Kaggle 에서 얻을 수 있는 건?

이유한

카이스트
생명화학공학과
분자 시뮬레이션 실험실
(Molecular Simulation Laboratory)

Kaggle 이란?

2010년 설립된
빅데이터 솔루션 대회 플랫폼 회사

2017년 3월 구글에 인수

Data Race for 데이터 과학자!

기업, 정부기관, 단체, 연구소, 개인

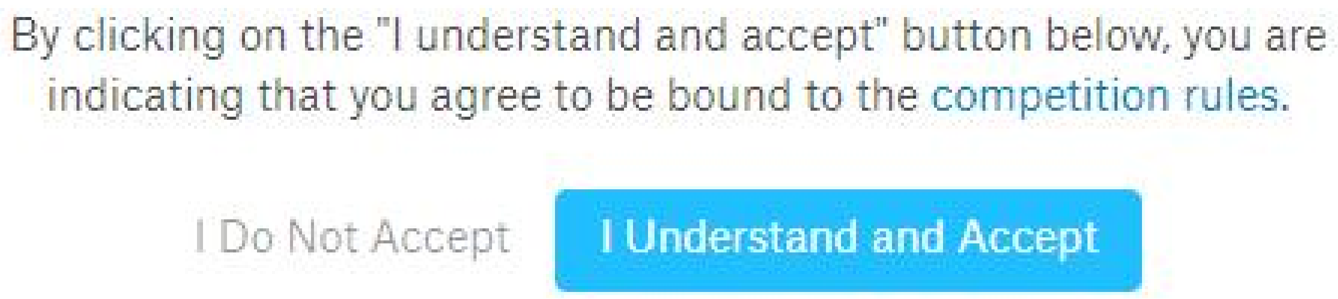
Dataset
With Prize

kaggle

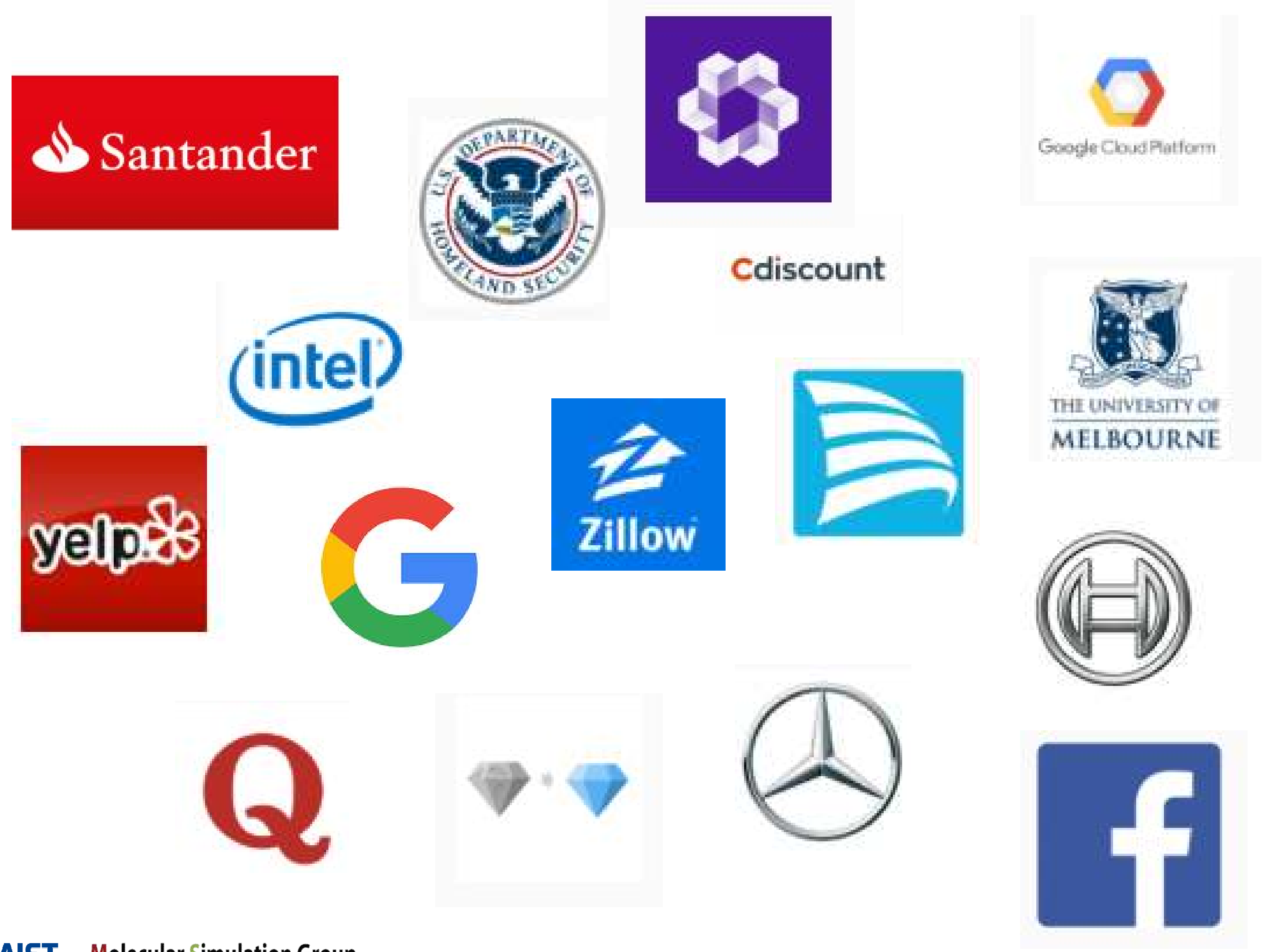
Dataset & Prize
개발 환경(kernel)
커뮤니티(follow, discussion)

전 세계 데이터 사이언티스트

참가하려면?


By clicking on the "I understand and accept" button below, you are indicating that you agree to be bound to the competition rules.

Kaggle 에서 competition 을 주최한 단체, 기업들


Santander
U.S. DEPARTMENT OF HOMELAND SECURITY
Google Cloud Platform
Cdiscount
intel
THE UNIVERSITY OF MELBOURNE
yelp
Zillow

