于中文机器翻译、汉语成语推荐系统等实际应用场景也会有所帮助。

为此，本次竞赛将基于选词填空的任务形式，提供大规模的成语填空训练语料。在给定若干段文本下，选手需要在提供的候选项中，依次选出填入文本中的空格处最恰当的成语。

- 评价指标：正确率
- 主办方：中国计算机学会、清华大学人工智能研究院
- 比赛数据：https://biendata.com/competition/idiom/data/
- 时间轴：2019-06-25 ~ 2019-09-18
- 比赛结果：https://biendata.com/competition/idiom/final-leaderboard/
- 赛后分享：

“达观杯”文本智能信息抽取挑战赛

https://biendata.com/competition/datagrand/

参赛人数：275 支队伍 • 2942 名参赛者，比赛类型：团队算法赛

- 比赛背景：信息抽取（information extraction），即从自然语言文本中，抽取出特定的事件或事实信息，帮助我们将海量内容自动分类、提取和重构。文本智能抽取是信息检索、智能问答、智能对话等人工智能应用的重要基础，它可以克服自然语言非形式化、不确定性等问题，发掘并捕获其中蕴含的有价值信息，进而用于业务咨询、决策支持、精准营销等方面，对产业界有着重要的实用意义。

  达观数据的文本信息抽取技术已应用于金融、制造、通信、法律、审计、媒体、政府等多种文字密集型行业，为企业自动化抽取文档的关键信息、对比不同版本的文档差异、纠正文档文字错误、以及发现文书中潜在的法律风险，以下分享三个实例。

- 评价指标：F1
- 主办方：达观数据
- 比赛数据：https://biendata.com/competition/datagrand/data/
- 时间轴：2019-06-28 ~ 2019-08-31
- 比赛结果：https://biendata.com/competition/datagrand/
- 赛后分享：

第一名:https://zhuanlan.zhihu.com/p/84717061

第九名:https://github.com/lonePatient/daguan_2019_rank9

十强答辩 ppt 下载地址： https://pan.baidu.com/s/1yvXFf5GzyvDksdBKNp9FKQ 提取码: svr2

SMP - ECISA “拓尔思杯”中文隐式情感分析评测 2019

https://biendata.com/competition/smpecisa2019/

参赛人数：223 支队伍 • 481 名参赛者，比赛类型：团队算法赛

SMP - ECISA “拓尔思杯”中文隐式情感分析评测 2019

主办方: SMP 2019，山西大学，拓尔思

时间轴:2019-07-14 ~ 2019-07-15

- 比赛背景：在本届 SMP 会议上，我们将举办“拓尔思杯”中文隐式情感分析评测（SMP-ECISA 2019）。近年来，显式情感分析已经取得了良好的进展与丰硕的成果，但对隐式情感分析研究仍处于起步阶段。隐式情感分析作为情感分析的重要组成部分，其研究成果将有助于更全面、更精确地提升在线文本情感分析的性能，可为文本表示学习、自然语言理解、用户建模、知识嵌入等方面研究起到积极的推动作用，也可进一步促进基于文本情感分析相关领域的应用和产业的快速发展。
- 评价指标：宏平均准确率（P）、召回率（R）及 F1 值
- 主办方：SMP 2019，山西大学，拓尔思
- 比赛数据：https://biendata.com/competition/smpecisa2019/data/
- 时间轴：2019-07-14 ~ 2019-07-15
- 比赛结果：https://biendata.com/competition/smpecisa2019/final-leaderboard/
- 赛后分享：

CCKS 2019 医疗命名实体识别

https://biendata.com/competition/ccks_2019_1/

参赛人数：92 支队伍 • 142 名参赛者，比赛类型：团队算法赛

Data Competition in 2019

- 比赛背景：对于给定的一组电子病历纯文本文档，任务的目标是识别并抽取出与医学临床相关的实体提及（entity mention），并将它们归类到预定义类别（pre-defined categories），比如疾病、治疗、检查检验等。
- 评价指标：精确率（Precision）、召回率（Recall）以及 F1-Measure
- 主办方：CCKS
- 比赛数据：https://biendata.com/competition/ccks_2019_1/data/
- 时间轴：2019-07-14 ~ 2019-07-15
- 比赛结果：https://biendata.com/competition/ccks_2019_1/final-leaderboard/
- 赛后分享：

CCKS 2019 公众公司公告信息抽取

https://biendata.com/competition/ccks_2019_5/

参赛人数：98 支队伍 • 326 名参赛者，比赛类型：团队算法赛

- 比赛背景：随着金融科技的发展和全球资本市场的不断扩大，在金融领域， 每一天都有海量的数据产生， 而与之形成强烈对比的是有限的人力以及人脑所能处理信息的极限能力。因此，依靠传统的人工方式已经无法应对投研分析、风险控制、金融监管和事件关联等需求，而亟需引入新的技术来提高信息处理效率，包括大数据分析、自然语言处理、知识图谱等技术，都已经开始被积极用于金融分析和金融监管领域。在监管方面，每一家公众公司都具有相关信息披露义务，由此也产生了大量的公告阅读和信息抽取需求。 据不完全统计，以沪深股市为例，2017 年共披露公告 44 万余篇，2018 年共 27 万余篇，并且随着上市公司数量的增加这一数字也在逐年增加。每年 3 月底、4 月底、8 月底、10 月底为定期报告披露高峰期，最多的一天所发布公告达 10297 篇。

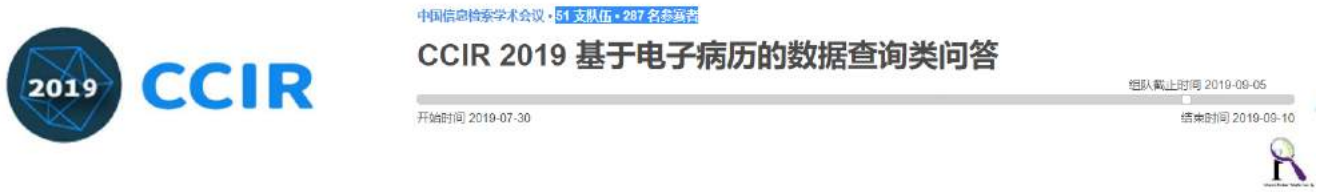
Data Competition in 2019

本次评测的主要目标是针对公告文件（均以 PDF 方式发布）中的信息抽取。作为知识图谱构建的基础，结构化数据是必不可少的。由此，如何通过自动化的技术来从各类公告中抽取信息，将非结构化数据转化为结构化数据是知识图谱领域所面临的一大挑战。

- 评价指标：精确率（Precision）、召回率（Recall）以及 F1-Measure
- 主办方：CCKS & 东南大学
- 比赛数据：https://biendata.com/competition/ccks_2019_5/data/
- 时间轴：2019-07-14 ~ 2019-07-15
- 比赛结果：https://biendata.com/competition/ccks_2019_5/final-leaderboard/
- 赛后分享：

CCIR 2019 基于电子病历的数据查询类问答

参赛人数：51 支队伍 • 287 名参赛者，比赛类型：团队算法赛



- 比赛背景：近年来，基于知识图谱的事实性问答研究有了长足的进展，相比而言，对于带有逻辑关系与运算操作的数据查询类问答，目前还没有很好的解决方法。后者在专业领域有广泛的应用前景和商业价值。医疗电子病历是记录病人信息的主要媒介，对电子病历的查询和分析，在医疗卫生管理和临床科研中有着广泛的用途。本次联合医疗人工智能企业，推出基于电子病历的数据查询类问答评测，希望能够用新一代人工智能技术完成对电子病历的查询与简单的分析和推理。
- 评价指标：宏观准确率(Macro Precision)，宏观召回率(Macro Recall)，Averaged F1 值。最终排名以 Averaged F1 值为基准。
- 主办方：中国信息检索学术会议
- 比赛数据：https://biendata.com/competition/ccir2019/data/

Data Competition in 2019

- 时间轴：2019-07-30 ~ 2019-09-10
- 比赛结果：https://biendata.com/competition/ccir2019/final-leaderboard/
- 赛后分享：

智源 - 看山杯 专家发现算法大赛 2019

https://biendata.com/competition/zhihu2019/

参赛人数：711 支队伍 • 1639 名参赛者，比赛类型：团队算法赛

- 比赛背景：本次比赛是智源 2019 人工智能大赛的任务之一。北京智源人工智能研究院将于 2019 年组织 10 次竞赛，详情请点击这里。

  知识分享服务已经成为目前全球互联网的重要、最受欢迎的应用类型之一。在知识分享或问答社区中，问题数远远超过有质量的回复数。因此，如何连接知识、专家和用户，增加专家的回答意愿，成为了此类服务的中心课题。本次比赛旨在解决这一问题。

  知乎是中文互联网知名的综合性社区平台。知乎自 2011 年创办至今，已经成为一个拥有 2.2 亿用户，每天有数以十万计的新问题以及 UGC 内容产生的网站。其中，如何高效的将这些用户新提出的问题邀请其他用户进行解答，以及挖掘用户有能力且感兴趣的问题进行邀请下发，优化邀请回答的准确率，提高问题解答率以及回答生产数，成为知乎最重要的课题之一。。
- 评价指标：AUC
- 主办方：北京智源人工智能研究院，知乎
- 比赛数据：https://biendata.com/competition/zhihu2019/data/
- 时间轴：2019-08-29 ~ 2019-12-17
- 比赛结果：https://biendata.com/competition/zhihu2019/final-leaderboard/
- 赛后分享：

智源&计算所-互联网虚假新闻检测挑战赛

https://biendata.com/competition/falsenews/

参赛人数：521 支队伍 • 970 名参赛者，比赛类型：团队算法赛

- 比赛背景：互联网虚假信息正在威胁着全球互联网的安全，其在规模、传播速度、造假手段三个方面呈现快速增长。2018 年顶级国际期刊《科学》指出，在 2016 年美国总统大选期间样本选民平均每人每天要接触 4 篇假新闻；要传播至 1500 个选民，假新闻的速度是真实新闻的 6-20 倍。研究认为互联网虚假新闻甚至影响了英国脱欧投票和 2016 年美国总统大选的结果。2018 年底陆续出来的 DeepFake 造假技术（图像视频换脸）和 DeepNude 造假技术（自动生成裸体照片）给各国政府带来了恐慌。国际咨询公司 Gartner 预测，到 2020 年，互联网虚假新闻将面临泛滥之势，基于人工智能技术的造假能力将远超于虚假检测的能力。

  这种现象引起了各国政府和社会群体的空前关切和担忧，其对国家安全、个人与企业声誉和媒体信任度带来了严重冲击。为此，2019 年 6 月，美国国会召开听证会，讨论 DeepFake（深度伪造）技术的风险和对策。呼吁国家加强虚假信息检测技术的研发，以及虚假信息治理执法。2019 年 8 月，人民网舆情中心也发出了同样的呼吁，目前辟谣滞后造成“空窗期”内谣言广泛传播，造假手段不断更新让人工审核力不从心，急需开展人工智能技术和人工审核结合的联合辟谣。

  本次虚假新闻检测由中国科学院计算技术研究所，以及北京智源人工智能研究院共同举办，旨在促进互联网虚假新闻检测技术的发展，营造清朗的网络空间。

- 评价指标：F1
- 主办方：智源研究院 · 中科院计算所
- 比赛数据：https://biendata.com/competition/falsenews/data/
- 时间轴：2019-08-30 ~ 2019-11-06

- 比赛结果：

  https://biendata.com/competition/falsenews/final-leaderboard/

  https://www.biendata.com/competition/falsenews_2/final-leaderboard/

  https://www.biendata.com/competition/falsenews_3/final-leaderboard/

- 赛后分享：

第一名:https://www.biendata.com/models/category/3529/L_notebook/

其他:https://github.com/deping-1/2019-false-news-detection-challenge

平安医疗科技疾病问答迁移学习比赛

https://biendata.com/competition/chip2019/

参赛人数：186 支队伍 • 441 名参赛者，比赛类型：团队算法赛

- 比赛背景：本次比赛是 chip2019 中的评测任务二，由平安医疗科技主办。chip2019 会议详情见链接：http://cips-chip.org.cn/evaluation

  迁移学习是自然语言处理中的重要一环，其主要目的是通过从已学习的相关任务中转移知识来改进新任务的学习效果，从而提高模型的泛化能力。

  本次评测任务的主要目标是针对中文的疾病问答数据，进行病种间的迁移学习。具体而言，给定来自 5 个不同病种的问句对，要求判定两个句子语义是否相同或者相近。所有语料来自互联网上患者真实的问题，并经过了筛选和人工的意图匹配标注.

- 评价指标：F1

- 主办方：平安医疗科技

- 比赛数据：https://biendata.com/competition/chip2019/data/

Data Competition in 2019

- 时间轴：2019-09-18 ~ 2019-11-03
- 比赛结果：https://biendata.com/competition/chip2019/final-leaderboard/
- 赛后分享：

OAG-WhoIsWho 赛道一

https://biendata.com/competition/aminer2019/

参赛人数：278 支队伍 • 576 名参赛者，比赛类型：团队算法赛

- 比赛背景：在许多应用中，同名消歧 (Name Disambiguation - aiming at disambiguating WhoIsWho) 一直被视为一个具有挑战性的问题，如科学文献管理、人物搜索、社交网络分析等，同时，随着科学文献的大量增长，使得该问题的解决变得愈加困难与紧迫。尽管同名消歧已经在学术界和工业界被大量研究，但由于数据的杂乱以及同名情景十分复杂，导致该问题仍未能很好解决。

  收录各种论文的线上学术搜索系统(例 Google Scholar, Dblp 和 AMiner 等)已经成为目前全球学术界重要且最受欢迎的学术交流以及论文搜索平台。然而由于论文分配算法的局限性，现有的学术系统内部存在着大量的论文分配错误；此外，每天都会有大量新论文进入系统。故如何准确快速的将论文分配到系统中已有作者档案以及维护作者档案的一致性，是现有的线上学术系统亟待解决的难题。

  由于学术系统内部的数据十分巨大（AMiner 大约有 130，000，000 作者档案，以及超过 200，000，000 篇论文），导致作者同名情景十分复杂，要快速且准确的解决同名消歧问题还是有很大的障碍。

  竞赛希望提出一种解决问题的模型，可以根据论文的详细信息以及作者与论文之间的联系，去区分属于不同作者的同名论文，获得良好的论文消歧结果。而良好的消歧结果是确保学术系统

中，专家知识搜索有效性、数字图书馆的高质量内容管理以及个性化学术服务的重要前提，也可影响到其他相关领域。

- 评价指标：Macro Pairwise-F1
- 主办方：智源、AMiner
- 比赛数据：https://biendata.com/competition/aminer2019/data/
- 时间轴：2019-09-30 ~ 2019-12-02
- 比赛结果：https://biendata.com/competition/aminer2019/final-leaderboard/
- 赛后分享：

第五名:https://biendata.com/models/detail/3599/L_notebook/

OAG-WhoIsWho 赛道二

https://biendata.com/competition/aminer2019_2/

参赛人数：233 支队伍 • 458 名参赛者，比赛类型：团队算法赛

- 比赛背景：在许多应用中，同名消歧 (Name Disambiguation - aiming at disambiguating WhoIsWho) 一直被视为一个具有挑战性的问题，如科学文献管理、人物搜索、社交网络分析等，同时，随着科学文献的大量增长，使得该问题的解决变得愈加困难与紧迫。尽管同名消歧已经在学术界和工业界被大量研究，但由于数据的杂乱以及同名情景十分复杂，导致该问题仍未能很好解决。

  收录各种论文的线上学术搜索系统(例 Google Scholar, Dblp 和 AMiner 等)已经成为目前全球学术界重要且最受欢迎的学术交流以及论文搜索平台。然而由于论文分配算法的局限性，现有的学术系统内部存在着大量的论文分配错误；此外，每天都会有大量新论文进入系统。故如何准

确快速的将论文分配到系统中已有作者档案以及维护作者档案的一致性，是现有的线上学术系统亟待解决的难题。

由于学术系统内部的数据十分巨大（AMiner 大约有 130，000，000 作者档案，以及超过 200，000，000 篇论文），导致作者同名情景十分复杂，要快速且准确的解决同名消歧问题还是有很大的障碍。

竞赛希望提出一种解决问题的模型，可以根据论文的详细信息以及作者与论文之间的联系，去区分属于不同作者的同名论文，获得良好的论文消歧结果。而良好的消歧结果是确保学术系统中，专家知识搜索有效性、数字图书馆的高质量内容管理以及个性化学术服务的重要前提，也可影响到其他相关领域。

- 评价指标：Weighted F1
- 主办方：智源、AMiner
- 比赛数据：https://biendata.com/competition/aminer2019_2/data/
- 时间轴：2019-09-30 ~ 2019-12-02
- 比赛结果：https://biendata.com/competition/aminer2019_2/final-leaderboard/
- 赛后分享：

第六名:https://biendata.com/models/detail/3598/L_notebook/

第七名:https://zhuanlan.zhihu.com/p/97171241

其他:https://www.ctolib.com/YanYangB-disambiguation.html

DigSci 科学数据挖掘大赛 2019

https://biendata.com/competition/digsci2019/

参赛人数：122 支队伍 • 414 名参赛者，比赛类型：团队算法赛

- 比赛背景：科学研究已经成为现代社会创新的主要动力。大量科研数据的积累也让我们可以理解和预测科研发展，并能用来指导未来的研究。论文是人类最前沿知识的媒介，因此如果可以理解论文中的数据，可以极大地扩充计算机理解知识的能力和范围。

  在论文中，作者经常会引用其他论文，并对被引论文做出对应描述。如果我们可以自动地理解、识别描述对应的被引论文，不仅可以加深对科研脉络的理解，还能在科研知识图谱、科研自动问答系统和自动摘要系统等领域有所进步。

  本次比赛将提供一个论文库（约含 20 万篇论文），同时提供对论文的描述段落，来自论文中对同类研究的介绍。参赛选手需要为描述段落匹配三篇最相关的论文。

- 评价指标：Weighted F1
- 主办方：AMiner · Microsoft
- 比赛数据：https://biendata.com/competition/digsci2019/data/
- 时间轴：2019-10-02 ~ 2019-10-12
- 比赛结果：https://biendata.com/competition/digsci2019/leaderboard/
- 赛后分享：

冠军:https://zhuanlan.zhihu.com/p/88664963

亚军:https://zhuanlan.zhihu.com/p/88257675

基于 Adversarial Attack 的问题等价性判别比赛

https://biendata.com/competition/2019diac/

参赛人数：666 支队伍 • 1164 名参赛者，比赛类型：团队算法赛

- 比赛背景：虽然近年来智能对话系统取得了长足的进展，但是针对专业性较强的问答系统（如法律、政务等），如何准确的判别用户的输入是否为给定问题的语义等价问法仍然是智能问答系统的关键。举例而言，“市政府管辖哪些部门？”和“哪些部门受到市政府的管辖？”可以认为是语义上等价的问题，而“市政府管辖哪些部门？”和“市长管辖哪些部门？”则为不等价的问题。

  针对问题等价性判别而言，除去系统的准确性外，系统的鲁棒性也是很重要、但常常被忽略的一点需求。举例而言，虽然深度神经网络模型在给定的训练集和测试集上常常可以达到满意的准确度，但是对测试集合的稍微改变（Adversarial Attack）就可能导致整体准确度的大幅度下降。

- 评测指标：Macro F1
- 主办方：百分点
- 比赛数据：https://biendata.com/competition/2019diac/data/
- 时间轴：2019-11-05 ~ 2019-12-17
- 比赛结果：https://biendata.com/competition/2019diac/data/
- 赛后分享：

U-RISC 神经元识别大赛

https://biendata.com/competition/urisc/

参赛人数：224 支队伍 • 421 名参赛者，比赛类型：团队算法赛

- 比赛背景：外界环境中的视觉信息通过眼球的光学系统到达眼底的视网膜，在这里信息形式发生转变，光学信号变成可以在神经系统传递的神经信号。毋庸置疑，视网膜是视觉形成的起点，但是视网膜不仅完成信息形式的转变，而且还对信息进行加工，这些加工过的信息才进一

步传输给大脑的视皮层，最终形成我们的视觉。完成信息加工的是视网膜上的多种细胞类型，他们排列成“3+2”的网络结构，即三层细胞层与两层连接层，传统的解剖学研究已经让我们对视网膜有所认识，随着研究技术的发展，现在我们可以在前所未有的分辨率下进一步探索那些未知区域。

除了在生物意义上有重要的作用，机器学习领域也对视网膜感兴趣。自哈佛大学的 David Hubel 和 Torsten Wiesel 深入研究了视网膜和视皮层的原理之后（他们两人获得了 1981 年的诺贝尔医学奖），麻省理工学院的 David Marr 进一步为视觉信息处理建立了数学模型，并影响了后来的人工神经网络研究。他的同事 Tomaso Poggio 至今仍然在麻省理工学院从事前沿人工智能研究。

此后，包括 Geoffery Hinton 的 Capsules 在内的很多模型都借鉴了视网膜和视皮层处理信息的方式。计算生物学家也发现，深度卷积神经网络比很多经典的计算神经生物学模型能更准确地捕捉视网膜对外界的反应，这说明人工神经网络和生物神经网络有一定的相似性。【Deep Learning Models of the Retinal Response to Natural Scenes., Adv Neural Inf Process Syst, 2016.】

对神经系统内细胞的分布和连接的研究，不仅有助于我们理解神经系统如何工作，同时也会推动人工智能的发展，更重要的是可以为目前难以治疗的神经系统疾病提供理论依据。神经科学把对细胞分布和连接的研究称作“连接组学”，目前在不同的研究尺度上，科研人员都努力地获取了大量的数据，例如，一个小鼠的脑部扫描的数据量就在 T 级别。对于这些海量数据，大多数情况下研究人员还只能手动从中获取信息，这无异于“大海捞针”，所以高效地、自动化地挖掘有价值的信息是一个重要而紧迫的任务。

本次比赛要求选手根据对小鼠视网膜的电镜成像图片，像素级地标注（即用不规则多边形，而是精确描绘出神经元的形状）每个神经元的外廓（只标注神经元的最外层细胞膜，神经元内部不进行填充标注）。