Competition Example – Tensorflow competition

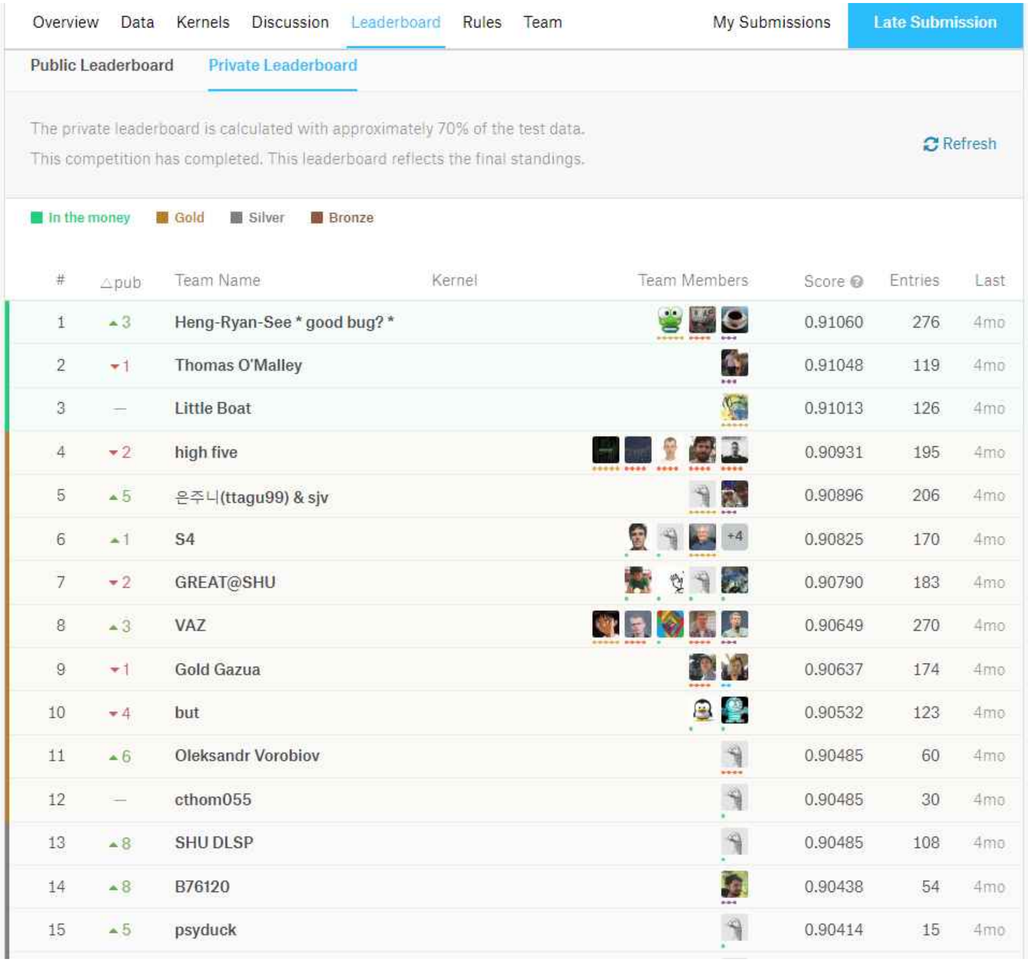
Dataset: 65,000개의 word audio file

Prize :
1st - $8,000
2nd – $6,000
3rd – $3,000
+ spectial price $8,000

Yes, no, up, down, left, right, on, off, stop, go, silence, others 로 이루어진 단어들을 구별하는 AI를 만들어달라!

Competition Example – Tensorflow competition

약 두달간의 RACE
1315 팀 참가



Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Late Submission

Public Leaderboard Private Leaderboard

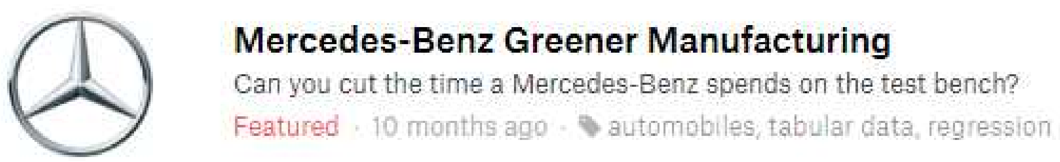
The private leaderboard is calculated with approximately 70% of the test data.
This competition has completed. This leaderboard reflects the final standings.

Refresh

■ In the money ■ Gold ■ Silver ■ Bronze

#	△pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲3	Heng-Ryan-See * good bug? *			0.91060	276	4mo
2	▼1	Thomas O'Malley			0.91048	119	4mo
3	—	Little Boat			0.91013	126	4mo
4	▼2	high five			0.90931	195	4mo
5	▲5	은주니(ttagu99) & sjv			0.90896	206	4mo
6	▲1	S4		+4	0.90825	170	4mo
7	▼2	GREAT@SHU			0.90790	183	4mo
8	▲3	VAZ			0.90649	270	4mo
9	▼1	Gold Gazua			0.90637	174	4mo
10	▼4	but			0.90532	123	4mo
11	▲6	Oleksandr Vorobiov			0.90485	60	4mo
12	—	cthom055			0.90485	30	4mo
13	▲8	SHU DLSP			0.90485	108	4mo
14	▲8	B76120			0.90438	54	4mo
15	▲5	psyduck			0.90414	15	4mo

또 다른 competition 들

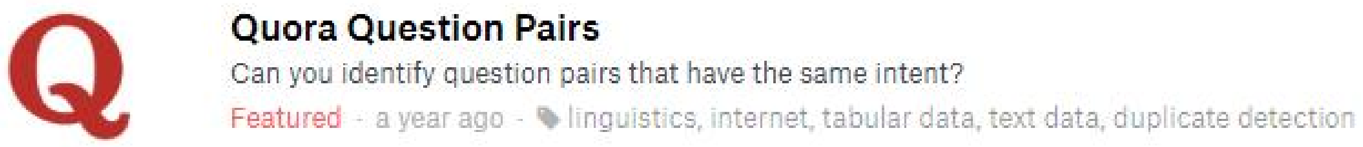


Mercedes-Benz Greener Manufacturing

Can you cut the time a Mercedes-Benz spends on the test bench?

Featured · 10 months ago · automobiles, tabular data, regression

$25,000

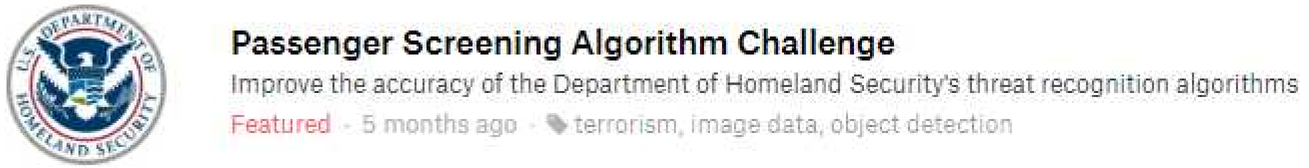


Quora Question Pairs

Can you identify question pairs that have the same intent?

Featured · a year ago · linguistics, internet, tabular data, text data, duplicate detection

$25,000

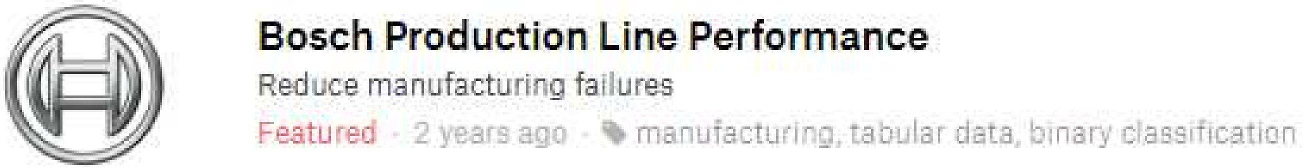


Passenger Screening Algorithm Challenge

Improve the accuracy of the Department of Homeland Security's threat recognition algorithms

Featured · 5 months ago · terrorism, image data, object detection

$1,500,000



Bosch Production Line Performance

Reduce manufacturing failures

Featured · 2 years ago · manufacturing, tabular data, binary classification

$30,000

여지껏 다뤄본 것이
IRIS dataset, MNIST 뿐인데

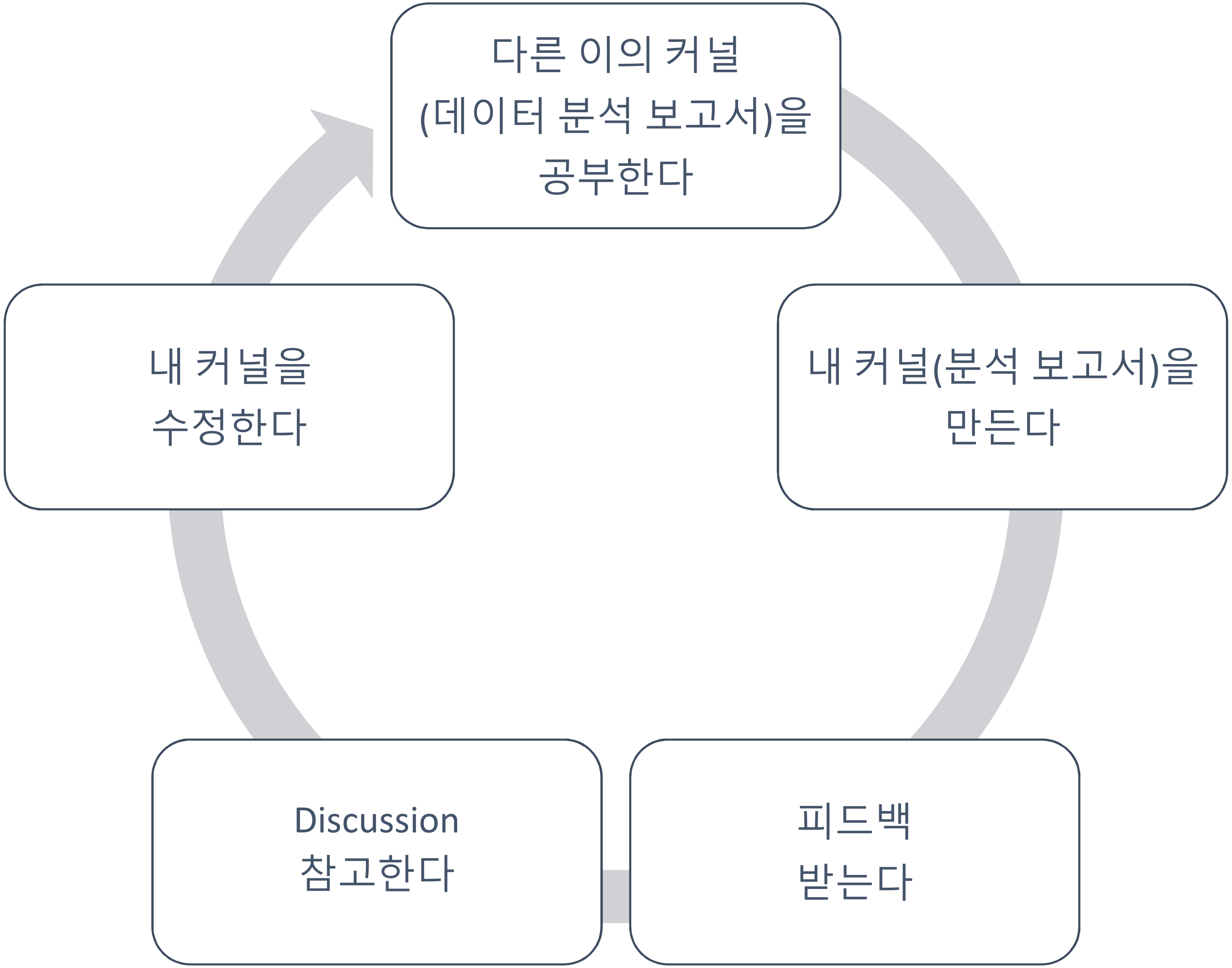
저런 걸 어떻게
분석해야 하나?

공부해서 함께 나누자!