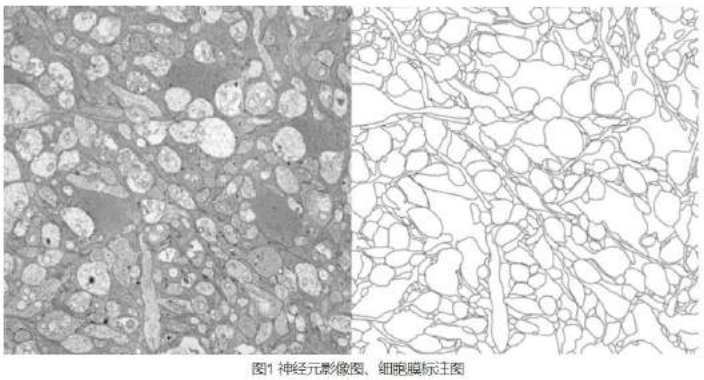


图1 神经元影像图、细胞膜标注图

Data Competition in 2019

- 评测指标：F-score
- 主办方：智源研究院
- 比赛数据：https://biendata.com/competition/urisc/data/
- 时间轴：2019-10-31 ~ 2020-01-15
- 比赛结果：

  https://biendata.com/competition/urisc/leaderboard/

  https://biendata.com/competition/urisc/final-leaderboard/
- 赛后分享：

四川航空—航班智能调整与机组资源协同决策

https://biendata.com/competition/airtraffic/

参赛人数：96 支队伍 • 247 名参赛者，比赛类型：团队算法赛

- 比赛背景：自 2002 年 8 月 29 日四川航空股份有限公司（以下简称“川航”）成立以来，除成都总部，已设有重庆分公司、北京分公司、云南分公司等九个分公司，深圳、南宁、绵阳等三个运行基地，并且开通温哥华、墨尔本、悉尼、莫斯科、迪拜、东京、大阪、新加坡、布拉格、洛杉矶、奥克兰、圣彼得堡、苏黎世等国际航线。在为大量旅客服务的同时，也面临着很多挑战。在夏季雷雨、大雾等极端天气对多基地运行造成影响的情况下，都有可能导致相关机场出现大面积航班延误的情况。为保障航班运行安全正点有序，签派员需要实时监控航班动态，对航班延误、流量控制、飞机故障等突发情况及时处理和调整计划。

  川航目前使用人工决策的方式来进行航班调整，但是对于涉及到多基地运行、可能导致的大面积延误等情况，仅依靠人工方式不能满足对现场运行高效性、准确性的要求。基于此，川航提出“航班智能调整决策”的研究，预期目标是：实现系统自动根据后续可能延误的航班情况推

荐出优化调整方案的核心算法。具体要求是：当遇到某些特定情况，导致航班发生延误时，该算法能够在满足多种实际约束条件的前提下，可以对航班计划进行恢复，并快速给出最优的航班调整替换方案，并且根据航班计划对机组排班计划进行调整，使得机组的资质等与航班计划可以匹配，从而使得航班与机组计划得到快速恢复、减少航班延误、提高航班正常率，使旅客有更好的出行体验，并提升公司的运行效率与经济效益。

已知川航一个月内的全部航班计划与机组排班计划，假设在下列场景描述的情况下，航班将发生延误，需要进行调整，使得后续航班的延误情况得以缓解，同时机组的排班计划也可以得到相应的恢复。调整需要满足航班运行与机组编排的各类约束条件，目的是将航班运行情况受到的影响降到最低（使目标函数值达到最小）。

- 评测指标：目标损失、适用性、完备性；
- 主办方：四川航空
- 比赛数据：https://biendata.com/competition/airtraffic/data/
- 时间轴：2019-11-27 ~ 2020-01-17
- 比赛结果：https://biendata.com/competition/airtraffic/final-leaderboard/
- 赛后分享：

1.7 JData

JD-AR & ARCore by Google 消费应用创新大赛

https://jdata.jd.com/html/detail.html?id=7

参赛人数：455，比赛类型：团队创新赛

- 比赛背景：JD-AR & ARCore by Google 消费应用创新大赛——国内首次针对 AR 技术的大规模竞赛，由京东和 Google 联手打造，面向国内外开发者团队或个人，致力于为全球开发者们提供

国际一流水平的开发与交流机会，并助力他们运用 ARCore 创造出针对消费应用场景下，创新、实用、可落地的解决方案。

本次大赛以无界零售为出发点，汇聚全球 AR 开发者，进行头脑风暴。通过创造性的技术碰撞，共同发力突破行业发展瓶颈，挖掘现有及潜在的 AR 发展价值，面向消费行业打造一站式解决方案。

- 评价指标：AR 技术创新、创景体验与互动和商业应用价值
- 比赛数据：
- 时间轴：2019 年 04 月 02 日 - 2019 年 06 月 11 日
- 比赛结果：
- 赛后分享：

用户对品类下店铺的购买预测

https://jdata.jd.com/html/detail.html?id=8

参赛人数：1401，比赛类型：团队算法赛，比赛数据：结构化

- 比赛背景：京东零售集团坚持“以信赖为基础、以客户为中心的价值创造”这一经营理念，在不同的消费场景和连接终端上，在正确的时间、正确的地点为 3 亿多活跃用户提供最适合的产品和服务。目前，京东零售集团第三方平台签约商家超过 21 万个，实现了全品类覆盖，为维持商家生态繁荣、多样和有序，全面满足消费者一站式购物需求，需要对用户购买行为进行更精准地分析和预测。

  基于此，本赛题提供来自用户、商家、商品等多方面数据信息，包括商家和商品自身的内容信息、评论信息以及用户与之丰富的互动行为。参赛队伍需要通过数据挖掘技术和机器学习算法，构建用户购买商家中相关品类的预测模型，输出用户和店铺、品类的匹配结果，为精准营销提供高质量的目标群体。同时，希望参赛队伍通过本次比赛，挖掘数据背后潜在的意义，为

电商生态平台的商家、用户提供多方共赢的智能解决方案。评价指标：AR 技术创新、创景体验与互动和商业应用价值。

- 比赛数据：https://jdata.jd.com/html/detail.html?id=8
- 时间轴：2019 年 04 月 18 日 - 2019 年 06 月 13 日
- 比赛结果：https://jdata.jd.com/html/detail.html?id=8
- 赛后分享：

雪豹识别全球挑战赛

https://jdata.jd.com/html/detail.html?id=9

参赛人数：334，比赛类型：团队算法赛，比赛数据：视频

- 比赛背景：雪豹被称为“雪山之王”，生活在世界高海拔的雪域高原。它们是高山生态系统的旗舰物种和伞护物种，也是气候变化和水资源安全的指示物种。然而目前，由于全球气候变暖、人类社会经济活动发展、人兽冲突等因素，全球范围内的雪豹正面临着日益严峻的生存威胁，而科学的保护应对方法，需要建立于充分的基础数据收集与分析之上。我国是当之无愧的雪豹大国，据估计，全球 50%以上的雪豹适宜栖息地及其种群分布在中国境内，更加翔实的科学调查研究仍待开展。传统监测作为一种有效的手段，也面临很多问题，包括：监测影像数据大量冗余，物种及雪豹个体识别带来大量人力负荷，数据时效性较差，预测分析不够智能。参赛队伍需要通过数据挖掘的技术和机器学习的算法，根据提供的雪豹视频数据设计和训练模型，实现对不同雪豹个体的识别。
- 比赛数据：https://jdata.jd.com/html/detail.html?id=9
- 时间轴：2019 年 8 月 30 日 - 2019 年 11 月 18 日
- 比赛结果：https://jdata.jd.com/html/detail.html?id=9
- 赛后分享：

1.8 点石

第二届中国“高分杯”美丽乡村大赛

https://dianshi.baidu.com/competition/28/rule

参赛队伍：537，比赛类型：团队赛，比赛数据：图像

- 比赛背景：为了加速推动我国高分辨率对地观测系统重大专项卫星遥感数据在农业农村领域的应用创新，促进乡村振兴，开拓现代农业与农村大众创业、万众创新的新局面，展现中国乡村文化与美丽国土的魅力，举办“中国高分杯”美丽乡村大赛。

  希望通过此次大赛展示美丽乡村、发现优秀作品、对接优质企业、链接社会大众、促进高分应用、服务乡村振兴。同时，通过竞赛引导学生关注人工智能在遥感农业产业的应用，激发学生对人工智能技术和产业的热爱，鼓励通过团队协作，综合运用所学知识，围绕农业应用场景，迸发创新智慧。另一方面为参赛者提供交流平台，进一步提升图像识别领域的研究水平，推人工智能领域技术和应用的发展。

- 评价指标：像素分类准确率
- 比赛数据：https://dianshi.baidu.com/competition/28/data
- 时间轴：2019 年 1 月 21 日 - 2019 年 3 月 20 日
- 比赛结果：https://dianshi.baidu.com/competition/28/rank
- 赛后分享：

第一名：https://blog.csdn.net/nima1994/article/details/89849773

Urban Region Function Classification

https://dianshi.baidu.com/competition/30/rule

Data Competition in 2019

参赛队伍：2312，比赛类型：团队赛，比赛数据：结构化+图像

- 比赛背景：Jointly organized by the International Knowledge Centre for Engineering Sciences and Technology under the Auspices of UNESCO (IKCEST), China Knowledge Centre for Engineering Sciences and Technology(CKCEST), Baidu and Xi'an Jiaotong University (XJTU), the big data algorithm contest aims to identify top talents through the contest format in the global big data and artificial intelligence sector, particularly the talents from "The Belt and Road" countries, in an effort to help the Government, Industry, and Higher Education Institutions jointly drive the research, application and development of the big data industry, consolidate the contest's theoretical and practical foundations and accelerate the nurturing of innovative AI talents.
- 评价指标：准确率
- 比赛数据：https://dianshi.baidu.com/competition/30/data
- 时间轴：April 30th, 2019 - September,2019
- 比赛结果：https://dianshi.baidu.com/competition/30/rank
- 赛后分享：

第二名：https://github.com/zhuqunxi/Urban-Region-Function-Classification

Context-Aware Multi-Modal Transportation Recommendation

https://dianshi.baidu.com/competition/29/rule

参赛队伍：1696，比赛类型：团队赛，比赛数据：结构化

Context-Aware Multi-Modal Transportation Recommendation

报名截止时间：2019-05-24 | 参赛队伍：1696 | 举办方：Baidu Inc. & KDD Cup

$45000

已结束

- 比赛背景：Context-aware multi-modal transportation recommendation has a goal of recommending a travel plan which considers various unimodal transportation modes, such as walking, cycling, driving, public transit, and how to connect among these modes under various contexts. The successful development of multi-modal transportation recommendations can have a number of advantages, including but not limited to reducing transport times, balancing traffic flows, reducing traffic congestion, and ultimately, promoting the development of intelligent transportation systems.

  Despite the popularity and frequent usage of transportation recommendation on navigation Apps (e.g., Baidu Maps and Google Maps), existing transportation recommendation solutions only consider routes in one transportation mode. Intuitively, in the context-aware multi-modal transportation recommendation problem, the transport mode preferences vary over different users and spatiotemporal contexts. For example, metros are more cost-effective than taxis for most urban commuters; economically disadvantaged people may prefer cycling and walking to others for local travel, if the transport options are inadequate. Imagine another scenario that the distance of the OD pair is relatively large, and the trip purpose is in no emergency. In this case, a cost-effective transportation recommendation that including multiple transport modes, e.g., taxi-bus, maybe more attractive.

- 评价指标：F1
- 比赛数据：https://dianshi.baidu.com/competition/29/data
- 时间轴：April 30th, 2019 - September,2019
- 比赛结果：https://dianshi.baidu.com/competition/29/rule
- 赛后分享：

第二名：https://mp.weixin.qq.com/s/CHLxzXo2dV6RY-JVam510w

“智荟杯”2019 全国高校金融科技创新大赛

https://dianshi.baidu.com/competition/32/rule

参赛队伍：341，比赛类型：团队赛，比赛数据：图像

- 比赛背景：为促进国家“互联网+”及“双创”战略的实施，加强金融科技领域沟通与协作，引导金融科技研究和应用创新，激发高校大学生的创新意识和创业精神，搭建金融科技产学研共享平台、合作交流平台、创新创业平台，助力高校推动金融科技相关学科的深化发展，助力金融科技产业人才培养，由教育部高等学校计算机类专业教学指导委员会、中国通信学会主办，浦发银行、百度智能云联合发起，金融科技创新联盟协办的“‘智荟杯’2019 全国高校金融科技创新大赛”将于 2019 年 10-12 月全面展开。

  本次大赛将紧盯国内外顶尖高校在金融科技领域创新板块中最具原创性和竞争力的研发成果，展开一场国际间超水平的科创竞赛。在全球顶尖研究类大学的学术平台上，与海内外创新领域充满朝气活力的专家、教授和创新创业者，共同探索未来金融科技领域的创新发展方向。同时本着为企业及社会发掘、培育和输送优质的科创人才的初心，深耕人才和企业、人才和社会产业互动的土壤，发掘最优秀的金融科技创新成果，对接市场，携手共进，树立标杆，“智”造未来。

- 评价指标：F1
- 比赛数据：https://dianshi.baidu.com/competition/32/data
- 时间轴：April 30th, 2019 - September,2019
- 比赛结果：https://dianshi.baidu.com/competition/32/rank
- 赛后分享：2019/10/15 - 2019/12/13

1.10 AI 研习社

200 种鸟类识别分类

https://god.yanxishe.com/4

参赛人数：108，比赛类型：单人赛，比赛数据：图像

- 比赛背景：数据集来自加利福利亚理工，200 种鸟类，已经重新分割。
- 评价指标：分类准确率
- 比赛数据
  https://static.leiphone.com/AI%E7%A0%94%E4%B9%A0%E7%A4%BE_%E9%B8%9F%E7%B1%BB%E8%AF%86%E5%88%AB%E6%AF%94%E8%B5%9B%E6%95%B0%E6%8D%AE%E9%9B%86.rar
- 时间轴：2019/08/27 - 2020/09/28
- 比赛结果：https://god.yanxishe.com/4
- 赛后分享

第一名：https://www.yanxishe.com/blogDetail/14749

第二名：https://www.yanxishe.com/resourceDetail/1040

中文对话情感分析

https://god.yanxishe.com/5

参赛人数：85，比赛类型：单人赛，比赛数据：文本

- 比赛背景：分析中文对话的情感，共两种状态：positive（积极）、negative（消极）。
- 评价指标：分类准确率
- 比赛数据：http://1t.click/bcn5
- 时间轴：2019/08/27 - 2020/09/27
- 比赛结果：https://god.yanxishe.com/5
- 赛后分享

第一名：https://www.yanxishe.com/resourceDetail/1042

第二名：https://www.yanxishe.com/resourceDetail/1043

猫狗大战--经典图像分类题

https://god.yanxishe.com/8

参赛人数：110，比赛类型：单人赛，比赛数据：图像

- 比赛背景：训练集猫和狗各 10000 张图片，没有任何标注，选手需要自己提取图像特征。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/cat_dog.rar
- 时间轴：2019/09/10 - 2020/10/10
- 比赛结果：https://god.yanxishe.com/8
- 赛后分享

呼吸声音识别呼吸系统疾病

https://god.yanxishe.com/9

参赛人数：33，比赛类型：单人赛，比赛数据：音频

- 比赛背景：通过呼吸声音判断呼吸系统疾病，是临床医生诊断的一个重要参考。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/respiratory-sound-dataset.tar.gz
- 时间轴：2019/09/21 - 2020/10/20
- 比赛结果：https://god.yanxishe.com/9
- 赛后分享

人脸年龄识别

https://god.yanxishe.com/10

Data Competition in 2019

参赛人数：178，比赛类型：单人赛，比赛数据：图像

- 比赛背景：年龄是人类重要的生物特征，根据人脸面部图像推测年龄，这将极大满足日常生活中各种基于年龄的人机交互系统。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/face_age_dataset.zip.gz
- 时间轴：2019/10/01 - 2020/11/01
- 比赛结果：https://god.yanxishe.com/10
- 赛后分享

英文垃圾信息分类

https://god.yanxishe.com/11

参赛人数：88，比赛类型：单人赛，比赛数据：文本

- 比赛背景：利用贝叶斯对垃圾短信、邮件进行分类，是机器学习经典实例。本次比赛不限算法、框架，选手们可以尽情发挥！
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/sms_spam.zip
- 时间轴：2019/10/01 - 2020/11/01
- 比赛结果：https://god.yanxishe.com/11
- 赛后分享

安全帽佩戴检测赛

https://god.yanxishe.com/12

参赛人数：289，比赛类型：单人赛，比赛数据：图像

Data Competition in 2019

- 比赛背景：安全帽是指对人头部受坠落物及其他特定因素引起的伤害起防护作用的帽子。安全帽能有效降低安全事故，因此安全帽佩戴检测是当前最具应用性价比的图像识别技术。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/Safety_helmet.zip
- 时间轴：2019/10/30 - 2020/11/30
- 比赛结果：https://god.yanxishe.com/12
- 赛后分享

胸腔 X 光肺炎检测

https://god.yanxishe.com/13

参赛人数：120，比赛类型：单人赛，比赛数据：图像

- 比赛背景：肺炎是一种常见且多发的呼吸系统疾病，X 光是医生判断肺炎的重要依据。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/Safety_helmet.zip
- 时间轴：2019/10/25 - 2020/11/25
- 比赛结果：https://god.yanxishe.com/13
- 赛后分享

肌肉活动电信号推测手势

https://god.yanxishe.com/14

参赛人数：253，比赛类型：单人赛，比赛数据：结构化

- 比赛背景：肌肉的舒张和收缩可以通过电极转化为电信号，从而控制假肢、机械臂，具有广泛的应用场景。
- 评价指标：分类准确率

Data Competition in 2019

- 比赛数据：https://static.leiphone.com/gesture.zip
- 时间轴：2019/11/08 - 2020/12/08
- 比赛结果：https://god.yanxishe.com/14
- 赛后分享

白葡萄酒品质预测

https://god.yanxishe.com/15

参赛人数：155，比赛类型：单人赛，比赛数据：结构化

- 比赛背景：评价一款葡萄酒时不外乎从颜色、酸度、甜度、香气、风味等入手，而决定这些就是葡萄酒的挥发酸度、糖分、密度等。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/winequality_dataset.zip
- 时间轴：2019/11/19 - 2020/01/19
- 比赛结果：https://god.yanxishe.com/15
- 赛后分享

美食识别挑战（1）：豆腐 VS 土豆

https://god.yanxishe.com/16

参赛人数：294，比赛类型：单人赛，比赛数据：图像

- 比赛背景：豆腐和土豆是中国美食常用的两种食材，本次挑战就是从众多美食图片中正确识别是否包含豆腐或者土豆。
- 评价指标：分类准确率
- 比赛数据：http://1t.click/bqUV
- 时间轴：2019/11/28 - 2020/12/28

Data Competition in 2019

- 比赛结果：https://god.yanxishe.com/16
- 赛后分享

肺炎 X 光病灶识别

https://god.yanxishe.com/18

参赛人数：318，比赛类型：单人赛，比赛数据：图像

- 比赛背景：通过 X 光检测肺炎病灶数量，是医生判断病进展情的重要依据。
- 评价指标：分类准确率
- 比赛数据：http://1t.click/bxSW
- 时间轴：2019/12/09 - 2020/01/11
- 比赛结果：https://god.yanxishe.com/18
- 赛后分享

喵脸关键点检测

https://god.yanxishe.com/19

参赛人数：148，比赛类型：单人赛，比赛数据：图像

- 比赛背景：关键点检测是许多计算机视觉任务的基础，例如表情分析、异常行为检测。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/cat_face.zip
- 时间轴：2019/12/19 - 2020/01/18
- 比赛结果：https://god.yanxishe.com/19
- 赛后分享

IMDB 评论剧透检测

https://god.yanxishe.com/20

参赛人数：137，比赛类型：单人赛，比赛数据：结构化+文本

- 比赛背景：“剧透”是非常影响大家观影追剧体验的行为，所以本次赛题就是检测影评是否包含剧透信息。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/IMDB.zip
- 时间轴：2019/12/30 - 2021/01/29
- 比赛结果：https://god.yanxishe.com/20
- 赛后分享

心跳异常检测

https://god.yanxishe.com/21

参赛人数：67，比赛类型：单人赛，比赛数据：结构化

- 比赛背景：本数据集来自 PTB 诊断心电图数据库，分为正常心跳数据和异常心跳数据。
- 评价指标：分类准确率
- 比赛数据：https://static.leiphone.com/heartbeat.zip
- 时间轴：2020/01/10 - 2021/02/09
- 比赛结果：https://god.yanxishe.com/20
- 赛后分享

1.11 图灵联邦

BONC Cloudiip 工业仪表表盘读数大赛

http://www.turingtopia.com/competitionnew/detail/53aa39e8d46048d8a4de2c6d21adafb1/sketch

Data Competition in 2019

参赛人数：67，比赛类型：团队赛，比赛数据：图像

- 比赛背景：在人工智能落地的领域中，工业自动化将是一个很有前景的领域，但同时也面临着独特的挑战。
- 评价指标：分类准确率
- 比赛数据：http://www.turingtopia.com/competitionnew/detail/53aa39e8d46048d8a4de2c6d21adafb1/dataset
- 时间轴：2019-11-12 - 2020-02-25
- 比赛结果：
- 赛后分享

视频点击预测大赛

http://www.turingtopia.com/competitionnew/detail/e4880352b6ef4f9f8f28e8f98498dbc4/sketch

参赛人数：400 类型：团队赛，比赛数据：结构化

- 比赛背景：通过用户行为数据，用户特征，以及视频特征，可以在充足数据基础上精准的推荐给用户喜欢的视频类型。
- 评价指标：F1
- 比赛数据：http://www.turingtopia.com/competitionnew/detail/e4880352b6ef4f9f8f28e8f98498dbc4/dataset
- 时间轴：2019-11-16 - 2020-02-25
- 比赛结果：
- 赛后分享