고수의 발자취를
따라가자

모방은 창조의 시작

공부해서 함께 나누자! – 캐글 속 선순환



다른 이의 커널
(데이터 분석 보고서)을
공부한다

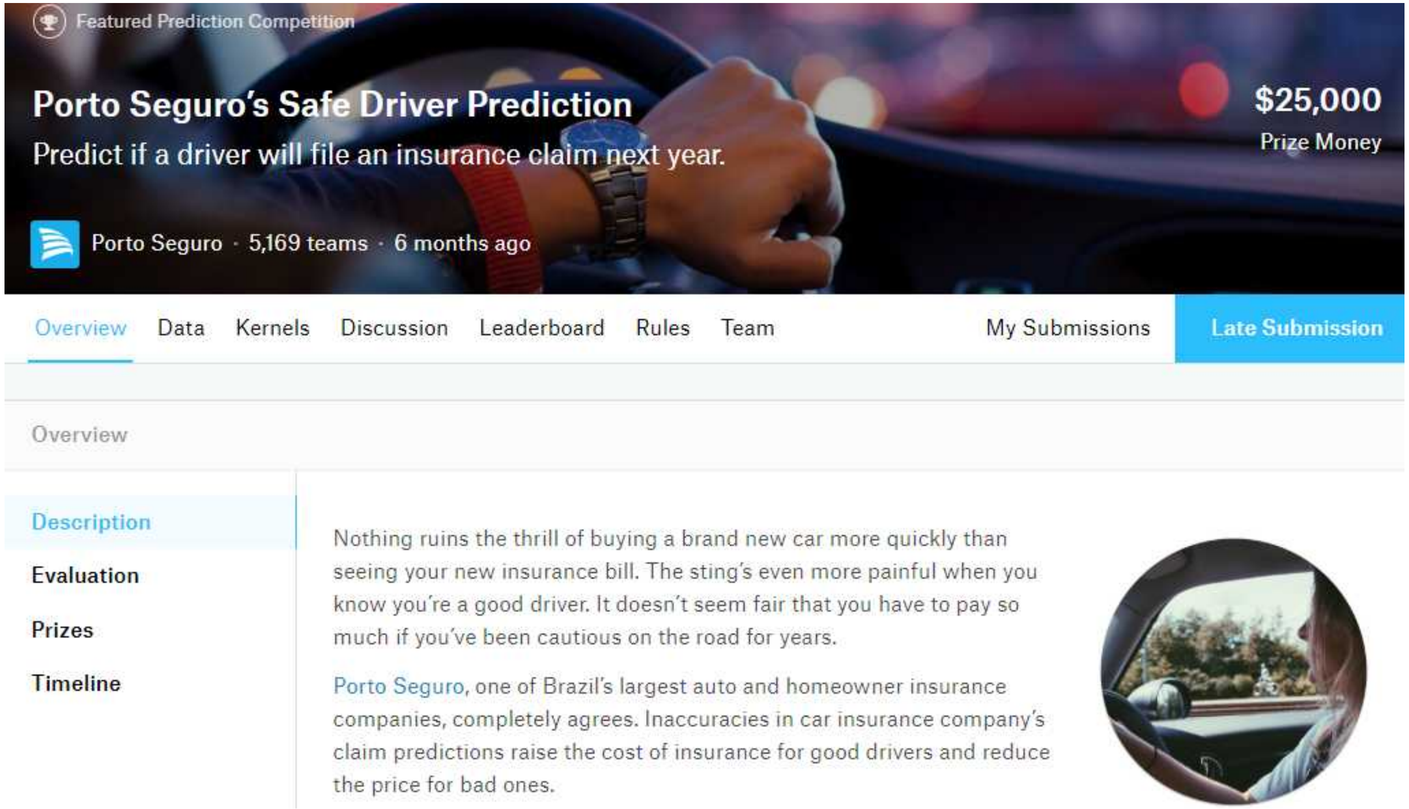
내 커널(분석 보고서)을
만든다

피드백
받는다

Discussion
참고한다

내 커널을
수정한다

My first kaggle story



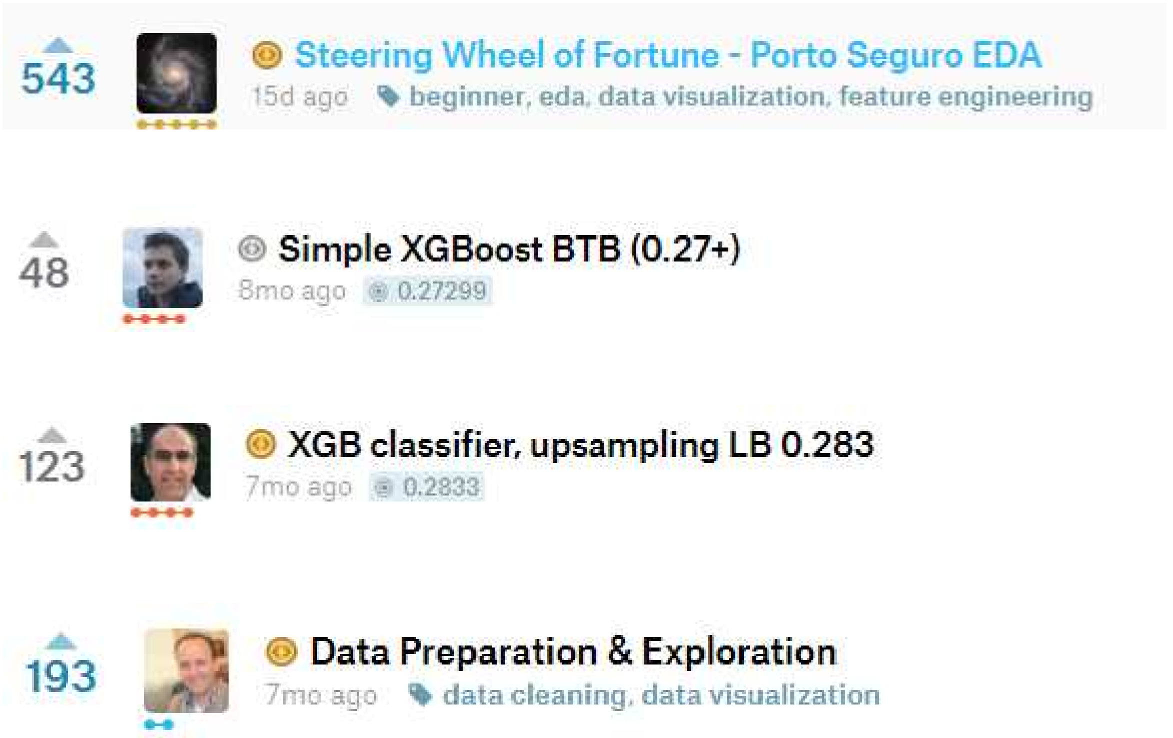
Featured Prediction Competition
Porto Seguro's Safe Driver Prediction
Predict if a driver will file an insurance claim next year.
$25,000
Prize Money
Porto Seguro · 5,169 teams · 6 months ago
Overview
Data
Kernels
Discussion
Leaderboard
Rules
Team
My Submissions
Late Submission
Overview
Description
Evaluation
Prizes
Timeline

Nothing ruins the thrill of buying a brand new car more quickly than seeing your new insurance bill. The sting's even more painful when you know you're a good driver. It doesn't seem fair that you have to pay so much if you've been cautious on the road for years.

Porto Seguro, one of Brazil's largest auto and homeowner insurance companies, completely agrees. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones.

약 60만명의 정보를 가지고 머신러닝 알고리즘을 만들어, 40만명의 개인이 향후에 보험을 계속 사용할 것인지 예측하라

My first kaggle story – learning the kernels



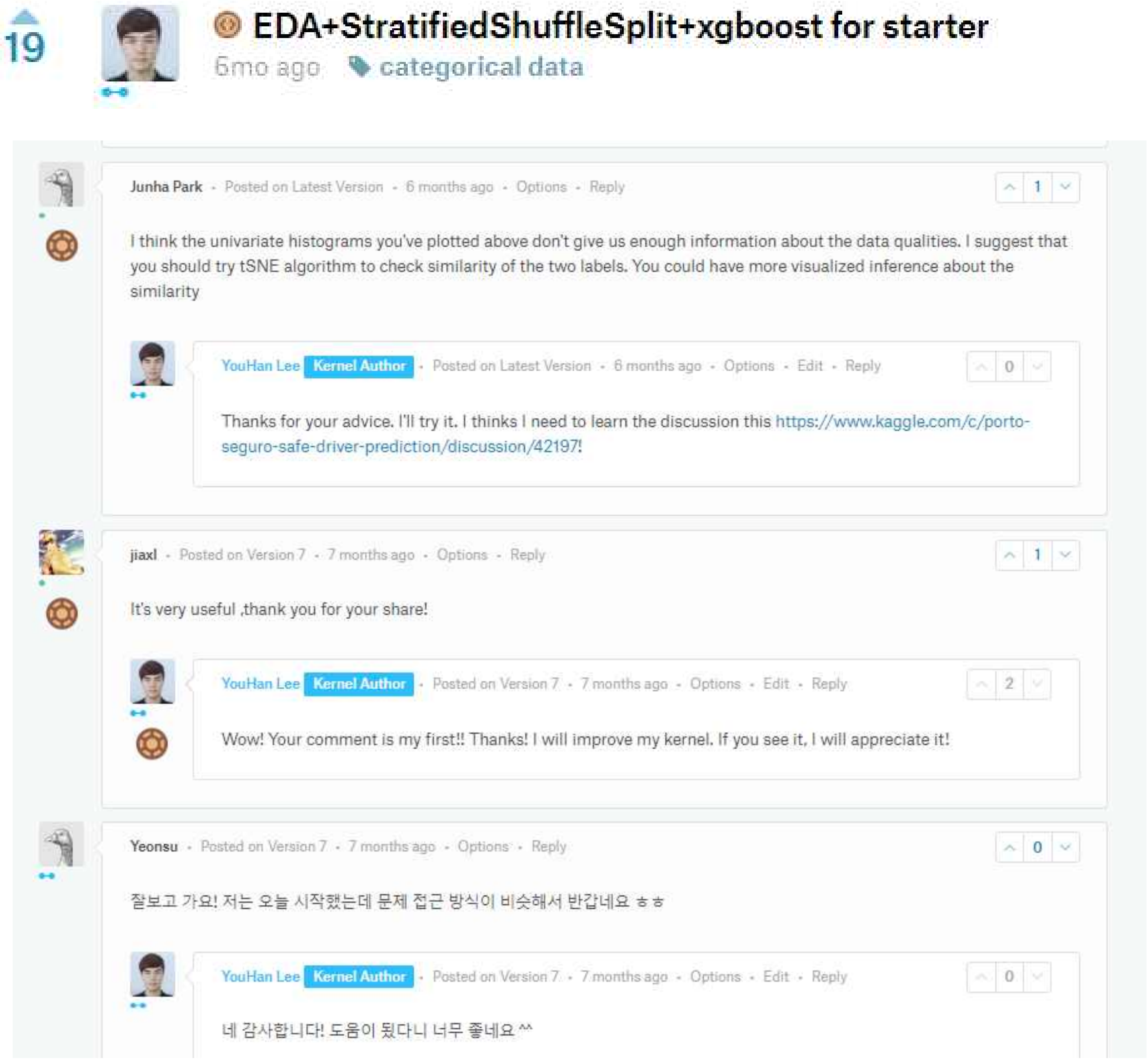
543
Steering Wheel of Fortune - Porto Seguro EDA
15d ago beginner, eda, data visualization, feature engineering

48
Simple XGBoost BTB (0.27+)
8mo ago 0.27299

123
XGB classifier, upsampling LB 0.283
7mo ago 0.2833

193
Data Preparation & Exploration
7mo ago data cleaning, data visualization

My first kaggle story – making my own kernel



19

EDA+StratifiedShuffleSplit+xgboost for starter

6mo ago · categorical data

Junha Park · Posted on Latest Version · 6 months ago · Options · Reply · 1

I think the univariate histograms you've plotted above don't give us enough information about the data qualities. I suggest that you should try tSNE algorithm to check similarity of the two labels. You could have more visualized inference about the similarity

YouHan Lee Kernel Author · Posted on Latest Version · 6 months ago · Options · Edit · Reply · 0

Thanks for your advice. I'll try it. I thinks I need to learn the discussion this https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/discussion/42197!

jiaxl · Posted on Version 7 · 7 months ago · Options · Reply · 1

It's very useful ,thank you for your share!

YouHan Lee Kernel Author · Posted on Version 7 · 7 months ago · Options · Edit · Reply · 2

Wow! Your comment is my first!! Thanks! I will improve my kernel. If you see it, I will appreciate it!

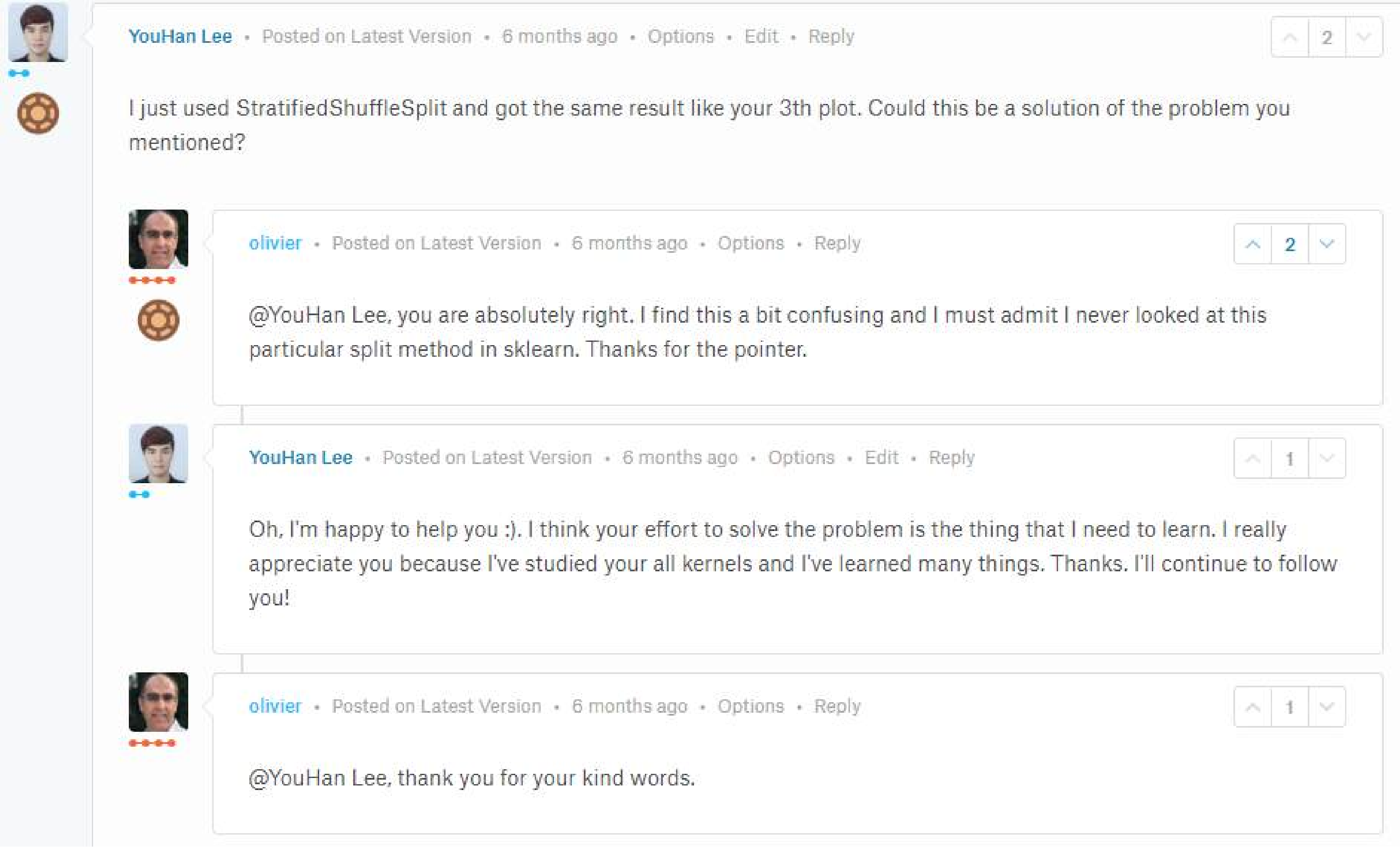
Yeonsu · Posted on Version 7 · 7 months ago · Options · Reply · 0

잘보고 가요! 저는 오늘 시작했는데 문제 접근 방식이 비슷해서 반갑네요 ㅎㅎ

YouHan Lee Kernel Author · Posted on Version 7 · 7 months ago · Options · Edit · Reply · 0

네 감사합니다! 도움이 됐다니 너무 좋네요 ^^

My first kaggle story – ask anything to authors



YouHan Lee • Posted on Latest Version • 6 months ago • Options • Edit • Reply

I just used StratifiedShuffleSplit and got the same result like your 3th plot. Could this be a solution of the problem you mentioned?

olivier • Posted on Latest Version • 6 months ago • Options • Reply

@YouHan Lee, you are absolutely right. I find this a bit confusing and I must admit I never looked at this particular split method in sklearn. Thanks for the pointer.