02 竞赛干货分享

DF | 多人种人脸识别冠军分享

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

天才儿童队伍由三名队员组成，成员由在读研究生和博士生组成，主要的研究方向是图像特征提取，目标识别与检测。团队有着丰富的项目经验和比赛经验。

文本是他们本次参赛的分享，比赛代码已经开源。

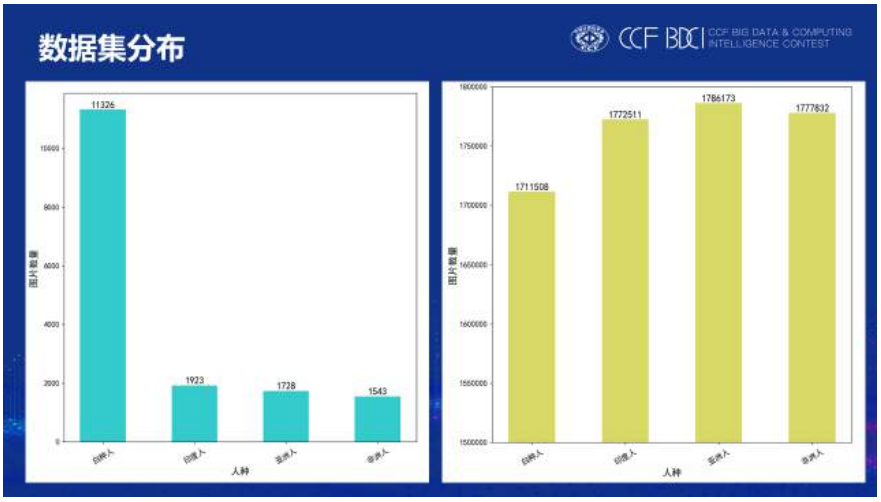
> 这次比赛图像领域的一些大佬并没有参加，给了我们一次咸鱼翻身的机会。
>
> 一般这种图像赛最少 2 张卡，4 张卡以上最佳，实验室或者公司应该有这种硬件条件，这也是相对于表格结构数据比赛的门槛（顺便吐槽下，奖金较少还要扣掉租服务器，但是去郑州玩得很开心，主办方非常用心在此表示感谢）。
>
> 拿到特等奖完全是意料之外的事情，八分运气在里面。要知道优秀的队伍很多，比我们强的也不少。技术之外的秘籍这种事情你认为有就有，认为没有就没有，只要相信自己的本心努力就好。
>
> by 天才儿童

- 方案摘要

随着人脸识别技术不断成熟，市场需求将加速释放，应用场景不断被挖掘。从社保领取到校园门禁，从远程授权到安检闸机检查，人脸识别正在不断打开市场。人脸识别应用在加速普及，行业也呈现出新的发展趋势。本文应用深度学习技术对人脸数据集进行建模，进行了数据与模型方面的探索，搭建了一套面向实际场景的人脸识别方案。我们首先提出了适合人脸图像的数据增强方法，构建了有监督与无监督结合的人脸识别模型，并搭建了一种推理模型，以此对人脸图片进行相似度判断。团队本次复赛 A 榜得分 0.6582，B 榜得分 0.6583，均为线上第一。

- 比赛背景&数据集

我们可以看到训练集中白种人数量远超过其他人种，而测试集中各人种数量均衡。



从数据集统计可以看到数据集具有多样性。

第 4 种是我们发现的主办方创造数据集的方法，如果有时间的话，可以根据这个思路继续做图像增强，训练一个使用主办方方法增强的模型，效果应该会更好。

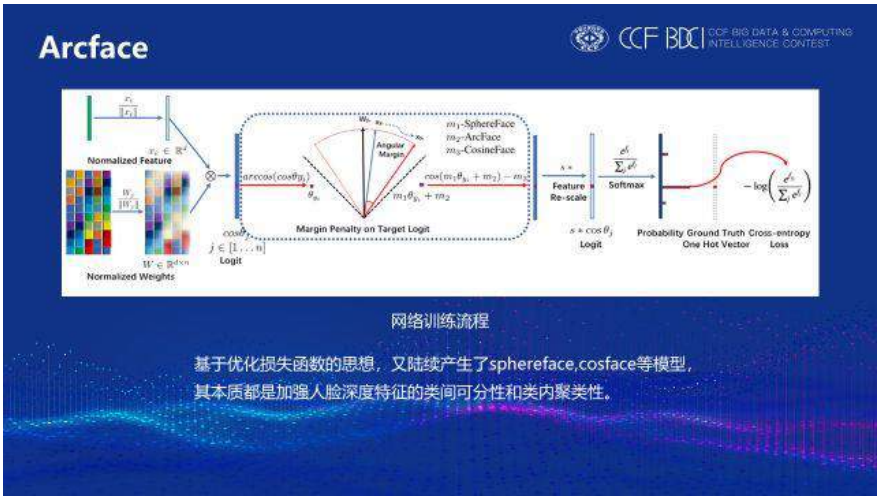
图像增强的方法非常多，大概包括这些方法：水平翻转、旋转、平移、缩放、直方图、平滑、滤波、色度抖动、锐度抖动。而且这些方法相互组合效果也不一样。除了线上提交以外，可以先建立几种指标线下判断。

现有的人脸识别测试过程，通常是提取人脸的特征向量，再通过向量距离比如余弦相似度进行对比，而非直接通过网络推理得到标签。特征提取作为人脸识别最关键的步骤，已经有了很多成熟的网络模型。接下来我们主要从神经网络的损失函数，骨干网络 2 个角度进行分析。

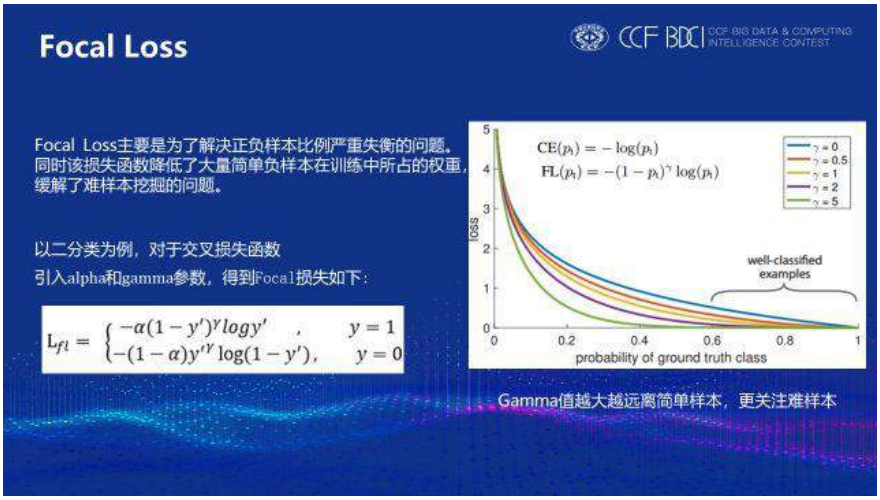
- 有监督模型

在 ResNet2015 年被提出后，越来越多优秀的网络基于 ResNet 进行优化更新也取得卓越成就，而在网络结构与进一步升级优化有困难的情况下，研究者逐步将目光转向损失函数上。2018 年提出的 arcface 的主要创新点就是优化损失函数。在传统的 Softmax 基础上提出了 Center Losas，最终 arcface 损失函数包含 Softmax loss 和 Center Loss。

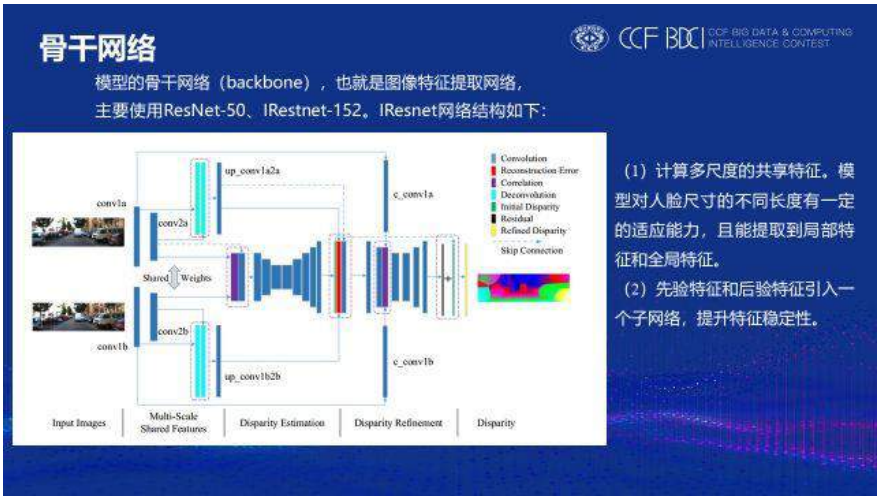
模型使用 arcface 效果最好，人脸识别方面就是 arcface 一枝独秀了，其他模型在比赛中效果有差距但可以用于模型融合。

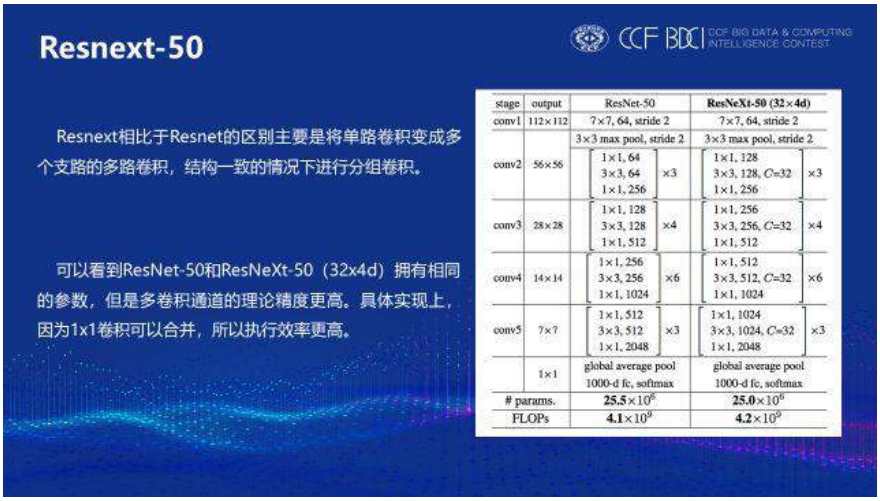


损失函数采用 Focal Loss 最佳。



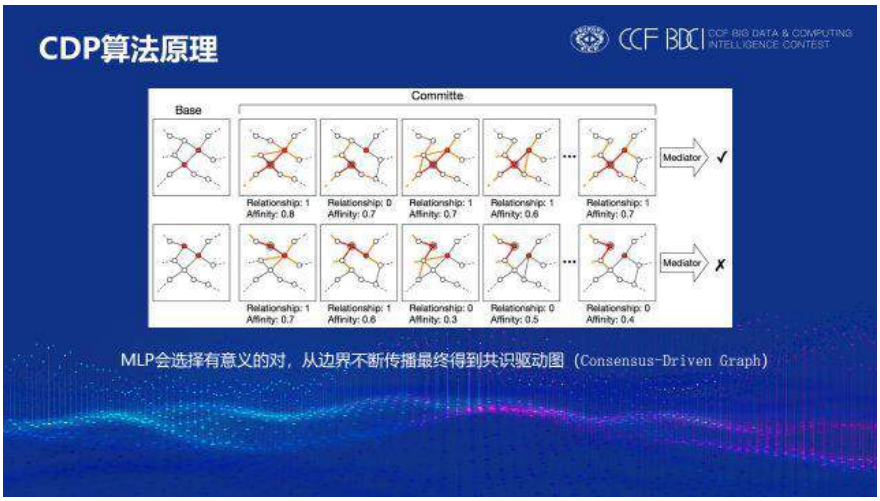
模型的骨干网络使用 Resnet 系列，当然还融合了其他的网络包括 Resnet,Resnext,Densenet。





- 无监督模型

然后我们还使用了无监督的模型进行后处理。ECCV2018 的论文《基于模型共识的大规模无标注数据标签传播》。其假设已经有了少量的标注图像，对于大量的未标注图像，生成其样本标签对，可以称其为半监督算法。



我们对其进行了改进，最终效果是一种无监督的人脸聚类模型。人脸聚类的方法不少，第二名也使用了类似聚类的方法。我们选择这个模型的主要原因是因为快，当然第二名的方法也许精度更高。

人种后验概率修正，利用人种的信息对预测概率进行修正。

当获得了聚类 ID 后：

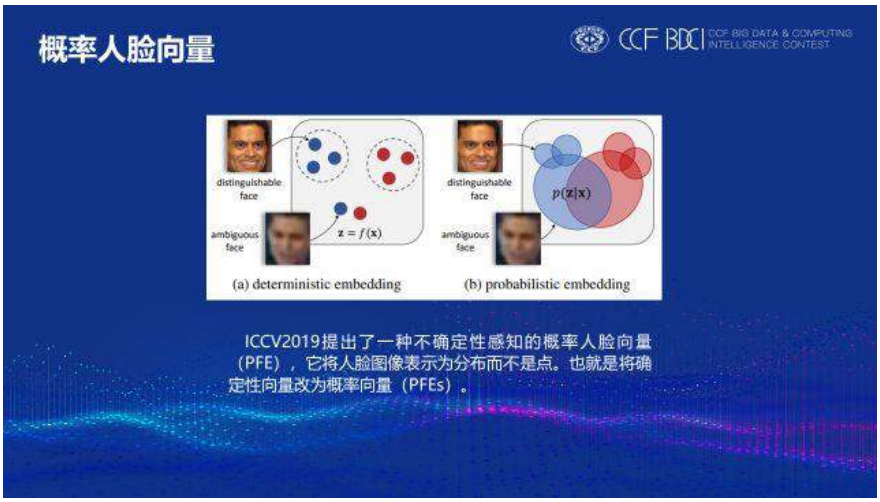
（1）对同一类 ID 判断为同一个人，

（2）不同类 ID 判断为不同人，

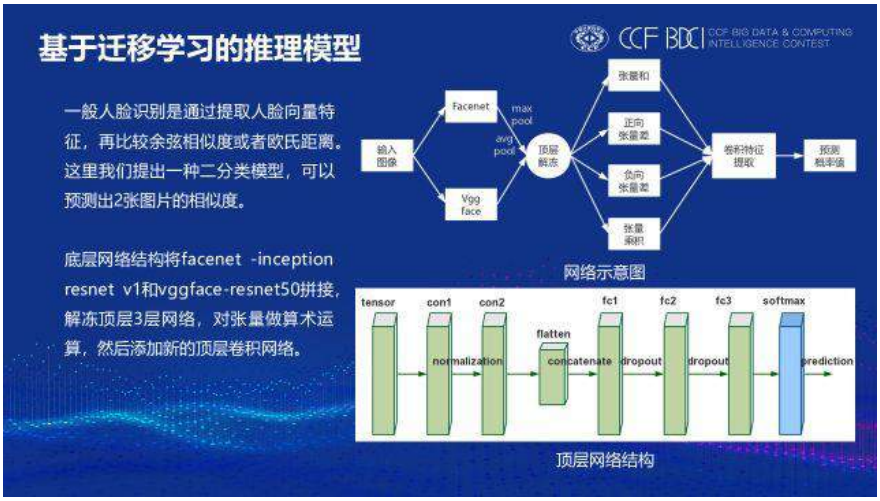
（3）对于没有划分的人，取原概率。

- 模型融合

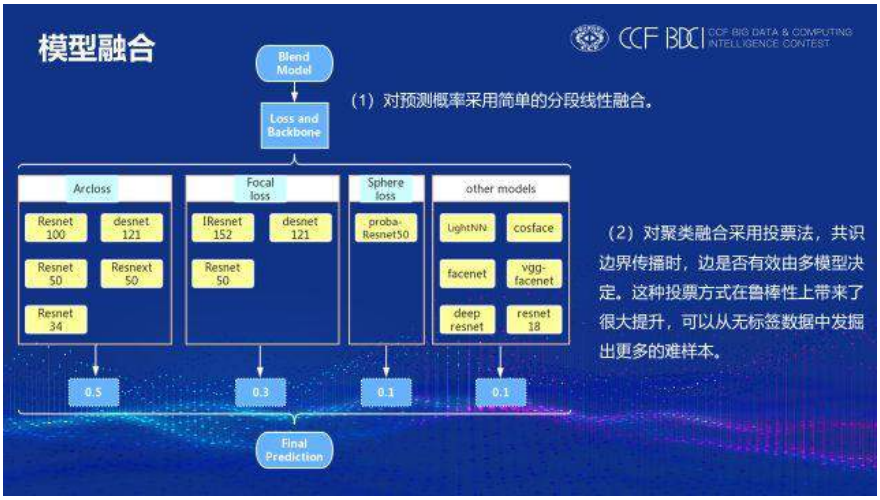
这里再介绍一些模型：这个模型分数并不高，但融合效果较好。



这个是我们搭建的模型，实际方案没有采用，这里提供给大家思路。

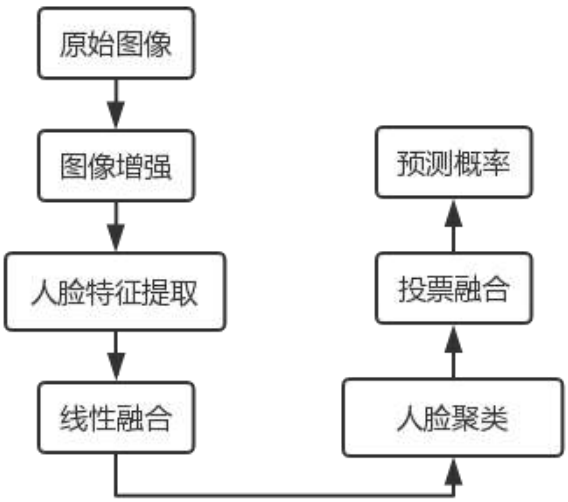


融合方案如下：



图中是 15 种模型，低分模型的权重已经是 1%左右了，提升只是万分位的，如果不刷分只需要效果最好的 5 种模型就可以了。

得到最终方案：



（1）提出了一种人脸图像优化算法，建立了优化指标，可以较为准确地选出有效的图像增强方法。

（2）对最新的人脸识别模型进行了分析与归纳，并进行了实验优化。

（3）无监督模型在保证精度的情况下，线下的理论复杂度与增量迭代，使得其可以用于大规模数据集中。并且可以有更多的业务拓展，比如人脸标注、数据清洗和多账号检测等。

（4）方案可移植性好，可复用性强。

比赛链接：

https://www.datafountain.cn/competitions/348

冠军代码：

https://github.com/themostnewone/2019ccf

下面的代码给出了所有图像增强方法与分数最高的 3 种 12 个有监督模型。

代码参考指北的小暇米的框架，在此表示感谢。

https://discussion.datafountain.cn/questions/1904/answers/22795

天池 | 安泰杯冠军法国南部分享

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

在 19 年 9 月下旬结束的"安泰杯"跨境电商智能算法大赛中，来自京东零售的法国南部队伍成功从 1960 支队伍中脱颖而出，在复赛阶段成功逆袭到第一，并通过答辩获得冠军。在接近 2 千只参赛队伍中他们如何取胜，并成功压制住植物的反击，他们获胜方案又有什么可取之处？本文将会给出完整的赛题解析和解题方案介绍，建议阅读全文并进行收藏。

- 赛题介绍

AliExpress 是阿里巴巴海外购物网站，其网站的海外用户可以在 AliExpress 挑选购买自己心意的商品。对于 AliExpress 来说，目前某些国家 A 的用户群体比较成熟，沉淀了大量的该国用户的行为数据。但是还有一些待成熟国家 B 的用户在 AliExpress 上的行为比较稀疏。对于这些国家 B 用户的推荐算法如果单纯不加区分的使用全网用户的行为数据，可能会忽略这些国家用户的一些独特的用户特点。而如果只使用国家 B 的用户的行为数据，由于数据过于稀疏，不具备统计意义，会难以训练出正确的模型。

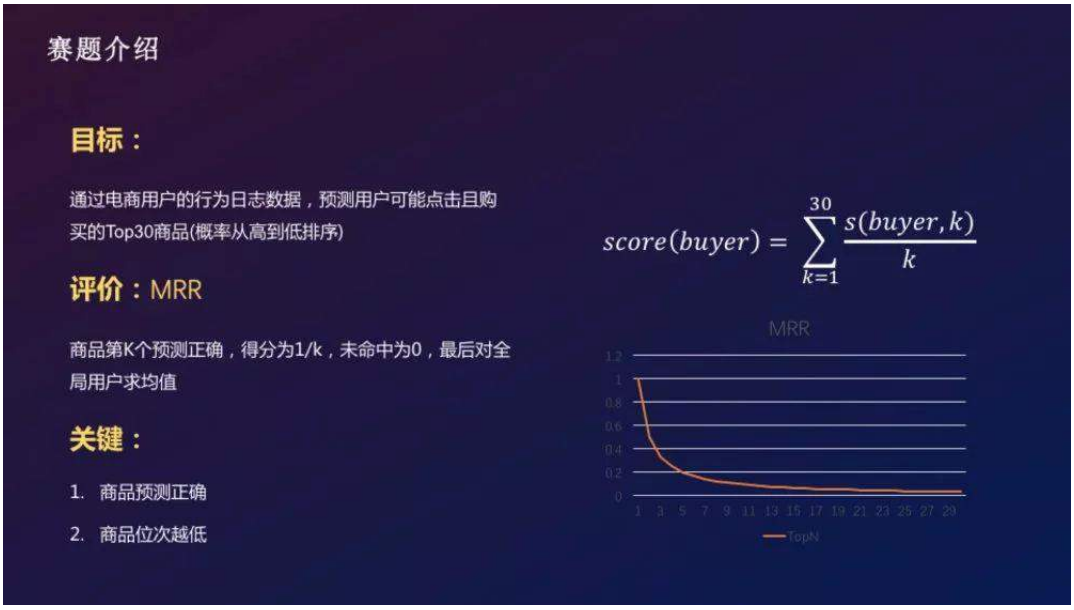
赛题难点是：怎样利用已成熟国家 A 的稠密用户数据和待成熟国家 B 的稀疏用户数据，训练出的正确模型对于国家 B 的用户有很大价值。

赛题数据给出若干日内来自成熟国家的部分用户的行为数据，以及来自待成熟国家的 A 部分用户的行为数据，以及待成熟国家的 B 部分用户的行为数据去除每个用户的最后一条购买数据，让参赛人预测 B 部分用户的最后一条行为数据。

赛题评价指标：赛题旨在通过海量数据挖掘用户下一个可能交互商品，选手们可以提交预测的 TOP30 商品列表，排序越靠前命中得分越高。赛题具体使用 MRR(Mean Reciprocal Rank)对选手提交的表格中的每个用户计算用户得分：

$$score(buyer) = \sum_{k=1}^{30} \frac{s(buyer,k)}{k}$$

其中，如果选手对该 buyer 的预测结果 predict k 命中该 buyer 的最后一条购买数据则 s(buyer,k)=1 s(buyer,k)=1；否则 s(buyer,k)=0 s(buyer,k)=0. 而选手得分为所有这些 score(buyer)的平均值。



- 初赛赛题数据

✓ 商品属性表

数据中共涉及 2840536 个商品，对于其中大部分商品，都会给出该商品的类目 id、店铺 id 以及加密价格，其中价格的加密函数 f(x)为一个单调增函数。

✓ 训练数据

给出 xx 国的用户的购买数据和 yy 国的部分用户的购买数据。

country_id	点击记录数	购买记录数	买家数
zz	4582953	1773429	45795
yy	5241393	1358329	58106
xx	42046596	5241625	511059