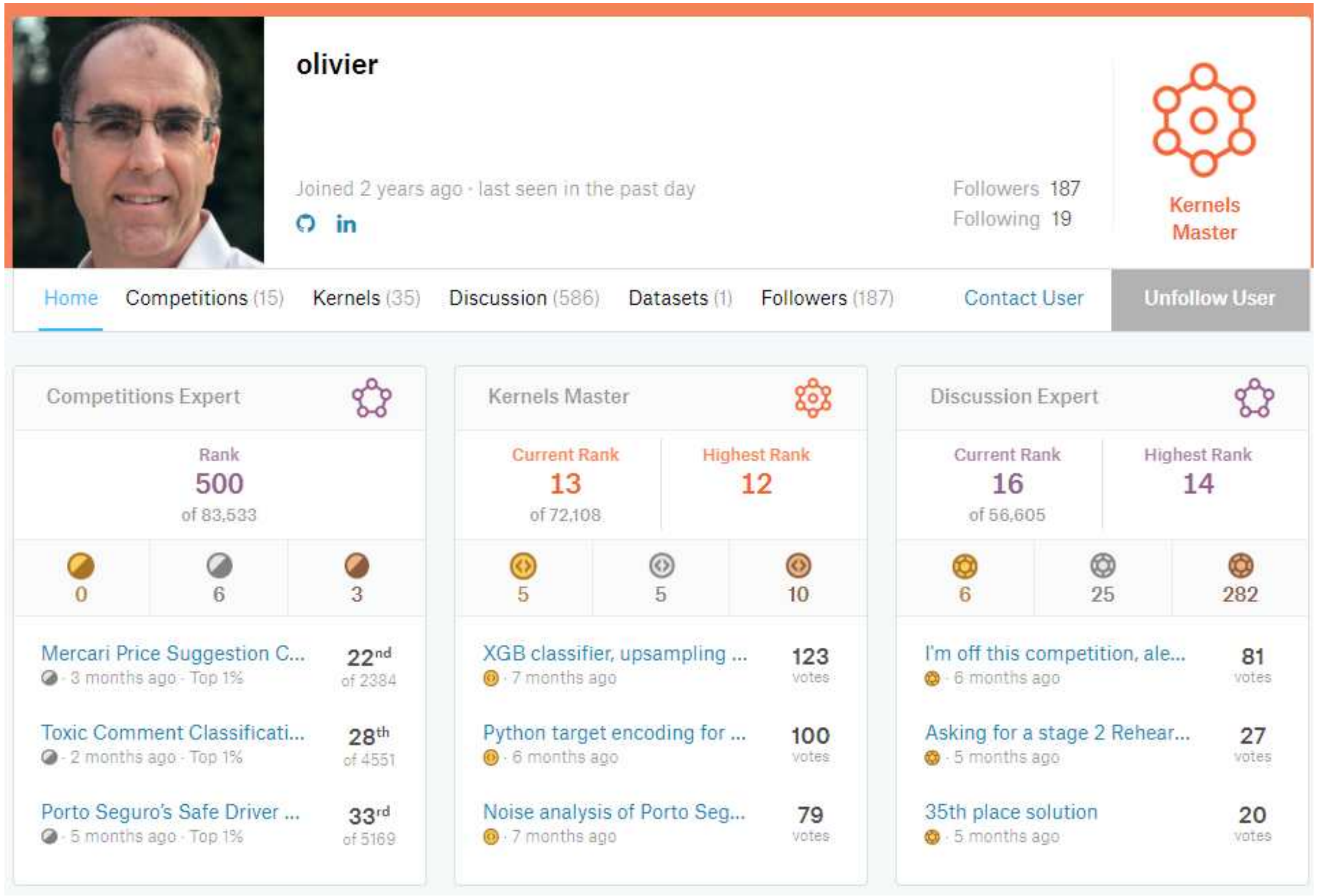
YouHan Lee • Posted on Latest Version • 6 months ago • Options • Edit • Reply

Oh, I'm happy to help you :). I think your effort to solve the problem is the thing that I need to learn. I really appreciate you because I've studied your all kernels and I've learned many things. Thanks. I'll continue to follow you!

olivier • Posted on Latest Version • 6 months ago • Options • Reply

@YouHan Lee, thank you for your kind words.

My first kaggle story – 친절한 올리비에 아저씨



olivier

Joined 2 years ago · last seen in the past day

Followers 187
Following 19

Kernels Master

Home | Competitions (15) | Kernels (35) | Discussion (586) | Datasets (1) | Followers (187) | Contact User | Unfollow User

Competitions Expert

Rank 500 of 83,533

●	●	●
0	6	3

Competition	Result
Mercari Price Suggestion C... 3 months ago · Top 1%	22nd of 2384
Toxic Comment Classificati... 2 months ago · Top 1%	28th of 4551
Porto Seguro's Safe Driver ... 5 months ago · Top 1%	33rd of 5169

Kernels Master

Current Rank 13 of 72,108 | Highest Rank 12

●	●	●
5	5	10

Kernel	Votes
XGB classifier, upsampling ... 7 months ago	123 votes
Python target encoding for ... 6 months ago	100 votes
Noise analysis of Porto Seg... 7 months ago	79 votes

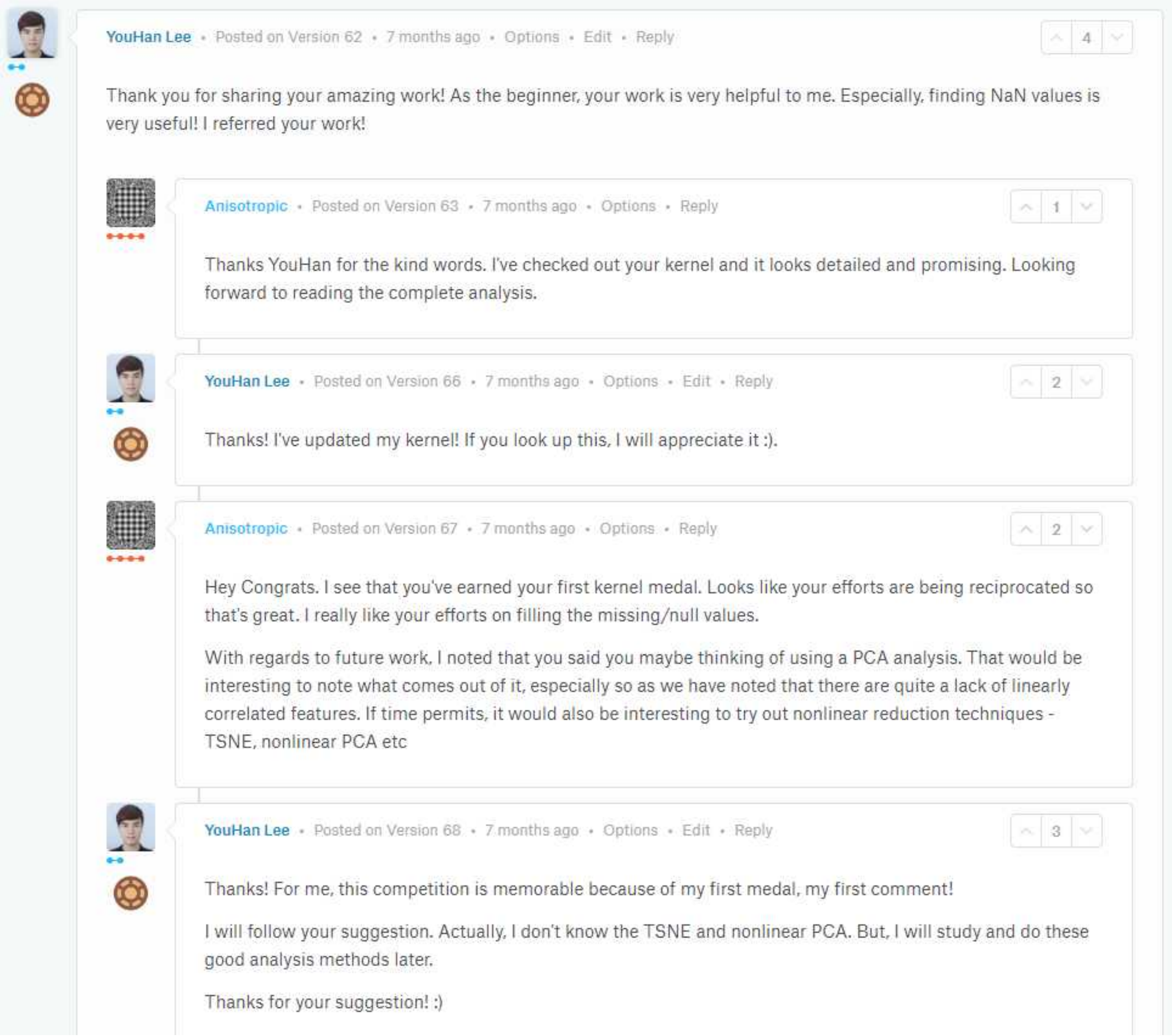
Discussion Expert

Current Rank 16 of 56,605 | Highest Rank 14

●	●	●
6	25	282

Discussion	Votes
I'm off this competition, ale... 6 months ago	81 votes
Asking for a stage 2 Rehear... 5 months ago	27 votes
35th place solution 5 months ago	20 votes

My first kaggle story – ask anything to author



YouHan Lee • Posted on Version 62 • 7 months ago • Options • Edit • Reply 4

Thank you for sharing your amazing work! As the beginner, your work is very helpful to me. Especially, finding NaN values is very useful! I referred your work!

Anisotropic • Posted on Version 63 • 7 months ago • Options • Reply 1

Thanks YouHan for the kind words. I've checked out your kernel and it looks detailed and promising. Looking forward to reading the complete analysis.

YouHan Lee • Posted on Version 66 • 7 months ago • Options • Edit • Reply 2

Thanks! I've updated my kernel! If you look up this, I will appreciate it :).

Anisotropic • Posted on Version 67 • 7 months ago • Options • Reply 2

Hey Congrats. I see that you've earned your first kernel medal. Looks like your efforts are being reciprocated so that's great. I really like your efforts on filling the missing/null values.

With regards to future work, I noted that you said you maybe thinking of using a PCA analysis. That would be interesting to note what comes out of it, especially so as we have noted that there are quite a lack of linearly correlated features. If time permits, it would also be interesting to try out nonlinear reduction techniques - TSNE, nonlinear PCA etc

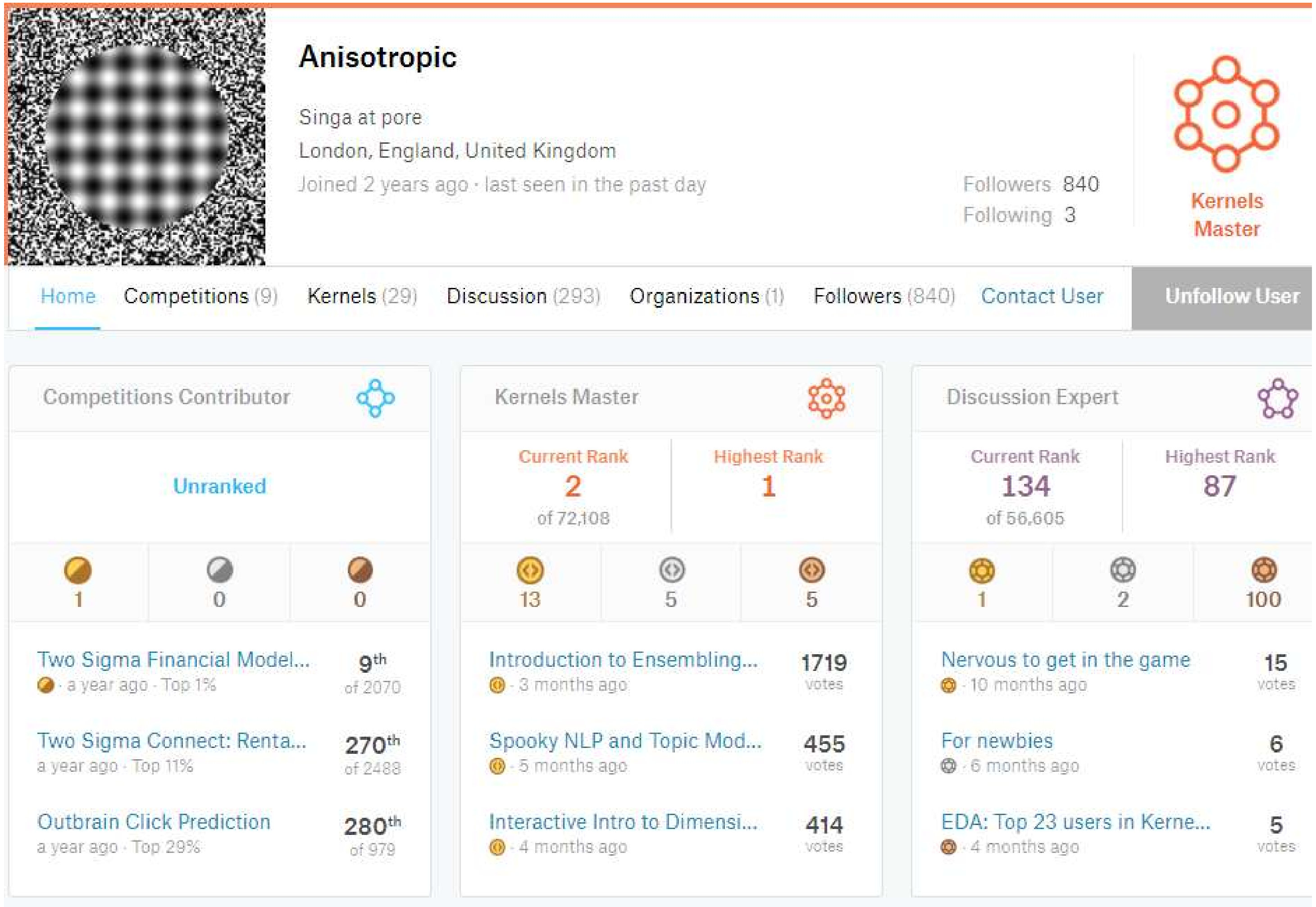
YouHan Lee • Posted on Version 68 • 7 months ago • Options • Edit • Reply 3