✓ 测试数据

给出 yy 国的 B 部分用户的购买数据除掉最后一条。数据的整体统计信息如下：

country_id	点击记录数	购买记录数	买家数
zz	424506	164081	4275
yy	429319	65887	5569

商品属性表、训练数据、测试数据对应的文件：item_attr, train 和 test。无论是训练数据还是测试数据，都具有如下的格式：

buyer_country_id	buyer_admin_id	item_id	log_time	irank	buy_flag
xx	817731	4033525	2018-06-12 07:12:58	1	1
xx	817731	98120	2018-06-11 07:12:58	2	0

其中各字段含义如下：

1. buyer_country_id: 买家国家 id, 有'xx'和'yy'两种取值；
2. buyer_admin_id: 买家 id；
3. item_id: 商品 id；

4. create_order_time: 订单创建时间；

5. irank: 每个买家对应的所有记录按照时间顺序的逆排序；

初赛数据集特点：

1）每个用户有至少 7 条购买数据；

2）测试数据中每个用户的最后一条购买数据所对应的商品一定在训练数据中出现过；

3）少量用户在两个国家有购买记录，评测中忽略这部分记录；

- 复赛赛题数据

在给出若干日内来自某成熟国家 xx 的部分用户的点击购买数据，以及来自某待成熟国家 yy 和待成熟国家 zz 的 A 部分用户的点击购买数据，以及国家 yy 和 zz 的 B 部分用户的截止最后一条购买数据之前的所有点击购买数据，让参赛人预测 B 部分用户的最后一条购买数据。

✓ 商品属性表

点击购买数据中涉及 9136277 个商品，对于其中大多数商品，我们都会给出该商品的类目 id、店铺 id 以及加密价格，其中价格的加密函数 f(x)为一个单调增函数。

✓ 训练数据

给出 xx 国的用户的点击、购买数据和 yy 国、zz 国的 A 部分用户的点击、购买数据。

buyer_country_id	buyer_admin_id	item_id	log_time	irank	buy_flag
xx	817731	4033525	2018-06-12 07:12:58	1	1
xx	817731	98120	2018-06-11 07:12:58	2	0

✓ 测试数据

给出 yy 国、zz 国的 B 部分用户的最后一条购买数据之前的点击购买数据。

buyer_country_id	buyer_admin_id	item_id	log_time	irank	buy_flag
xx	817731	4033525	2018-06-12 07:12:58	1	1
xx	817731	98120	2018-06-11 07:12:58	2	0

无论是训练数据还是测试数据，都具有如下的格式：

buyer_country_id	buyer_admin_id	item_id	log_time	irank	buy_flag
xx	817731	4033525	2018-06-12 07:12:58	1	1
xx	817731	98120	2018-06-11 07:12:58	2	0

其中各字段含义如下：

1. buyer_country_id: 买家国家 id, 只有'xx','yy','zz'三种取值 2. buyer_admin_id: 买家 id

3. item_id: 商品 id

4. log_time: 商品详情页访问时间

5. irank: 每个买家对应的所有记录按照时间顺序的逆排序

6. buy_flag: 当日是否购买

复赛数据集特点：

1）每个用户有若干条点击数据和至少 1 条购买数据 （但测试数据中该条购买记录可能未给出到选手；

2）每个用户的最后一条数据的 buy_flag 一定为 1 （但测试数据中该条数据未给出到选手；

3）测试数据中每个用户的最后一条点击数据（也是购买数据）所对应的商品一定在训练数据中出现过；

4）可能存在少量跨国买家.

- 赛题分析

赛题分析是深入理解赛题的最有效的方法，也是构建有效特征和模型的先驱条件。

根据零售行业的人货场概念，赛题提供了关于用户行为日志的常见字段可分为如下部分：

用户：用户标识、用户国籍

商品：商品标识、店铺、品类、价格

场景：点击时间、访问排序、购买标记

通过对赛题数据进行探索和分析，我们发现可以根据预测商品是否在历史交互过分成两种不同分布的用户：

- 历史交互用户（68%）：即预测商品用户曾经已交互过，在召回-排序阶段：

  召回：可通过 buy_flag=1，将交互商品全量召回

  排序：基于用户商品交互信息，解决排序问题，预测精度高

- 冷启动用户（32%）：即预测商品用户从未交互过，在召回-排序阶段：

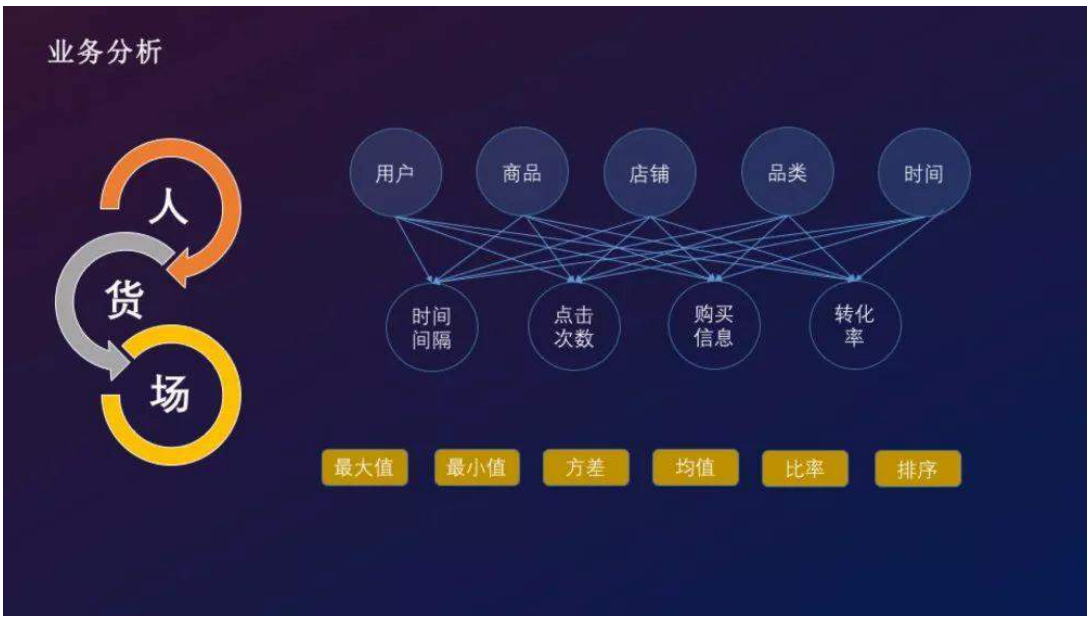
  召回：基于商品关联信息召回，召回难度大

  排序：基于用户最近交互商品与关联信息进行排序，预测精度较低

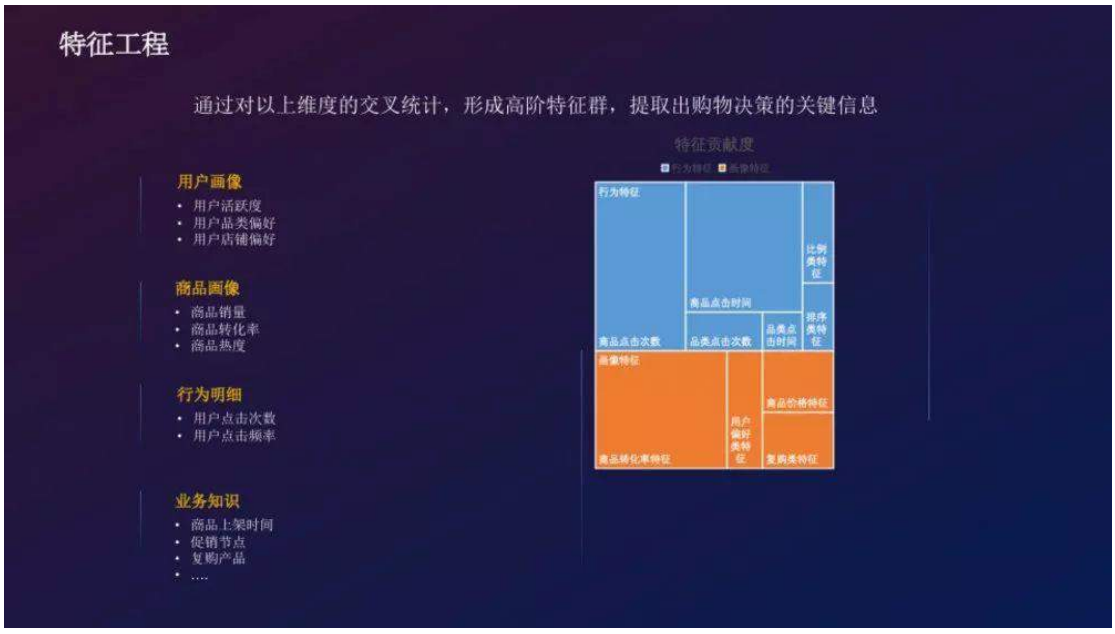
方案思路：面对两种不同分布的用户，我们因地制宜基于不同样本和特征分别建立两个排序模型，然后再通过用户判断模型对两个排序的结果进行优化。

- 特征工程

赛题所给的字段相对简单，主要可分为：用户-商品-场景，我们通过对不同类型因素进行交叉复合，并使用基础统计手段进行计算，构造出高阶特征，提取出购物决策的相关信息：



通过对以上维度的交叉统计，形成高阶特征群，提取出购物决策的关键信息，下图给出了所提取特征的贡献度：



- 构建模型

根据上述的分析我们构建了两个模型：

- ✓ 历史交互商品模型
- ✓ 关联商品模型

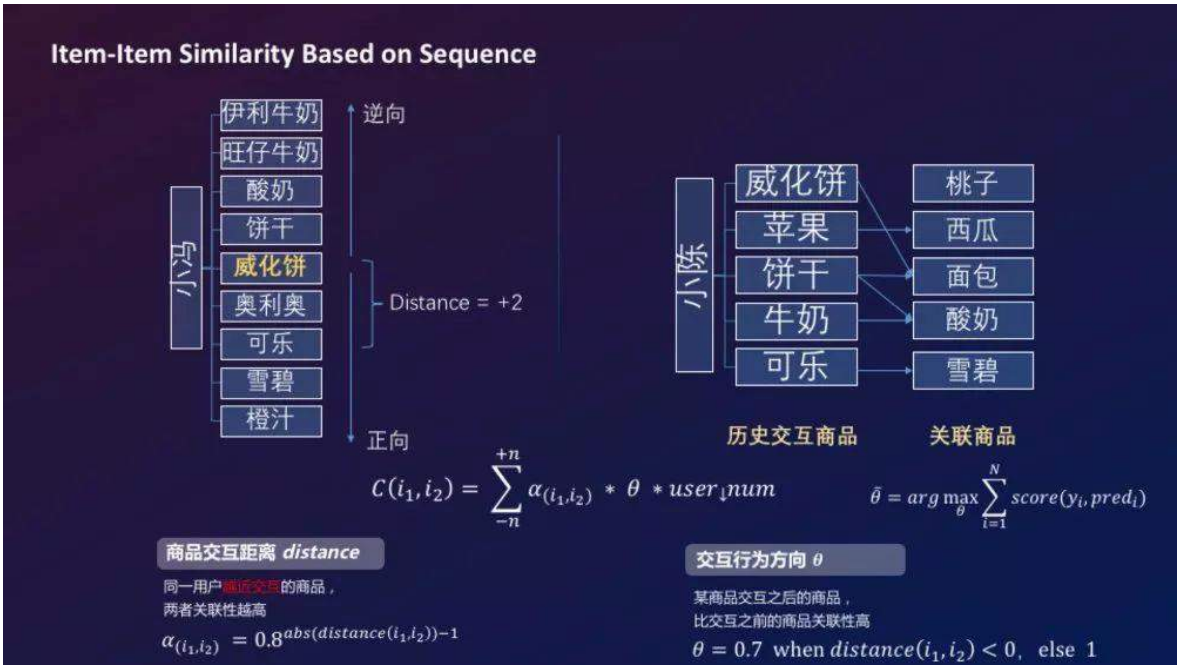
历史交互商品模型

- ✓ 样本构造：提取 buy_flag=1 的 user-item 作为样本，用户最后交互的设为正样本，其他为负样本
- ✓ 模型信息：这里使用的 LightGBM 模型；
- ✓ 样本信息：
  - ✓ 训练集样本数：137W
  - ✓ 训练集用户数：55W
  - ✓ 测试集样本数：3.1W
  - ✓ 测试集用户数：8944
  - ✓ 特征数量：425
  - ✓ Model：LightGBM
  - ✓ loss function：AUC
- ✓ 模型效果：
  - ✓ AUC: 0.9493
  - ✓ MRR: 0.8922
- ✓ Recall Rate 样本召回率：
  - ✓ Top1 item：81%
  - ✓ Top3 item：92.5%
  - ✓ Top10 item：99%

这个时候线上达到 0.6085 的成绩，排名第五，当然目前仅是考虑到了历史有过交互的商品，接下来将建立关联商品模型。

关联商品模型

那么我们如何找到用户未来可能交互的商品？比较好的方法是挖掘关联商品，根据用户历史交互商品，找到这些商品的关联品。



- ✓ Item-Item similarity Based on Sequnence
  商品相似性：基于用户行为序列计算，假设用户越近交互的两个商品相似性越高，并且考虑先后次序，通过线性搜索得到如下相似度计算公式：
- ✓ 样本构造：对用户最近 5 个交互商品的关联商品(加上时间衰减权重)，选取每个用户 TOP50 关联商品，之前得到的关联度中间结果直接作为特征训练排序模型
- ✓ 模型信息：
  - ✓ 训练集样本数：940W

- ✓ 训练集用户数：18W
- ✓ 测试集样本数：47W
- ✓ 测试集用户数：8944
- ✓ 特征数量：222
- ✓ Model：LightGBM
- ✓ loss function：AUC

✓ 模型效果：
AUC: 0.9736
MRR: 0.0554

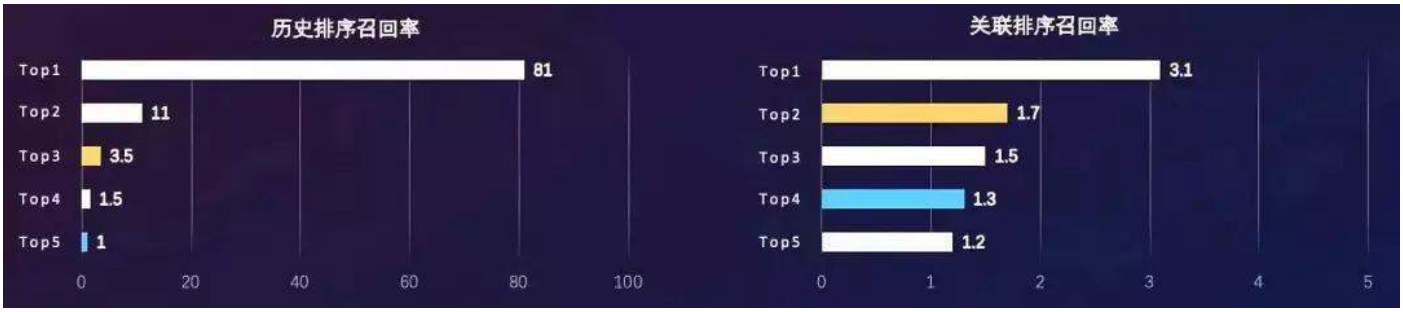
✓ Rcall Rate：

- ✓ Top1 item：3.1%
- ✓ Top3 item：6.3%
- ✓ Top30 item：15%

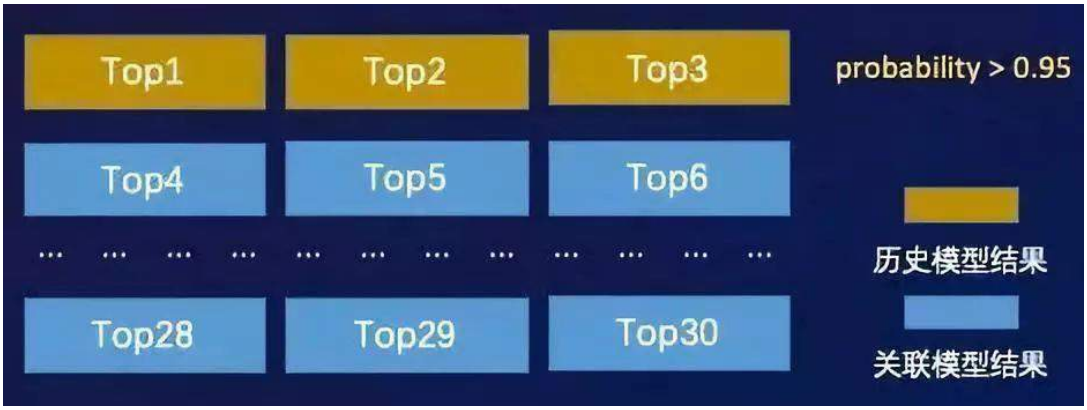
经过关联商品模型解决冷启动的问题，我们的成绩也由 0.6085 提高到 0.6187，排名提升到第一。

- 模型融合/优化

排序优化



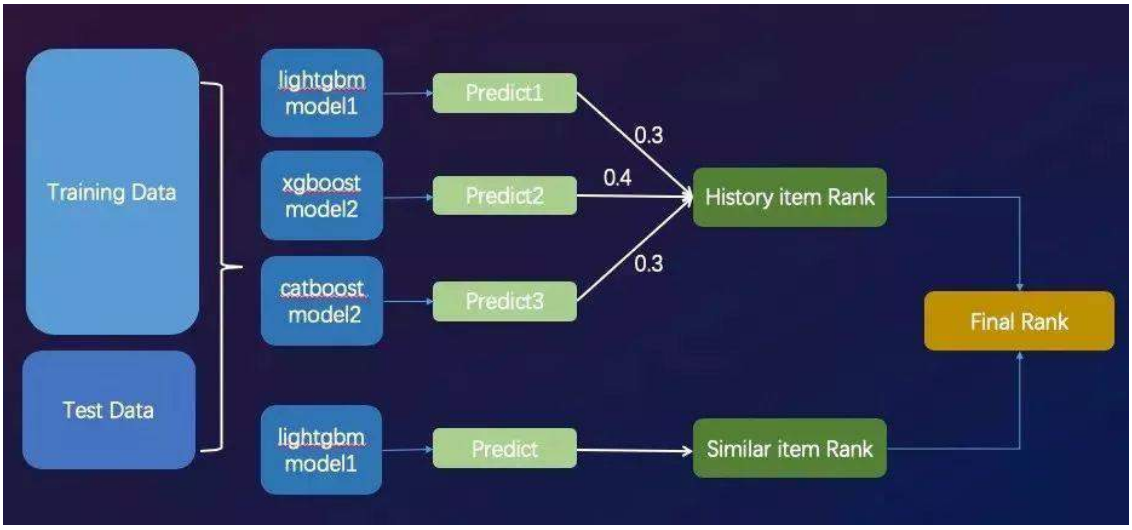
如图所示，历史商品模型排序第 4 的商品召回率仅有 1.5，而关联模型排序第一位召回率为 3.1。为了优化排序结果，优化两部分模型的结果，通过用户判别模型(预测用户是否为冷启动用户)，对概率大于 0.95 的高置信度用户直接截取掉历史 TOP3 后，与商品与关联模型的结果进行拼接，得到最终的 Top30 商品排序。



模型融合

✓ Ensemble

对于历史交互商品模型，训练了 LightGBM、Xgboost、CatBoost 三个模型，通过对预测结果简单加权进行融合。



✓ Stacking

此外，在每个模型训练过程中通过简单 Stacking 将其他 4 折的预测结果作为特征反喂模型，进一步拟合结果。通过排序优化和 stacking 后，我们将分数从 0.6198 提升到 0.6256，进一步拉开了与其他队伍的差距。

排名	参与者	组织	score	最优成绩提交日
1	法国南部	rain	0.6256	2019-09-16
2	江离	江离数据挖掘俱乐部	0.6185	2019-09-15
3	聪明甜甜傻歪歪	甜甜圈咔机	0.6128	2019-09-11
4	汤庄火箭队	字节跳动	0.6091	2019-09-15
5	#问题小哥哥#	IBM	0.6082	2019-09-15

- 比赛总结

在本次比赛前期，我花费了大量精力进行数据探索和分析，基于对数据和业务的了解才确立了最终的方案和优化路线。当对数据之间的联系了然于心后，开始进行细致的特征工程以提取各种信息，在也是历史交互模型得分提升的关键。

此后，为了提高召回率，尝试了 Embedding、协同过滤等方法，但是由于数据量和 category 字段少的限制，都没取得太好的效果，开始基于业务理解，尝试建立关联度计算公式，通过不断搜索参数，取得了不错的召回率，由此建立关联商品模型，此时成绩也上升到第一名。

最后阶段，我们开始提高模型的精度和稳定性，一是建立了用户判别模型对排序进行了优化，二是对模型进行了融合以及 stacking，得到了 0.6256 分数，进一步扩大了领先优势。

取胜关键 = 充分理解业务 + 完备的特征工程 + 合理建模方法 + 细致结果优化 + 坚持就是胜利！

比赛 PPT 和代码分享关注我们的开源项目：

https://github.com/datawhalechina/competition-baseline

优秀的数据挖掘比赛如何定义？

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

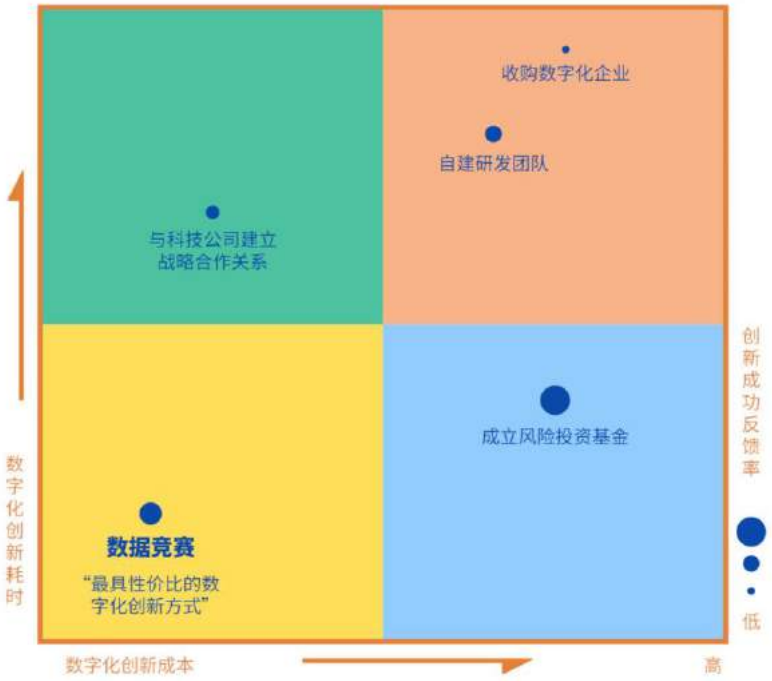
数据竞赛的创新价值均以数据为基础，除了以丰厚的奖金投入来 吸引优质人才的加入，越来越多的主办方也在数据安全与法规限 定的范畴内不断加大竞赛数据的投入力度，为创新提供更充足的 养料。

本文主以一场数据挖掘比赛为例，从几个点来评价数据挖掘比赛的优劣。顺便总结下数据挖掘比赛中需要注意的地方，从出题步骤到最终的比赛评测。

01 找准比赛目标

几年来随着数据挖掘比赛的风潮，很多互联网公司都办起了比赛。但很多公司在办比比赛之前并没有考虑到自己举办比赛的目的，而只是更风凑热闹。这种方式举办的比赛，自然质量不高，企业自身不重视。

当然数据竞赛也有它相应的有点，据统计数据竞赛是创新成本中较低的创新方式：



不同数字化创新方式的成本、耗时和成功率对比，来源和鲸科技

总体说来，企业举办数据挖掘比赛，无外乎以下几个目标：

1. 用比赛来宣传自己公司，宣称公司的技术；
2. 用比赛来吸引专项人才，打造公司形象；
3. 根据上级领导思想，或者被安排举办比赛；

各类赛事的难点，来源和鲸科技

上述的第三条其实在国内还是比较常见，比如在一些政府或者教育相关组织的比赛，被领导安排来做一场比赛。

主办方	赛题内容	奖金池(元)	参赛团队(支)
京东数科	借助电信用户地理位置信息预测各个区县未来15天总人口变化情况	2,200,000	2,038
中国联通研究院	面向电信行业存量用户的智能套餐个性化匹配模型	1,000,000	2,546
中国联通	根据用户的历史手机使用行为，预测其未来3个月的换机概率	100,000	768
中国移动研究院	基于文本内容识别垃圾短信	40,000	378

工业制造场景的数据竞赛典型赛题，来源和鲸科技

02 找准比赛的题目和立意

一个比赛注重要的就是比赛的立意，所以一个比赛的背景和所提供的数据也是要相符合的。其次比赛的题目最好与举办方公司内部需要解决的业务问题相符合，是举办方需要解决的问题，或者急需解决的问题。

交通出行领域数据竞赛赛题关键词，来源和鲸科技

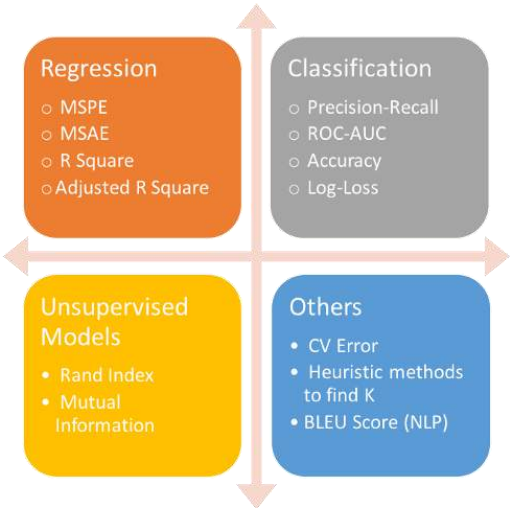
国内的数据挖掘比赛，其实进一步可细分为算法比赛和创意创业比赛。前者算法赛题可以通过具体的评价指标来评价，而后者则难以定量评价。所以在比赛出题阶段，就需要找准比赛的立意和目的，且找到一个合适的评价标准。

03 比赛数据和评价指标

数据挖掘比赛，核心就是数据。而对于一个合格的数据挖掘比赛，赛题数据又需要从各个方面进行考量：

1. 数据是否公开：数据是非常宝贵和具有商业机密的，举办方应该考虑如何选择比赛数据；
2. 数据字段是否匿名：数据字段是够包含敏感信息，是否需要匿名处理；
3. 数据是否规整：赛题数据是够完备，是否存在缺失情况；数据是否规整，是否还需要选手进行清洗；
4. 数据标签是否完备：数据训练标签是否完备，是否需要选手进行标注等；
5. 数据量是否足够大：数据量是否足够满足数据挖掘的要求；

当然评价指标也至关重要，评价指标应该契合赛题的目标，符合赛题的任务，并能够定量评判选手的模型精度。



04 明确比赛规则

比赛既然是比赛，就需要考虑如何进行评价，也应该明确一些比赛规则。

- ✓ 评价指标：在进行比赛立意和准备工程中，就需要考虑如何来设计评价指标。评价指标也应该和具体的赛题目标进行符合，例如分类问题可以使用分类正确率或者信息熵，二分类问题可以使用 AUC 或 KS 指标，回归问题可以用均方差等。评价指标应该具体明确，最好可以量化可解释；
- ✓ 数据规则：比赛在数据使用上应该考虑，比赛数据是否可分享、是否允许使用外部数据（模型）等；
- ✓ 初赛和复赛：一般的比赛可分为初赛和复赛两个阶段，并通过最终的复赛来决定最终的排名。但在这两个阶段需要注意的是，不要轻易改变比赛的评价指标，也不要轻易改变数据的整体分布，同时也要明确最终的排名的规则；

05 简化交流机制

比赛在举办工程中，选手肯定会对比赛的数据和规则等存在很多疑问，这就需要举办方进行答疑。所以在准备比赛过程中，竟可能提前把选手会遇到的问题想清楚，通过注意事项等方式进行表达告知，通过比赛规则提前明确提高的方式（包含如何提交、提交的格式）和合并队伍的方法

优质数据科学竞赛的关键要素，来源和鲸科技

现在在国内外存在很多数据挖掘平台，举办方可以选择与这些平台进行合作，可以得到更多曝光度和推广。而且此类比赛平台有历史举办比赛的经验，可以协助更好进行比赛。但此时需要注意的是，如果和此类平台合作，则需要明确双方具体的责任，和具体应急相应的流程，以免耽误比赛进度。

天池 | 心电异常事件预测冠军解决方案

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

- 赛题背景

心电图是医院心脏疾病常用辅助诊断指标。心电图由于其价格低、无创的特性被广泛用于心脏疾病的预筛查以及体检中，每天的检测量巨大。目前，多导联的心电图设备已经广泛用于临床当中，部分设备已经具有自动分析诊断功能，但自动分析对于多心电异常事件的判别还不够精确，需要医生做进一步修改。

近年来，人工智能在心电图预测领域有了应用。AI 技术、深度学习的发展有望助力心电图波形、心电异常事件的预测，从而达到提升预测精度的目标。

本次大赛要求选手以心电图异常事件预测为赛题方向，依据心电图机 8 导联的数据，以及病患年龄、性别等因素，用统计学、机器学习、深度学习等方式探索挖掘心电波形与心电异常事件之间的关系，构建精准预测模型。

本文为冠军“随便丨逛逛”的分享，PPT 和比赛分享。

- 赛题回顾

大赛包含有杭州师范大学移动健康管理系统教育部工程研究中心提供的 4 万个医疗心电样本。每个样本有 8 个导联，分别是 I，II，V1，V2，V3，V4，V5 和 V6。

问题描述：依据心电图机 8 导联的数据和年龄、性别特征，预测心电异常事件

比赛数据：32142 条初赛数据和 20036 条复赛数据（初赛数据有重复，初复赛标签分布差异大）选手提交结果与实际检测到的心电事件结果进行对比，以 F1 为评价指标，结果越大越好，具体计算公式如下：

$$F1 = \frac{2 * P * R}{P + R}$$

$$P = \frac{\text{预测正确的心电异常事件数}}{\text{预测的心电异常事件数}}$$

$$R = \frac{\text{预测正确的心电异常事件数}}{\text{总心电异常事件数}}$$

- 数据分析

数据分析与处理

1. 初赛数据存在大量重复

处理方案：对初赛数据去重

2. 初赛数据标签有55类，而复赛数据标签仅有34类，且部分标签对应样本极少，或难以训练

处理方案：综合考虑标签对应样本数量与模型验证效果，最终选取20类标签训练模型

数据分析与处理

4. 初赛数据存在标签问题

4.1 部分数据同时出现**窦性心律**和**窦性心律不齐**

处理方案：删除这些数据

4.2 部分数据标记为**不完全性右束支传导阻滞**或**完全性右束支传导阻滞**，却没标记为**右束支传导阻滞**

处理方案：人工增加标签

4.3 部分数据标记为**完全性左束支传导阻滞**，却没标记为**左束支传导阻滞**

处理方案：人工增加标签

数据分析与处理

5. 初赛数据与复赛数据标签分布差异过大

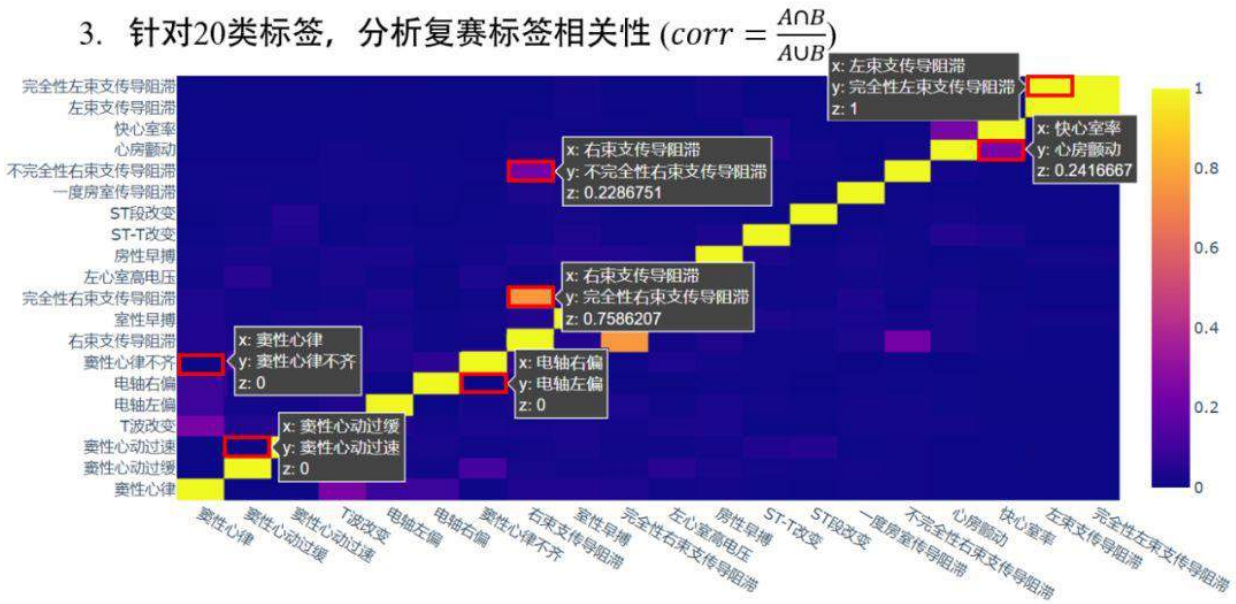
处理方案：

1. 先用初赛数据预训练模型
2. 再在预训练模型基础上用复赛数据微调
3. 最终对1和2的模型进行融合

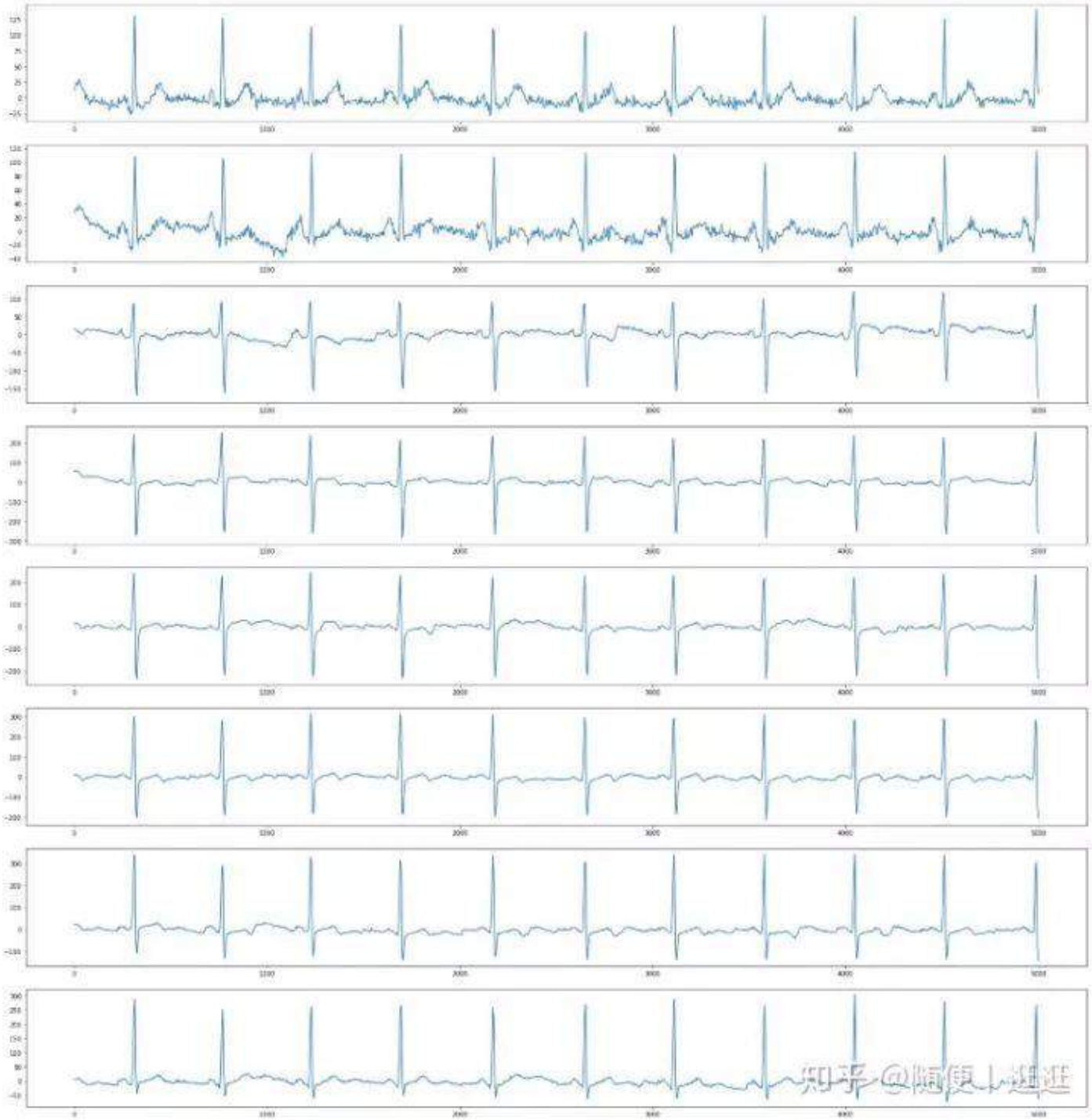
6. 评价指标为MicroF1

处理方案：训练时对所有20种类别采用同样的权重

标签相关性：计算公式为两标签交集数量除以两标签并集数量，0 表示完全互斥。该相关性也可视为一种“特征”。



不同导联节拍一致：尖峰位置一致。



不同导联十分相似：将不同导联画在同一坐标轴上，可看出相似性。如何构建模型以利用这种相似性是最为关键的思路。

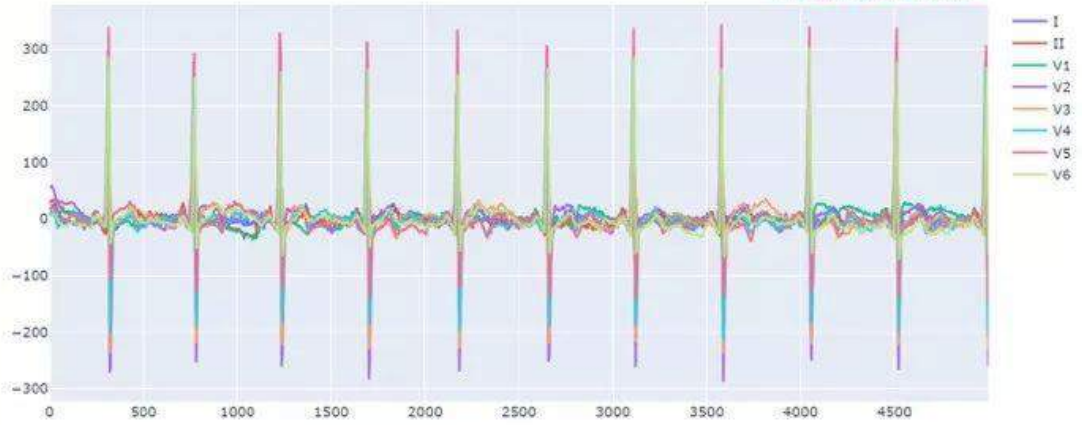
- 模型构建与融合

针对多导联心电图分类任务，我创新地提出一种网络结构，将其称之为 ECGNet: Multi-scale ResNet for Multi-lead ECG Data。该模型是本次比赛的致胜关键。（细节可见 PPT）

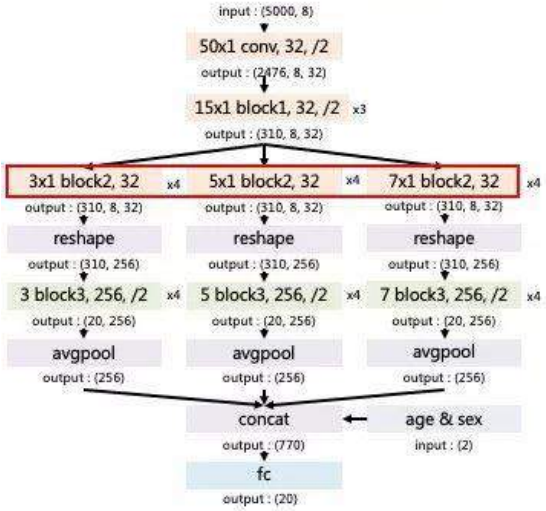
数据分析与处理

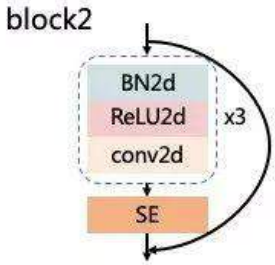
★ 9. 观察8导联图像，发现不同导联之间存在相似性

在模型构建时利用不同导联间的相似性

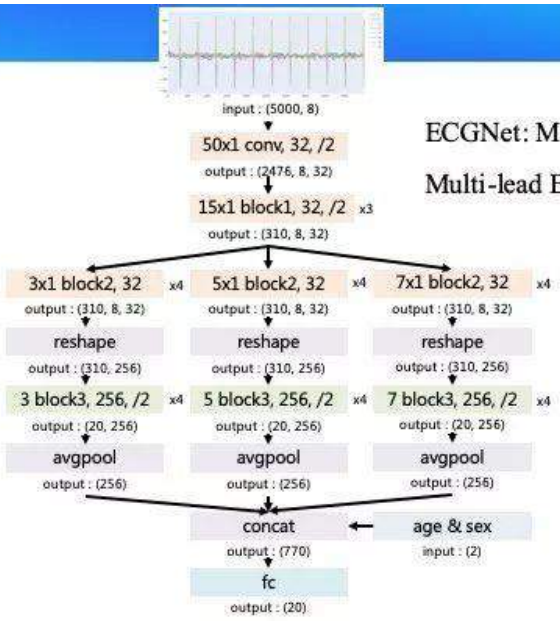


模型构建——block2



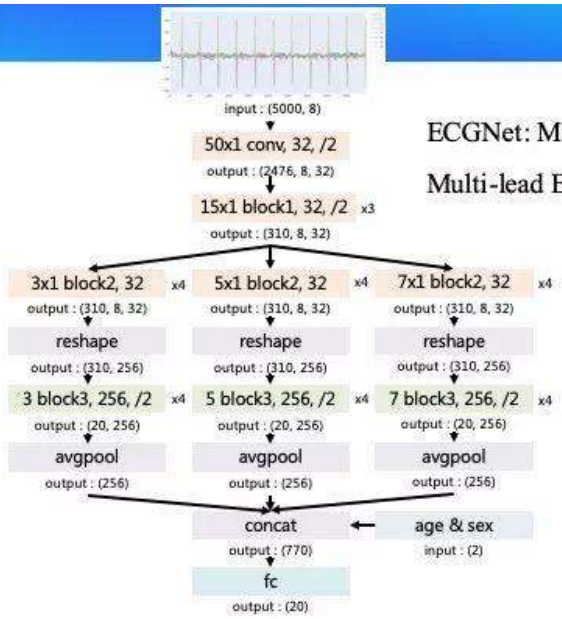


模型构建——总览



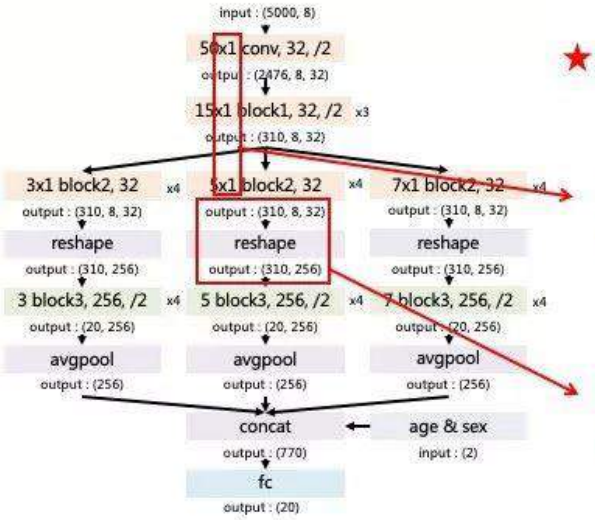
ECGNet: Multi-scale ResNet for Multi-lead ECG Data

模型构建——总览



ECGNet: Multi-scale ResNet for Multi-lead ECG Data

模型构建——利用导联相似性



★ 如何利用不同导联间的相似性？

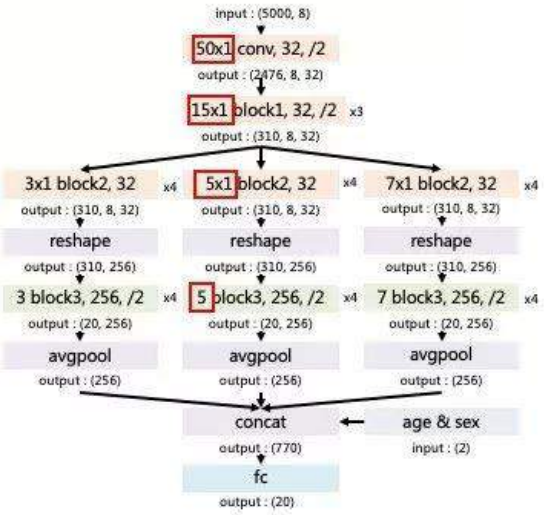
初期：将一维的心电图数据当作二维数据处理，在不同导联采用相同的卷积核。

作用:A.提取不同导联的共性特征

B.减小参数量

后期：将不同导联提取的特征合并，重新转变为一维数据

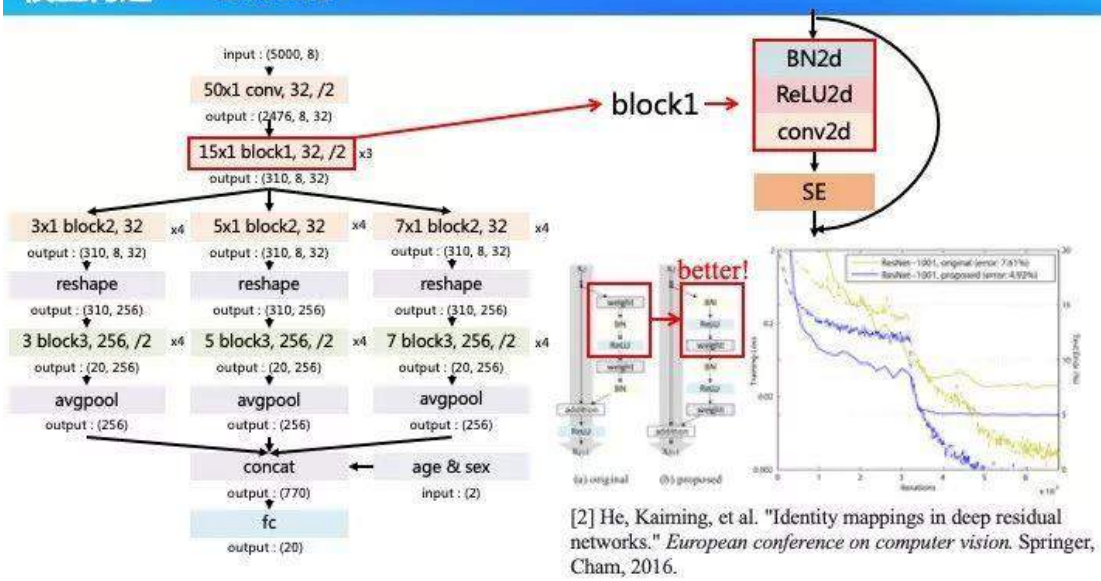
模型构建——卷积核长度



1. 处理ECG信号时，前期采用较长的卷积核效果较好[1]。
2. 随着特征图尺寸的减小，可以减小卷积核长度（同时减小运算量）

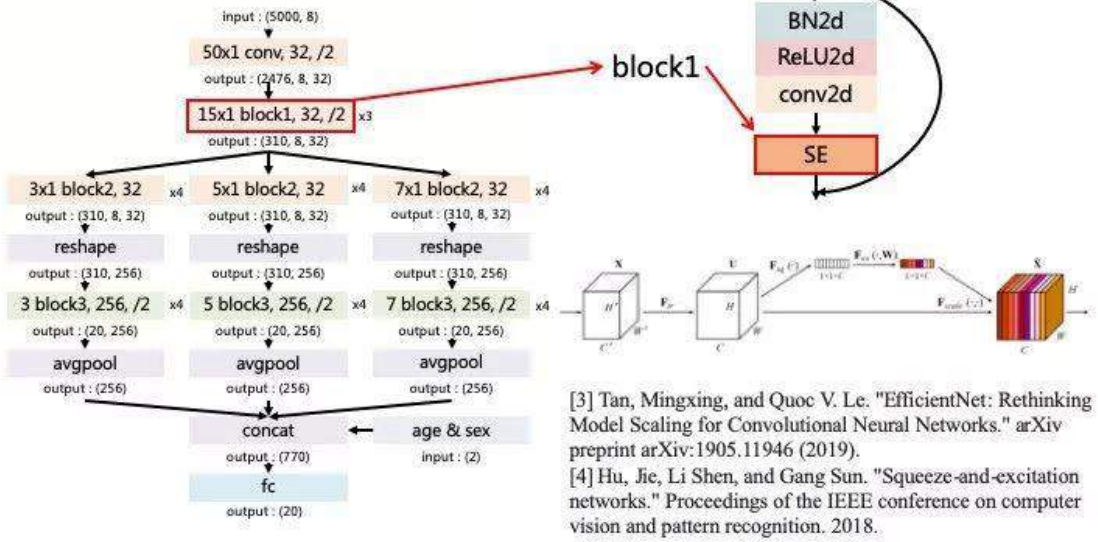
[1] Hannun, Awni Y., et al. "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network." Nature medicine 25.1 (2019): 65.

模型构建——block1



[2] He, Kaiming, et al. "Identity mappings in deep residual networks." *European conference on computer vision*. Springer, Cham, 2016.

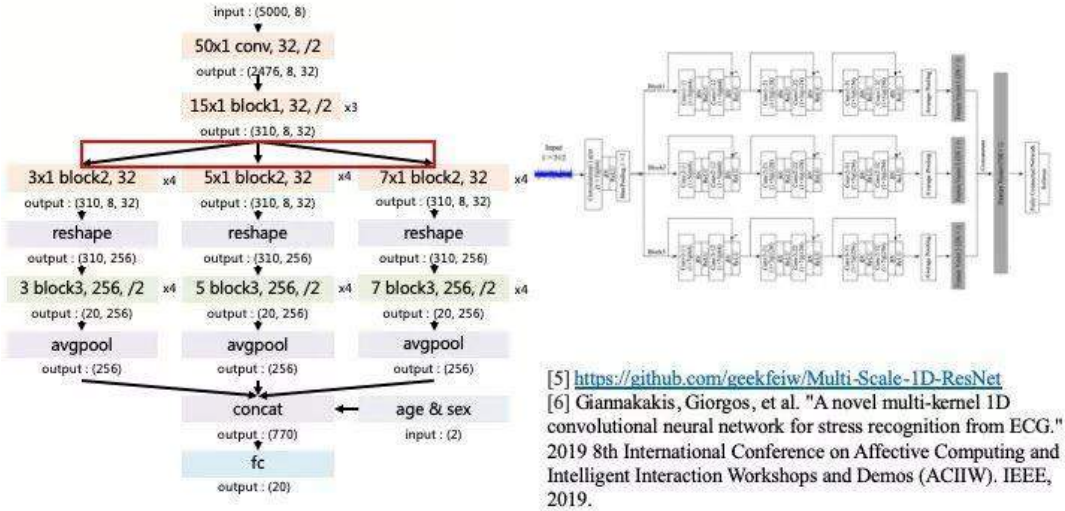
模型构建——block1



[3] Tan, Mingxing, and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks." arXiv preprint arXiv:1905.11946 (2019).

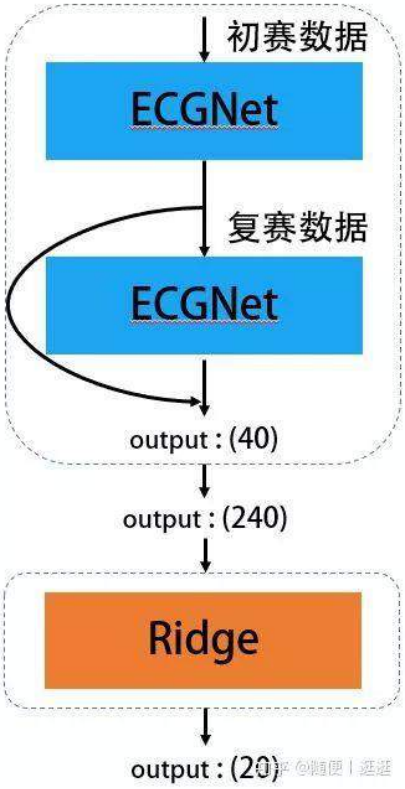
[4] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

模型构建——多尺度网络



[5] https://github.com/geekfeiw/Multi-Scale-1D-ResNet

[6] Giannakakis, Giorgos, et al. "A novel multi-kernel 1D convolutional neural network for stress recognition from ECG." 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2019.



模型融合阶段效果提升，我认为主要有两点原因：

✓ 充分利用初赛和复赛的数据

✓ “隐含”地利用了不同标签的相关性

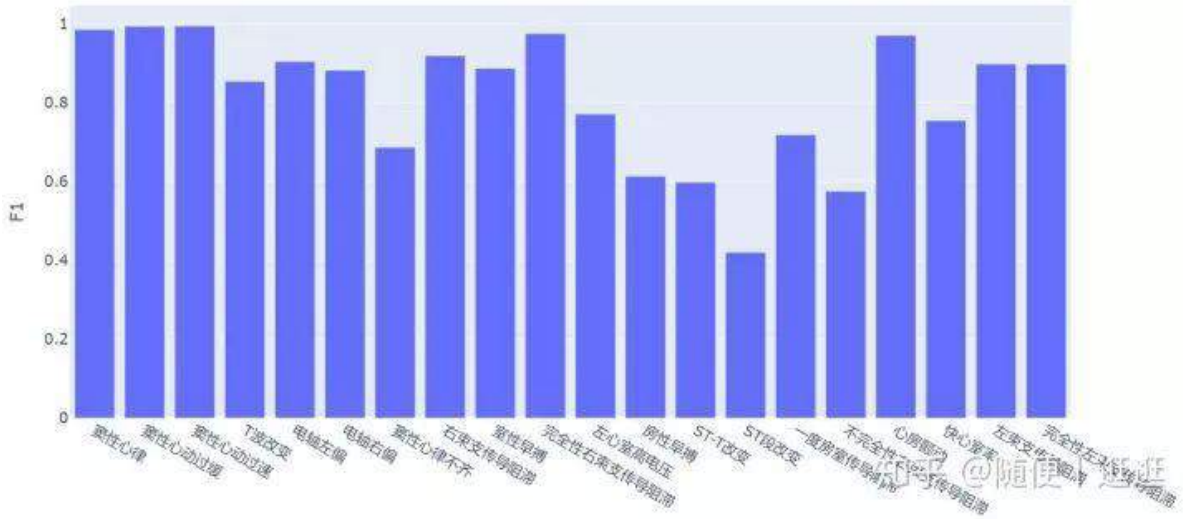
不足之处：

✓ 模型缺乏多样性

✓ 没有用到传统特征和树模型

此外，植物提出的嫁接学习也是种很有意思的思路。

线下对 20 类标签用单模型评估效果，反正比我自己判断的（仅限电轴偏转方向）好得多。



- 赛后总结

✓ 充分利用提供的数据很重要，尤其分布差异很大时。

✓ 多思考多尝试很重要。

在处理多导联心电图数据时：

✓ 对不同导联应用相同的卷积核，能在减小参数量的同时，很好地提升模型的效果。

✓ 采用多尺度网络能捕捉不同尺度的特征，较好地提升模型的效果。

✓ 网络初期可以采用较大的卷积核，后期可以采用较小的卷积核。

✓ BN-ReLU-Conv 要优于 Conv-BN-ReLU。

✓ Squeeze-and-excitation 结构也能提升模型的效果。

✓ 传统特征也很重要。

失败尝试，以下是一些失败的尝试（不代表这些方法真的不行，也许是我的打开方式有问题）：

✓ 傅里叶变换

✓ 小波变换

✓ 频谱图

✓ DenseNet

✓ EfficientNet

✓ Attention

✓ LSTM

✓ 窗口

✓ 各种去噪

✓ 特征工程

总之还是要多思考多去尝试吧，没什么事能一帆风顺的。

- 赛后感想

本次比赛收获颇丰，除了实质性奖励以外，还锻炼了我赛题思考、数据分析、模型构建、论文阅读、编程实现以及答辩的能力，且与其他选手交流了一些有趣的思路。

客观上来看，本次比赛有很大的运气成分的，自己还有很多不足。

总之，继续努力，再接再厉吧。（单人参赛好累啊，还有复现阶段需要抓紧时间，我差点没在期限内整出来。）很多细节我没有详细介绍，有兴趣地可以看 PPT 或答辩视频。

Data Competition in 2019

PPT 和比赛分享：

https://github.com/RandomWalk-xzq/Hefei_ECG_TOP1

DF | 技术需求与技术成果关联度冠军分享

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

CCF BDCI “技术需求”与“技术成果”项目之间关联度计算比赛由中国计算机学会和八六三软件发起，赛题目的是发现好的方法、算法或模型，并提供用于验证的程序源代码，可应用于平台模拟人工，实现“需求——成果智能匹配服务”。

文本将介绍冠军团队"莽就完事了"的赛题分享，冠军团队由马凯欣一人组成。凯欣来自东北林业大学，本文将包括凯欣的参赛方案分享和代码分享。

> *我叫马凯欣，来自东北林业大学，目前是计算机技术专业一年级在读，专业方向是自然语言处理。我之前没有参加过大数据与人工智能的相关比赛，这次是我的头一次，CCF BDCI 的赛题我也不知道挑哪一个好。*
> *说到参赛经历我倒是有一些，本科参加过 ACM-ICPC 竞赛，也取得过一些奖项。*
>
> *by 马凯欣*

- 赛题和数据介绍

人工判断技术需求和技术成果关联度的方法是：从事技术转移工作的专职工作人员，阅读技术需求文本和技术成果文本，根据个人经验予以标注。

“技术需求”与“技术成果”项目之间关联度计算模型技术需求与技术成果之间的关联度分为四个层级：强相关、较强相关、弱相关、无相关。

文件名称	说明
DataSet.zip	包含两个csv文件， 其中Achievements为技术成果表，Requirements为技术需求表。
Train_Label.csv	为关系表，标注了部分Achievements技术成果表跟Requirements技术需求表的关联关系。
Test.csv	为待预测的关系表，需要预测技术成果表跟技术需求表的关联关系。

Achievements 为技术成果表，文件中的（.csv）（UTF-8 编码）一行对应于一个技术成果，以“，”分割不同的识别字段，具体描述具体格式如下：

字段信息	类型	描述
Guid	string	ID列
Title	string	技术成果的标题
Content	string	技术成果的具体内容

Requirements 为技术需求表，文件中的（.csv）（UTF-8 编码）一行对应于一个技术需求，以“，”分割不同的识别字段，具体描述具体格式如下：

字段信息	类型	描述
Guid	string	ID列
Title	string	技术需求的标题
Content	string	技术需求的具体内容

Train_Label.csv 是 Achievements 技术成果表跟 Requirements 技术需求表的关联关系：

字段信息	类型	描述
Guid	string	ID列
Aid	string	技术成果的表的ID
Rid	string	技术需求的表的ID
Level	string	1无相关、2弱相关、3较强相关、4强相关

- 评测方法

本次竞赛初赛评价指标使用 MAE 系数。平均绝对差值是用来衡量模型预测结果对标准结果的接近程度一种衡量方法。计算方法如下：

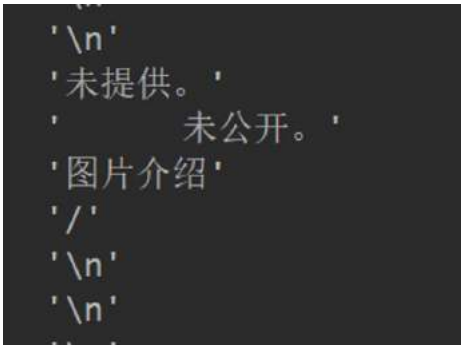
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|pred_i - y_i|$$

其中 pred_i 为预测样本，y_i 为真实样本。MAE 的值越小，说明预测数据与真实数据越接近。最终结果为:

$$Score = \frac{1}{1 + MAE}$$

- 数据清洗和数据增广

经过对赛题文本的内容进行筛选查看，发现其中存在一定量的空白、"\n"、"未提供"等无效信息。简单的使用对应标题对无用信息进行替换即可。



对问题进一步化简，可以简化成两个文本之间的关联度计算。

1. 那么 A 文本与 B 文本之间关联度，同样也是 B 文本与 A 文本之间关联度。该方法在仅取标题时可以提升成绩。当加入内容时会造成过拟合，最终未采用该方法。
2. 那么假设 A 文本与 B 文本之间关联度为 4，A 文本与 C 文本之间关联度为 3，那么可以假定 B 文本与 C 文本之间关联度为 3，按照这个思路可以假设关联矩阵

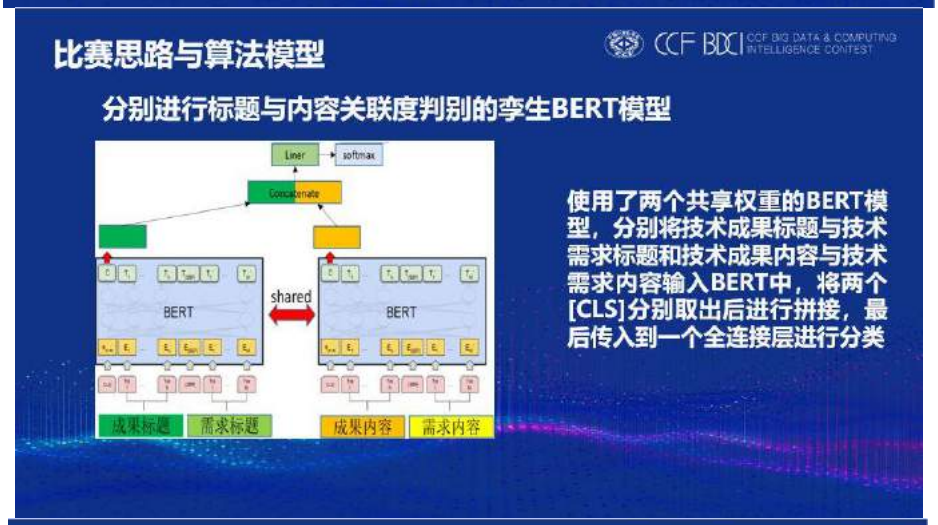
$$R = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

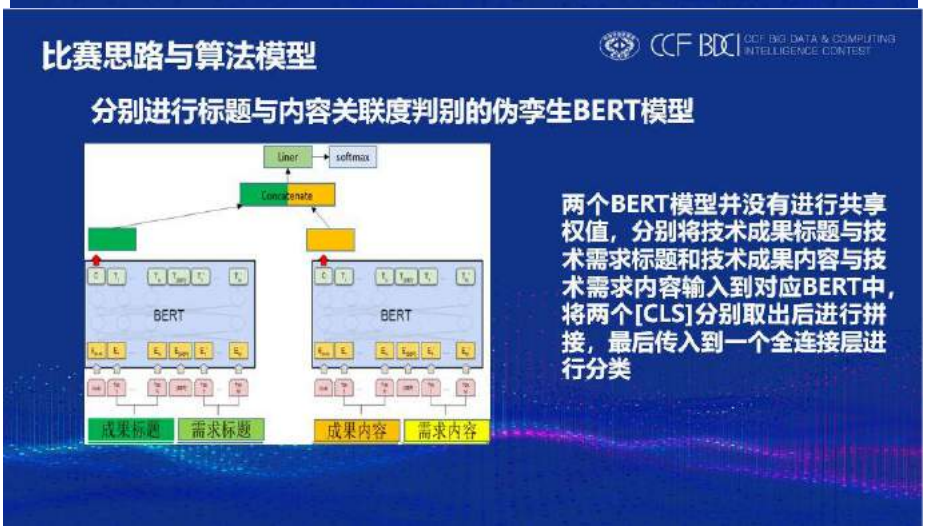
其中 A 文本与 B 文本之间关联度为 i，A 文本与 C 文本之间关联度为 j，那么 B 文本与 C 文本之间关联度为 R_(i,j)。此方法可增加数据 295994 条，从中按照原数据集各个关联度等级的比例从中随机取出 10000 条。

该方法我认为具有一定的可能性，但由于训练时间过长、提交次数有限，尝试过的参数均会造成过拟合现象。最终模型中未对数据进行数据增广。

- 构建模型

经过一定量的实验对比最终的模型确定为分别进行标题与内容关联度判别的孪生 BERT 模型，其中进行技术成果标题和技术需求标题关联度计算的 BERT 采用谷歌开源的 BERT-base；进行技术成果内容与技术需求内容关联度计算的 BERT 采用哈工大提出的 BERT-wmm。





最终只采用这个模型，也没有进行融合。当然可以很简单的认为它就是分别使用两个 BERT 进行相似度判别然后进行拼接。

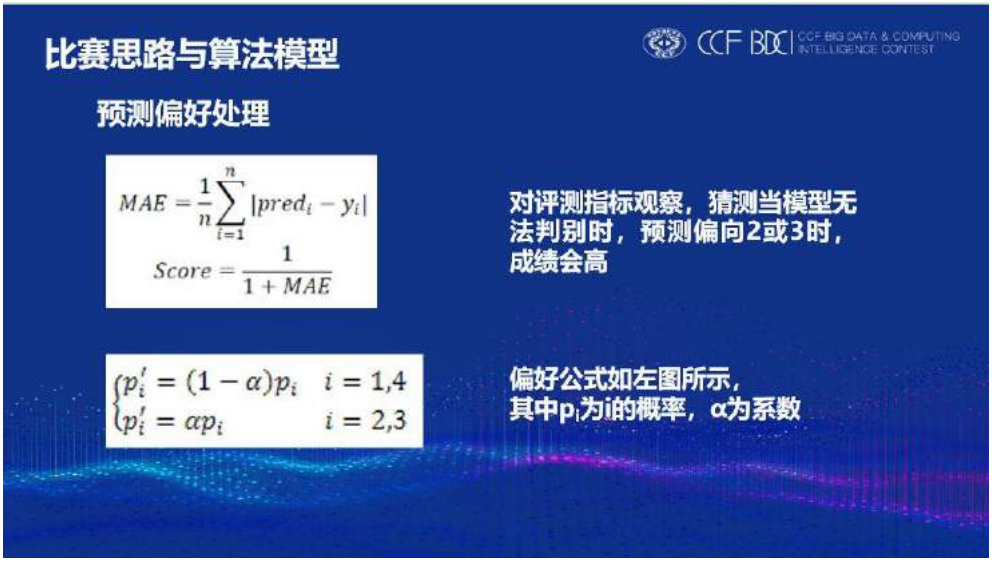
其中进行技术成果标题与技术需求标题关联度计算的 BERT 采用谷歌开源的 BERT-base；进行技术成果内容与技术需求内容关联度计算的 BERT 采用哈工大讯飞联合实验室发布基于全词覆盖的 BERT-WWM。该预训练由于采用了全词覆盖，在多数情况下可以取得更好的效果。

在第一个进行技术成果标题与技术需求标题关联度计算的 BERT 中输入最大长度 MaxLenT 设置为 128，两个标题拼接最大长度也没有超过 128 个字，同时这样可以减少训练时间和显存需求；在第二个进行技术成果内容与技术需求内容关联度计算的 BERT-WWM 中输入最大长度 MaxLenC 设置为 512，尽可能多的读取数据内容。

两个 BERT 都采用 12layers, 768hidden states, 12heads 版本，该模型采用 7 折交叉验证，其中 batch size 取 16，epoch 取 8，并在训练时保存较好的模型权值，初始学习率设置成 5e-5，后续学习率设置成 1e-5。

- 预测后处理

通过观测评测指标发现，当模型判断关联度为 1 和 2 的概率非常接近时，输出为 2 更加合理。所以当模型无法判别时，通过修正可以将输出偏向 2 或 3。



- 模型对比

序号	模型	初赛成绩	复赛成绩
1	BERT-base	0.78585178	0.79595751
2	RoBERTa-base	0.78077936	/
3	孪生BERT-1	0.78604090	0.79607499
4	孪生BERT-2	0.78509617	0.79843128
5	BERT+数据增广-1	/	0.80163449
6	BERT+数据增广-2	/	0.77996242
7	BERT+数据增广-3	/	0.79548806
8	BERT-T128C512	0.79079902	0.79866767
9	BERT-WWM-T128C512	0.79099053	0.80008900
10	最终模型	0.79175758	0.80642748

1.其中 BERT-base、RoBERTa-base、BERT+数据增广-1、BERT+数据增广-2、BERT+数据增广-3 模型中输入均只有技术成果标题与技术需求标题，MaxLenT 为 128，其余超参数与最终模型中基本相同。

2.孪生 BERT-1 模型为标题与内容拼接的孪生 BERT 模型，MaxLen 为 512，其余超参数与最终模型中基本相同。

3.孪生 BERT-2 模型为分别进行标题与内容关联度判别的孪生 BERT 模型，MaxLen 为 512，其余超参数与最终模型中基本相同。

4.BERT+数据增广-1 模型中，数据增广采用第一种方式。

5.BERT+数据增广-2 模型中，数据增广采用第二种方式，且取全部增广数据。

6.BERT+数据增广-3 模型中，数据增广采用第二种方式，但按照原数据集各个关联度等级的比例从中随机取出。

7.BERT-T128C512 模型中 BERT 均采用谷歌发布的 BERT-base，其余超参数与最终模型中相同。

8.BERT-WWM-T128C512 模型中 BERT 均采用采用哈工大讯飞联合实验室发布的 BERT-WWM，其余超参数与最终模型中相同。

9.最终模型中标题采用谷歌发布的 BERT-base，内容采用哈工大讯飞联合实验室发布的 BERT-WWM。

- 比赛总结

我个人认为 BERT-WWM 预训练相比于 BERT 预训练对中文效果应该更好，而得到这样的结果，可能的原因是两个预训练在训练时使用的语料库不同，标题部分中专业名词比重较大且短小，BERT 对此比较敏感，而 BERT-WWM 对常规文本比较敏感。当然这个成绩中也有预测偏好处理的功劳。

本次比赛十分感谢华南理工大学 Chevalier 同学在知乎上分享的 BaseLine。本代码修改于该代码。由于刚开始接触深度学习，也是头一次参加比赛，本人水平有限欢迎批评指正。邮箱：1239977613@qq.com

冠军代码分享：

https://github.com/Makaixin/Correlation-between-requirements-and-achievements

DF | 工件负荷率预测冠军分享

微信公众号：Coggle 数据科学

Coggle 全称 Communication For Kaggle，专注数据科学领域竞赛相关资讯分享。

团队“突然 ping 通”由 5 小伙伴组成，其中队长黄超为华东师范大学软件工程 2018 级研究生。本文将分享他们团队在离散工件质量负荷率预测比赛中得分 Top1 的分享，代码已经开源。

为了解决离散工件质量检测问题，我们首先对赛题数据进行了深入探索分析，得到 5 点具有重要指导意义的结论。继续深入研究 P 类特征与 A 类特征对结果的影响，确定了特征工程的方向，采用 XGBoost Regressor 预测 A 类特征，测试集中 P9 空值采用 LightGBM Regressor 预测填充。

在模型构建阶段为提升方案的稳定性，最终模型采用 5 折交叉验证，并配合 3 个不同的随机数设置。最后对结果做后处理，生成提交结果。

- 赛题背景

在高端制造领域，随着数字化转型的深入推进，越来越多的数据可以被用来进行分析和学习，进而实现制造过程中重要决策和控制环节的智能化，例如生产质量管理。从数据驱动的方法来看，生产质量管理通常需要完成质量影响因素挖掘及质量预测、质量控制优化等环节。

本赛题将关注第一个环节，基于对潜在的相关参数及历史生产数据的分析，完成质量相关因素的确认和最终质量符合率的预测。在实际生产中，该环节的结果将是后续控制优化的重要依据。

- 赛题数据

本赛题要求参赛者对给定的工艺参数组合所生产工件的质检标准符合率进行预测。

本赛题提供的数据包括两类特征：

- ✓ 工艺参数（Parameter）10 项，表示生产工件的设备加工参数，以下称为 P 类特征，分别为 P1、P2...P10；

- ✓ 质量指标（Attribute）10 项，表示产出工件的质量，以下称为 A 类特征，分类为 A1、A2...A10；

所有的特征均以脱敏处理。复赛训练集包括 12934 个样本，测试集包括 6000 个样本，其中 3000 个样本的 P 类特征的第 9 项（Parameter9）缺失，每 50 个样本分为一组，共 120 组。

- 评测方法

本赛题的预测目标为质检指标（不合格、合格、良、优），评价指标采用平均绝对误差（MAE）系数，计算方法如下：

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|pred_i - y_i|$$

其中 pred_i 为预测样本，y_i 为真实样本。MAE 的值越小，说明预测数据与真实数据越接近。最终结果为：

$$Score = \frac{1}{1 + 10 * MAE}$$

最终结果越接近 1 分数越高.

- 数据分析

本文提出的参赛建模策略基于以下数据分析中得出的结论，可总结为 5 点：

1. P1~P4 和 A1~A3 可视为连续性特征；P5~P10 和 A4~A10 可视为离散性特征；
2. 训练集和测试集中部分特征取值的差异是由于保留位数不同造成；
3. 大部分特征经过对数变换后为正态分布，便于后续处理；
4. P 类特征中，P1 和 P4 在训练集和测试集上分分布存在差异，且在后续特征选择中发现 P1~P4 对模型的干扰较大；
5. A 类特征中，A4~A10 与预测目标均有显著性关系，其中 A4~A6 特征对预测目标有决定性影响；

以上结论对本文提出的预测防范有重要的指导意义。

特征的连续性与离散性

尽管 P 类特征与 A 类特征均以数值形式提供，在分析过程中发现，P 类特征中 P1~P4 和 A 类特征中 A1~A3 的唯一值数量和样本及成本相同，可视为连续行为特征；其他特征的唯一值数量均未超过样本数的 10%，可视为离散型特征。

保留位数对 P 类特征的影响

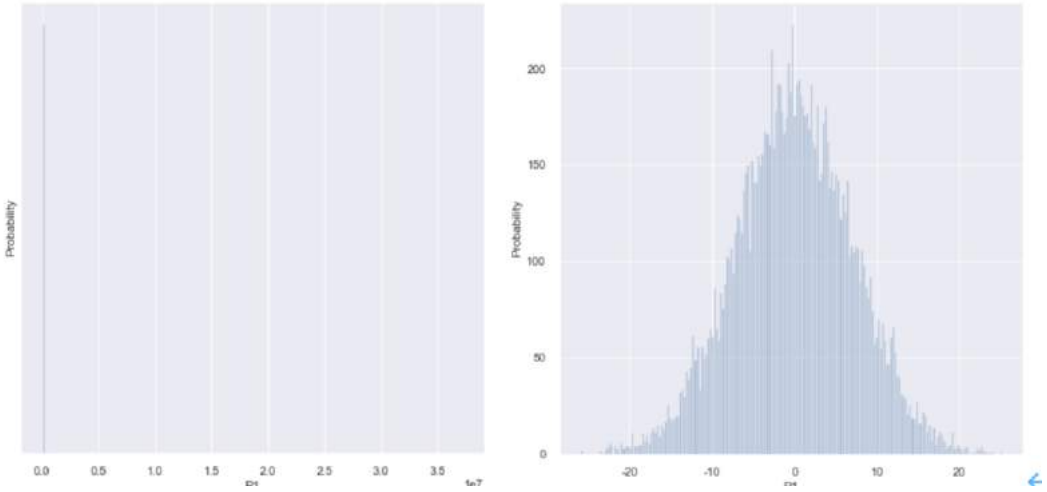
基于以上的发现，在比较特征在训练集和测试集上的分布发现，离散型特征 P5~P10 中，部分特征的部分取值仅在训练集和测试集中出现。进一步探索后发现，这种情况大部分是由于某些取值的保留位数不同造成的。

例如 P9 特征，在训练集中取值 0.0045328891125472222，在测试集中存在取值 0.0045328891125472216，但可以假设两者是相同的，其差异由导出时保留的位数不同造成。

特征名	训练集		测试集	
	仅存在于训练集的值的数量	对应样本数量	仅存在于测试集的值的数量	对应样本数量
P5	16	690	7	270
P6	12	1741	8	976
P7	0	0	0	0
P8	4	7	1	1
P9	6	592	3	145
P10	4	1273	4	627

对数变换对特征的影响

原始数据集中各项特征的取值范围比较大，最小值一般接近 0 而最大值超过 10^5，呈典型长尾分布形态。为了更好的分析特征，本队对所有的特征进行底数为 2 的对数变换，变换后的连续型特征和部分离散型特征呈正态分布。



P 类特征对结果的影响

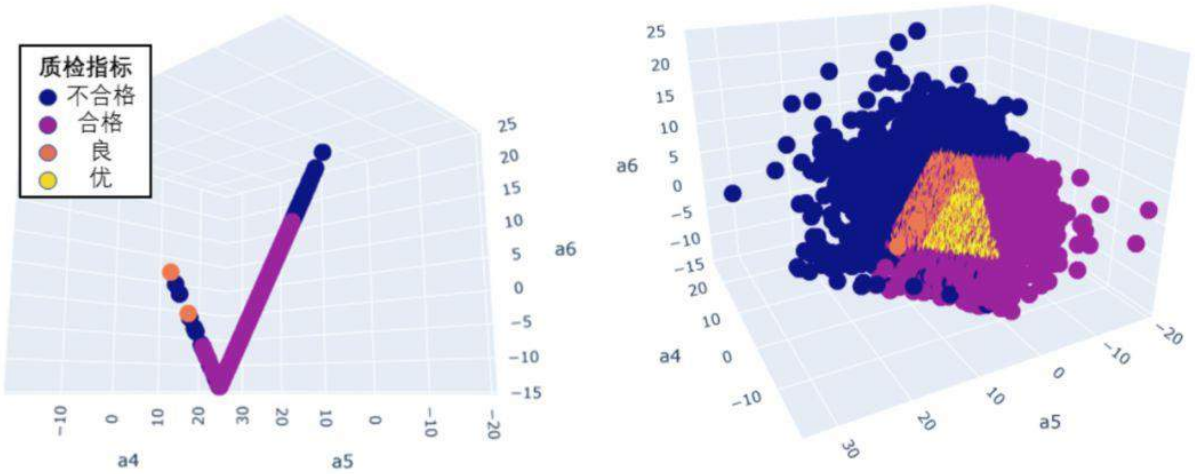
分析过程中发现 P 类特征中，经过对数变换后 P1 与 P4 在训练集和测试集上存在分布差异，在训练集上的分布方差较大。由于数据集规模较小，这种差异可能会对结果造成影响。

在实际线下验证中发现，当模型加入 P1~P4 特征时，在验证集上的评分会下降，在进一步用特征选择方法分析后发现，由于 P1~P4 特征取值较多，对于基于树模型的机器学习方法更加容易被选择为分裂条件，因此导致模型中此类特征重要性比较高。但在线下验证集中重要度显著下降，即这 4 项特征对模型有一定的误导性。

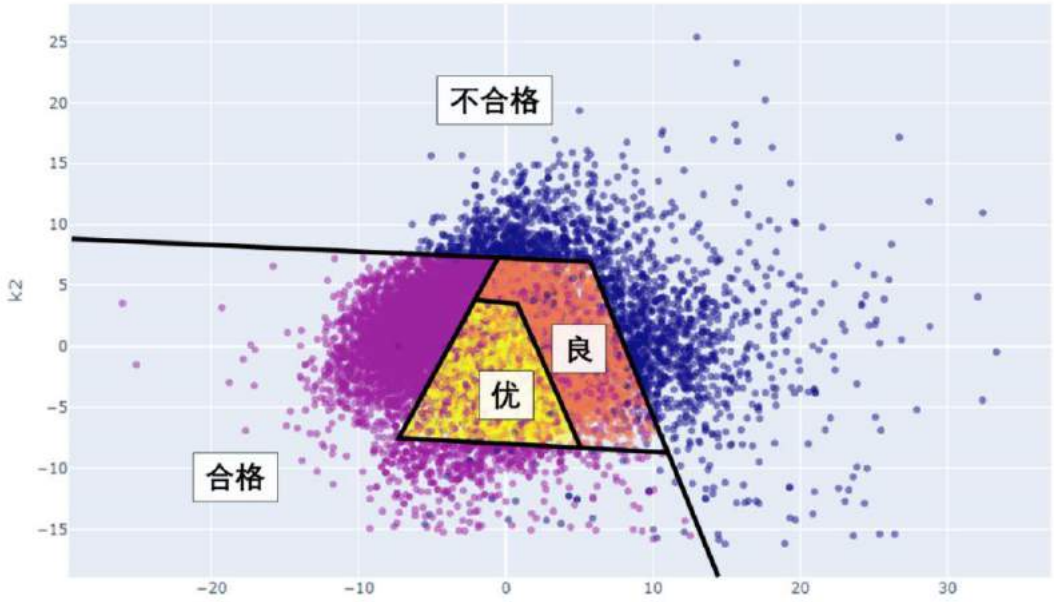
A 类特征对结果的影响

首先除 A1~A3 之外，其他 A 类特征对预测目标均有良好的区分度，显然是由于预测目标由工件质量指标综合得到的原因，而 A1~A3 特征是否加入与预测结果的影响较小。

其次 A4~A6 特征对预测目标有非常显著的影响。如下图所示，在经过对数变换后 A4、A5 和 A6 特征构成的三维空间中，训练集样本的散点图，颜色代表样本质检指标。从图中可以发现：（1）所有样本分布在两个相交的平面上；（2）绝大多数样本在同一个平面，且质检指标分布存在极为明显的规律。



基于上述发现，对 A4~A6 特征进行降维处理，得到特征 K1 和 K2，如下图散点图所示。以上分析说明 A4~A6 特征对预测结果有决定性影响。在后续分析中，我们尝试将 K1 和 K2 与其他 P 类和 A 类特征叠加，但未能出现进一步区分质检指标的组合。



- 预处理&特征工程建模

基于数据分析的结果，本队的建模方案包含 4 项内容：数据预处理、特征工程、机器学习模型构建和结果后处理。

数据预处理环节包括以下 2 项操作：

1. 对 P 类特征进行化简，即数值保留一定位数。若保留位数过小，可能导致部分接近的取值被合并，若过大则无法起到消除训练集和测试集差异的作用。经过尝试本队最终选择 P 类特征保留 11 位小时。
2. 对 P 类特征和 A 类特征进行底数为 2 的对数变换；

特征名	训练集		测试集	
	仅存在于训练集的值的数量	对应样本数量	仅存在于测试集的值的数量	对应样本数量
P5	9	20	0	0
P6	5	8	1	1
P7	0	0	0	0
P8	3	4	0	0
P9	3	8	0	0
P10	0	0	0	0

特征工程：P9 预测值

由于测试集中 P9 特征存在缺失值，且该特征在本队构建的模型中重要性较高，因此本队采用 P5~P8 和 P10 特征预测 P9 特征，并且使用预测结果填充测试集中的缺失值。根据数据分析中的结

论，可将 P9 特征视为离散型特征，因此在预测前先对 P9 先按照升序排序后进行编码，共 17 个取值。经过对多种模型的尝试与比较，最终采用 LightGBM Regressor 作为预测模型。

特征工程：A4~A6 预测值

根据数据分析中的结论，A4~A6 特征与预测目标有显著相关性。由于测试集不包含 A 类特征，因此考虑先采用训练集 P 类特征分别预测 A4~A6 特征，再将预测到的 A4~A6 特征作为训练集合测试集的输入。

具体操作：

1. 将训练集样本样本平均分为 6 份；
2. 每次使用其中 5 份训练集数据，以 P5~P10 特征作为输入，建立模型预测 A4~A6 特征，再用训练到的模型预测剩余 1 份训练集数据和测试集数据中的 A4~A6 特征；
3. 重复 b 步骤 6 次，得到所有训练集 A4~A6 特征的预测值，以及测试集数据集中 A4~A6 特征预测值的均值。

经过对多种模型的尝试与比较，最终采用 LightGBM Regressor 作为预测模型。

特征工程：其他统计特征

本队在比赛期间共尝试构建 106 个统计特征，经过特征筛选后，最终方案保留一下 2 类共 18 个统计特征：

1. 以一项特征作为分组条件，对另一项特征计算分组后的平均值、标准差、综合、唯一数计数，最终保留 15 项该特征；
2. 频率编码，最终保留 3 项特征；

特征工程：特征选择

根据数据分析中的结论，最终作为模型输入的特征为：P 类特征的 P5~P10，A 类特征的 A4~A6（预测值），以及 18 项统计特征。

模型构建

经过多次线下与线上验证后发现，在采用不同随机数时结果评分存在较为明显的差异。为提升方案最终的稳定性，最终模型采用 5 折交叉验证，并配合 3 个不同的随机数涉及，即 3 次 5 折交叉验证，最终得到 15 次预测值，将其平均值作为最终的预测结果，并取得相对稳定的线上评分。

经过对多种模型的尝试与比较，最终采用 XGBoost Regressor 作为预测模型。

结果后处理

根据赛题中的描述，在实际生产中，同一组工艺参数设置下生产的工件会出现多种质检结果，因此最终分类器模型的输出结果采用各分类的预测概率，则测试集中单组的预测结果预测为组内所有样本分布预测概率之和。

- 方案效果

本文提出的预测方案在比赛符合的 B 榜显示功能评分 0.7079336，位列第一名。

- 致谢

感谢 CCF 大数据与计算智能大赛 DataFountain 平台给我们提供一次近距离接触人工智能应用的宝贵机会。感谢活跃在 DataFountain 平台的数据科学家以及工作人员，你们的无私奉献促进了国内整个数据科学竞赛的水平，也让更多年轻人感受人工智能的魅力。我们对在本次参赛过程中学习到很多，我们能有这样的成绩离不开你们的帮助，再次以我们最诚挚的谢意和最衷心的感谢。比赛链接：

https://www.datafountain.cn/competitions/351

冠军代码：

https://github.com/CcIsHandsome/-TOP1-

乘用车细分市场销量预测

出题单位：深瞳云涂

Data Competition in 2019

赛题类型：第一赛道-算法题

技术方向：预测回归， 数据挖掘

赛题背景：消费者购车决策的过程正从线下向线上转移，找到消费者在互联网上的行为数据与销量之间的相关性，可为汽车行业带来更准确有效的销量趋势预测。

赛题任务：要求参赛者根据给出的 32 款车型，在 15 个细分市场的 24 个月销量数据，建立销量预测模型，预测同一款车型和相同细分市场在接下来连续 4 个月份的销量。

团队名称：秋名山车神

成员介绍：

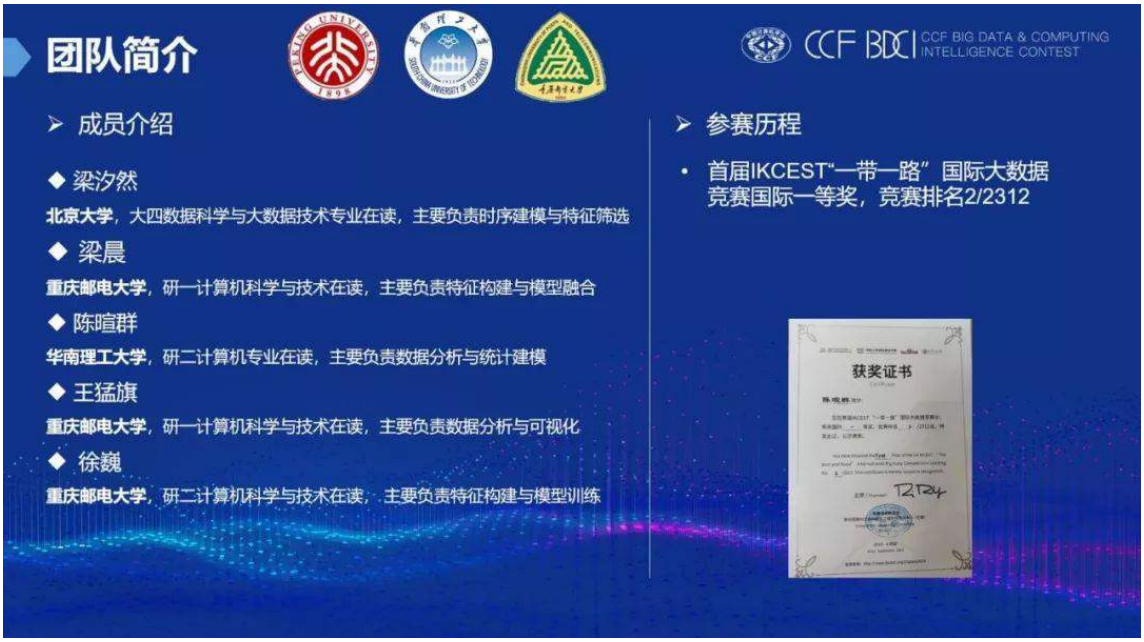
梁晨：队长，重庆邮电大学，计算机科学与技术 2019 级。

梁汐然：队员，北京大学，数据科学与大数据技术 2016 级。

陈暄群：队员，华南理工大学，计算机技术 2018 级。

王猛旗：队员，重庆邮电大学，计算机科学与技术 2019 级。

徐巍：队员，重庆邮电大学，计算机科学与技术 2018 级。



如你们所见，我们来自三个不同的学校，通过本次比赛聚集在一起，在此也十分感谢 DF 平台与主办方 CCF。我们团队中大多数人都是第一次参加数据科学竞赛。

本次比赛要特别感谢的人是鱼佬，他的框架太强了。换成是我我不一定会在比赛中开源这么强的思路，从某种层面上来说这可能会坑到自己。所以敢于开源的人都应该值得称赞，开源与分享可以让整个环境进步。

另外还要感谢月月鸟，阿道，焕明（校友，就是 54 的那个规则开源，实际上我发现很多队伍的规则都是基于他做的），他们的开源也让我们学习到了很多。

我们本次的方案一共约 500 行代码，主要的工作在于特征工程与规则构造，思路，代码都很简单，运行只需 3min，请放心食用。开源源码：

https://github.com/cxq80803716/2019-CCF-BDCI-Car_sales

接下来我会详细介绍一下本次的赛题与解决方案。

近几年来，国内汽车市场由增量市场逐步进入存量市场阶段，2018 年整体市场销量首次同比下降。在市场整体趋势逐步改变的环境下，消费者购车决策的过程也正在从线下向线上转移，我们希望能在销量数据自身趋势规律的基础上，找到消费者在互联网上的行为数据与销量之间的相关性，为汽车行业带来更准确有效的销量趋势预测。

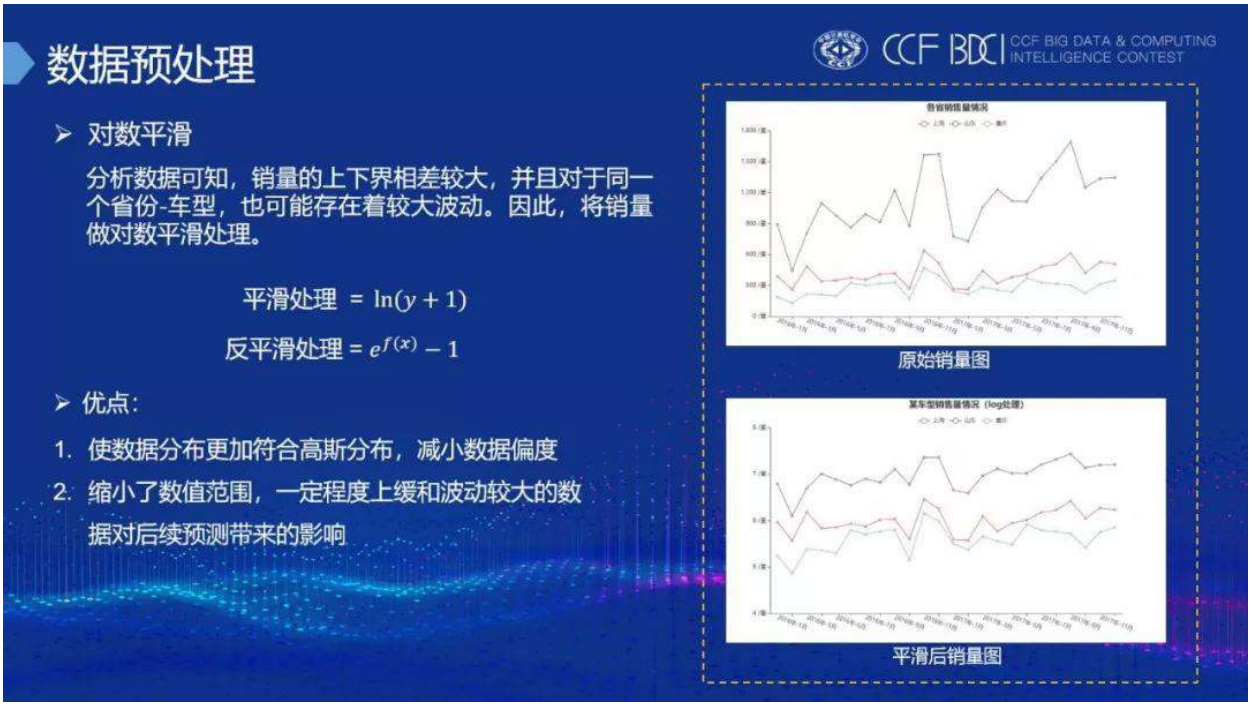
比赛链接：https://www.datafountain.cn/competitions/352

本赛题需要参赛队伍根据给出的 60 款车型在 22 个细分市场（省份）的销量连续 24 个月（从 2016 年 1 月至 2018 年 12 月）的销量数据，建立销量预测模型；基于该模型预测同一款车型和相同细分市场在接下来一个季度连续 4 个月份的销量；除销量数据外，还提供同时期的用户互联网行为统计数据，包括：各细分市场每个车型名称的互联网搜索量数据；主流汽车垂直媒体用户活跃数据等。参赛队伍可同时使用这些非销量数据用于建模。

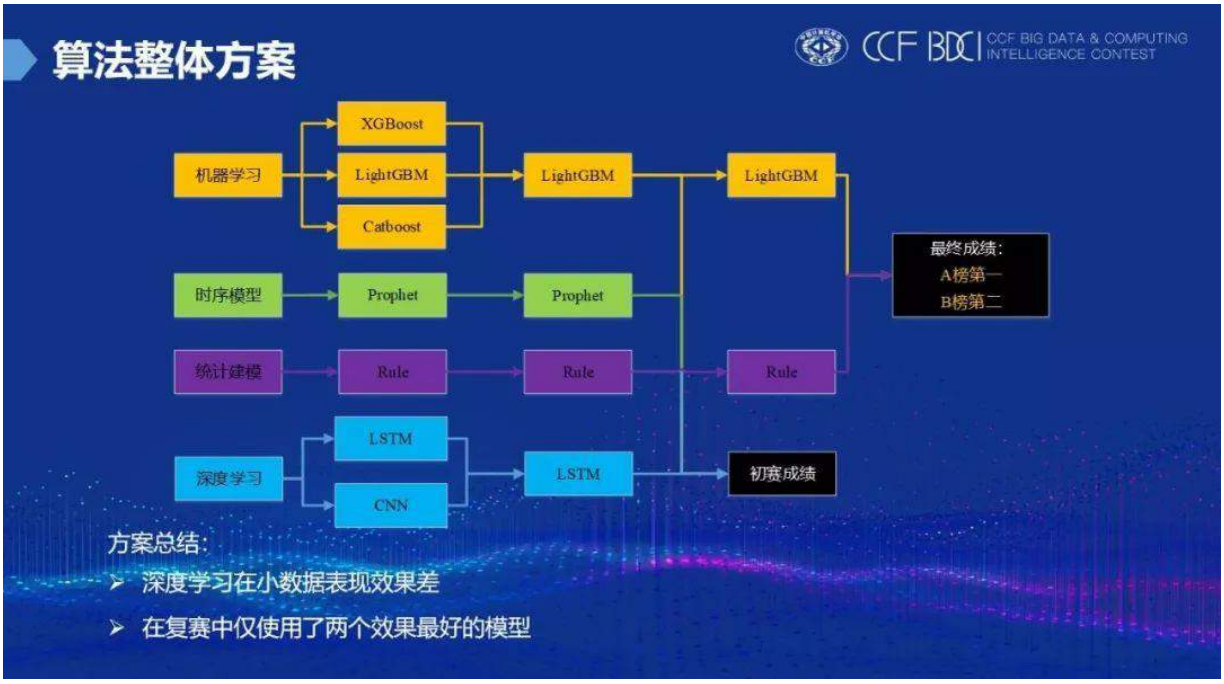
简单来说，本次赛题给出 2016.1-2017.12 的省份，车型，车身，销量，搜索量，评论量，评价量等，要求预测 2018.1-2018.4 的销量。

通过初步分析数据可以发现省份-车型所组成的类别特别多，并且对于同一车型，波动也挺大的，销量的范围较大等。

顺便说一下，我们经过多次尝试发现评论量与评价量几乎起不到作用，所以这两个特征我们并没有使用，如果你有办法处理这两个特征，欢迎评论讨论。



我们在初赛的时候尝试了 xgb, lgb, cat, prophet, rule, lstm, cnn，并且初赛的最终结果是由 lgb, prophet, rule 与 lstm 融合而来。但是后来我们发现就算只用 lgb 和 rule 也能得到差不多的分数，时序模型与深度学习模型在这道小数据时序问题上并不适用，又考虑到工业环境中模型越少，越简单越好，因此复赛时，我们只使用了差异性足够大的 lgb 与 rule 两个模型。



直接使用鱼佬的框架能带来相当不错的效果，但是可以发现，越往后的月份误差的累计会越来越严重，因此在模型中，我们只拼接了 1 月份预测出来的值。

在构造特征之前，对特征一些分析。画出省份销量的热力图可以发现销量与省份的发达程度，临海程度是成一定正相关的，这与我们的直觉相符。因此各省份销量的差异明显，建模时必须对省份进行区别。

对于同一个月份，2016 年与 2017 年的销量类似，即销量有周期性变化的特性。而且对于宏观变化而言，2017 年的销量较 2016 年销量呈现下降趋势。春节月的销量较低，春节后逐渐上升，年末达到最高。

这一块是本次工作的重头戏，也是我们花费了最多时间的地方，最终，我们的模型特征由以下几部分组成。考虑到春节等节假日，我们构造了与节假日相关的一些特征。

考虑到每个月的天数，工作日不同，我们构造了相关的特征。由于是个时序问题，因此历史销量与销量的变化趋势是我们应该考虑的重点。针对此，我们在多个不同粒度下构造了历史平移特征，差分特征，同比/环比特征与趋势特征。

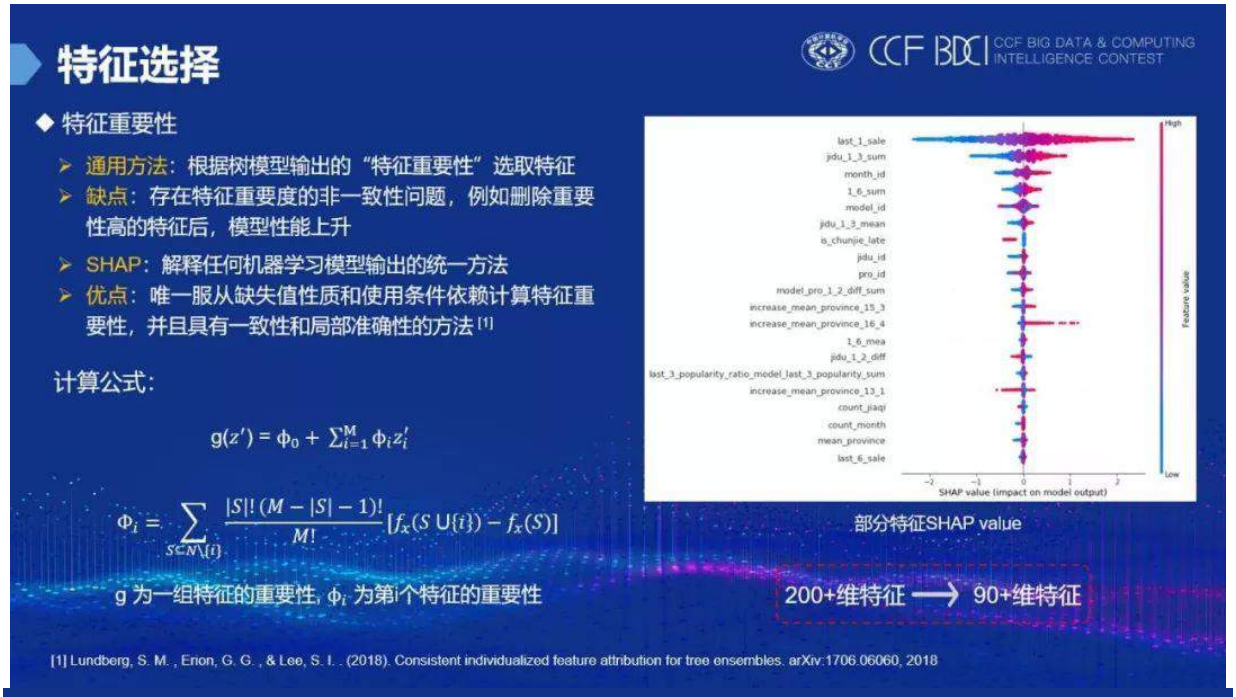
不过经过我们的尝试，同比的效果不怎么好。更加详细的内容请看代码。

我们所构造的特征实际上不止这些，不过由于信息重叠与毒特等原因，我们使用了一些方法对特征进行筛选。我们使用了树模型的特征重要性排序，均值判断与 SHAP 进行特征筛选。

特征重要性排序：根据树模型输出的特征重要性进行筛选，去掉重要性低的特征。

均值判断：由于本道赛题中，1234 月具有相对固定的均值比例与均值大小，因此可以通过添加/删除特征后 1234 月的平均均值来大概判断特征的好坏。

SHAP：利用了组合博弈论的知识，防止因为信息重叠而导致的特征重要度不公平的情况。



规则部分的框架是使用历史同期销量，历史最近销量与指数平滑进行加权得到一个初步预测销量，然后用上下半年的销量变化趋势与平滑构造年度销量变化趋势，最后两者相乘即可得到规则的预测结果。

由此可见，规则带有相当多的超参数。据我了解不少团队复赛时规则血崩，这也是时序题里面规则的泛性问题。我们初赛时规则可以达到 0.633，是一个绝对的主力，复赛提交次数太少，最终规则也只有 0.598，只能以 lgb 为主，规则为辅。

初步销量预测

以省份-车型作为主键，利用历史数据得到初步预测的销量：

$$y_{2018,m} = (\bar{y}_{2016,m} + \bar{y}_{2017,m} + \bar{y}_{near,m})$$

历史同期销量
- 2016年
  - 1月：$\bar{y}_{2016,m} = 0.8 * y_{2016,m} + 0.2 * y_{2016,m+1}$
  - 2,3,4月：$\bar{y}_{2016,m} = 0.2 * y_{2016,m-1} + 0.6 * y_{2016,m} + 0.2 * y_{2016,m+1}$
- 2017年：$\bar{y}_{2017,m} = 0.2 * y_{2017,m-1} + 0.6 * y_{2017,m} + 0.2 * y_{2017,m+1}$

历史最近销量
- 1,2月：$\bar{y}_{near,m} = 0.2 * y_{m+24-3} + 0.2 * y_{m+24-2} + 0.3 * y_{m+24-1} + 0.3 * y_{exp3,m+24}$
- 3,4月：$\bar{y}_{near,m} = 0.2 * y_{m+24-3} + 0.2 * y_{m+24-2} + 0.6 * y_{m+24-1}$

注：$y_{yyyy,m}$ 表示yyyy年m月的历史销量　　y_m 表示从2016.1开始第m月的历史销量,即 $y_{2018,1} = y_{25}$

$y_{exp3,m}$ 表示三次指数平滑得到的第m个月的预测销量

三次指数平滑

基本原理：时间序列的态势具有稳定性或规则性，所以时间序列可被合理地顺势推延。较大的平滑因子可以较好地反映出数据变化的趋势，分析数据可知销量波动较大，因此所提出的方法中，平滑因子α=0.95。三次指数平滑可以反映出数据的二次曲线趋势。

一，二，三次指数平滑计算公式：

$$y_{exp1,m} = \alpha * y_{m-1} + (1-\alpha) * y_{exp1,m-1}$$

$$y_{exp2,m} = \alpha * y_{exp1,m-1} + (1-\alpha) * y_{exp2,m-1}$$

$$y_{exp3,m} = \alpha * y_{exp2,m-1} + (1-\alpha) * y_{exp3,m-1}$$

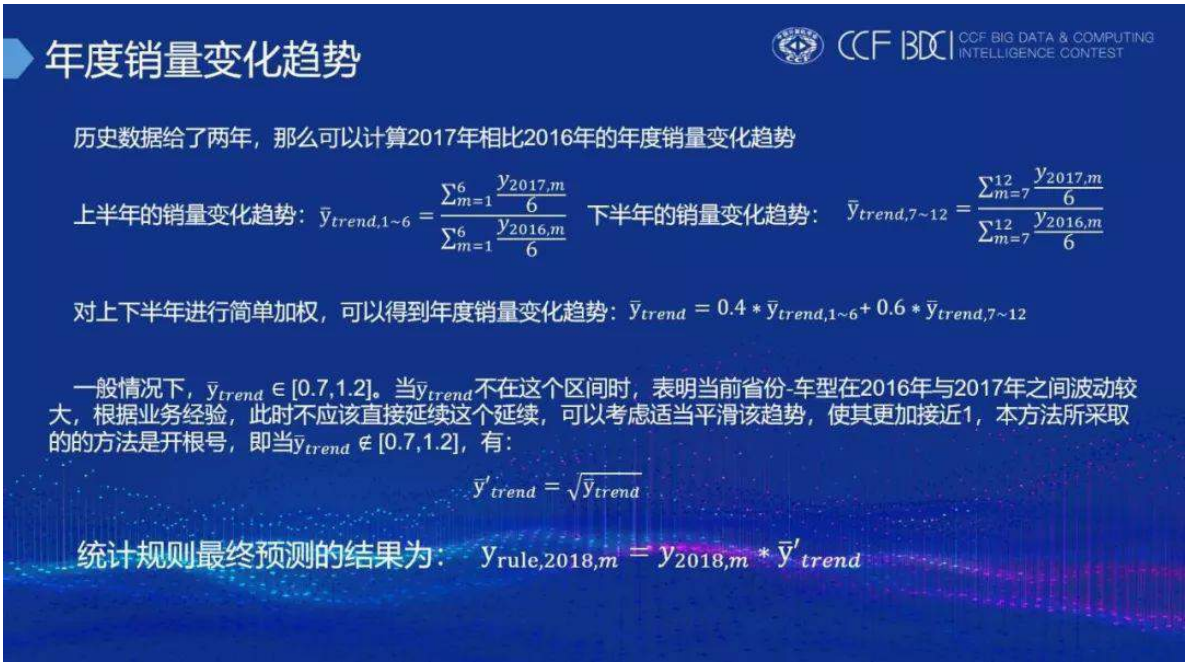
三次指数平滑预测公式（可预测后续的多期，但越往后越不准）：

$$a_m = 3 * y_{exp1,m} - 3 * y_{exp2,m} + y_{exp3,m}$$

$$b_m = \frac{\alpha * [(6-5\alpha) * y_{exp1,m} - 2 * (5-4\alpha) * y_{exp2,m} + (4-3\alpha) * y_{exp3,m}]}{2 * (1-\alpha)^2}$$

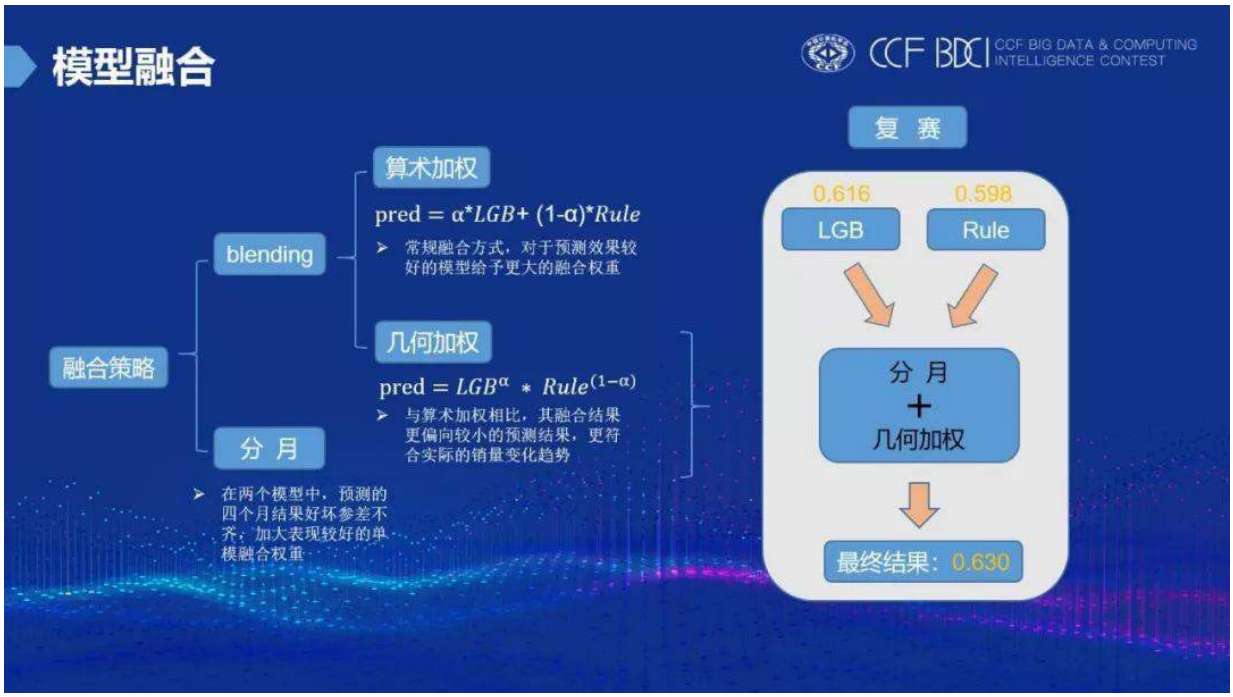
$$c_m = \frac{\alpha^2 * [y_{exp1,m} - 2 * y_{exp2,m} + y_{exp3,m}]}{2 * (1-\alpha)^2}$$

$$y_{exp3,m+T} = a_m + b_m * T + c_m * T^2$$



因为只有两个模型，所以模型融合基本上不用考虑太多，直接进行简单的算数/几何加权即可，由于几何加权可以使预测值偏小，而 2018 年的销量理应是较之前低的，所以我们使用了几何加权进行融合。

另外 1234 月分开进行融合，可以带来微小的提升。



这里特别提一下模型的数量与运行时间的优点，据我们决赛观察，大多数队伍都有超过 2 个的模型，而且有的队伍需要超过一个小时的运行时间。从工业角度来讲，我们的方案应该是更加适用的。

不过还有很多可以提升的地方，比如评论，评价的使用方式，规则中超参数过多问题，从某个时间点开始的某项政策对后续的影响等。如果赛题能提供更多的特征与数据，相信可做的地方还有相当多。