Thanks! For me, this competition is memorable because of my first medal, my first comment!

I will follow your suggestion. Actually, I don't know the TSNE and nonlinear PCA. But, I will study and do these good analysis methods later.

Thanks for your suggestion! :)

My first kaggle story – 멋진 스승님



Anisotropic

Singa at pore
London, England, United Kingdom
Joined 2 years ago · last seen in the past day

Followers 840
Following 3

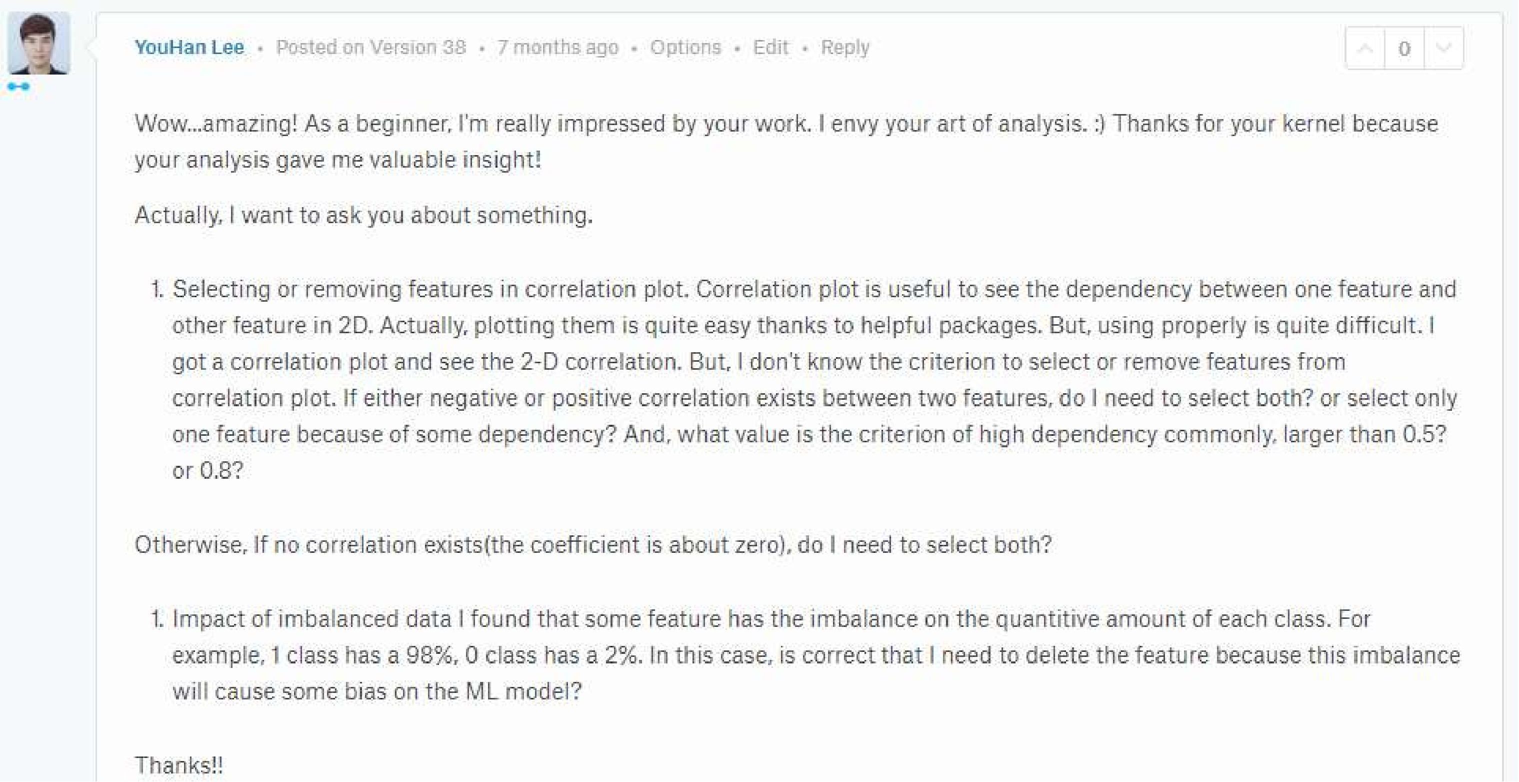
Kernels Master

Competitions Contributor		
Unranked		
● 1	● 0	● 0
Two Sigma Financial Model... ● · a year ago · Top 1%	9th of 2070	
Two Sigma Connect: Renta... a year ago · Top 11%	270th of 2488	
Outbrain Click Prediction a year ago · Top 29%	280th of 979	

Kernels Master		
Current Rank 2 of 72,108	Highest Rank 1	
● 13	● 5	● 5
Introduction to Ensembling... ● · 3 months ago	1719 votes	
Spooky NLP and Topic Mod... ● · 5 months ago	455 votes	
Interactive Intro to Dimensi... ● · 4 months ago	414 votes	

Discussion Expert		
Current Rank 134 of 56,605	Highest Rank 87	
● 1	● 2	● 100
Nervous to get in the game ● · 10 months ago	15 votes	
For newbies ● · 6 months ago	6 votes	
EDA: Top 23 users in Kerne... ● · 4 months ago	5 votes	

My first kaggle story



YouHan Lee • Posted on Version 38 • 7 months ago • Options • Edit • Reply

Wow...amazing! As a beginner, I'm really impressed by your work. I envy your art of analysis. :) Thanks for your kernel because your analysis gave me valuable insight!

Actually, I want to ask you about something.

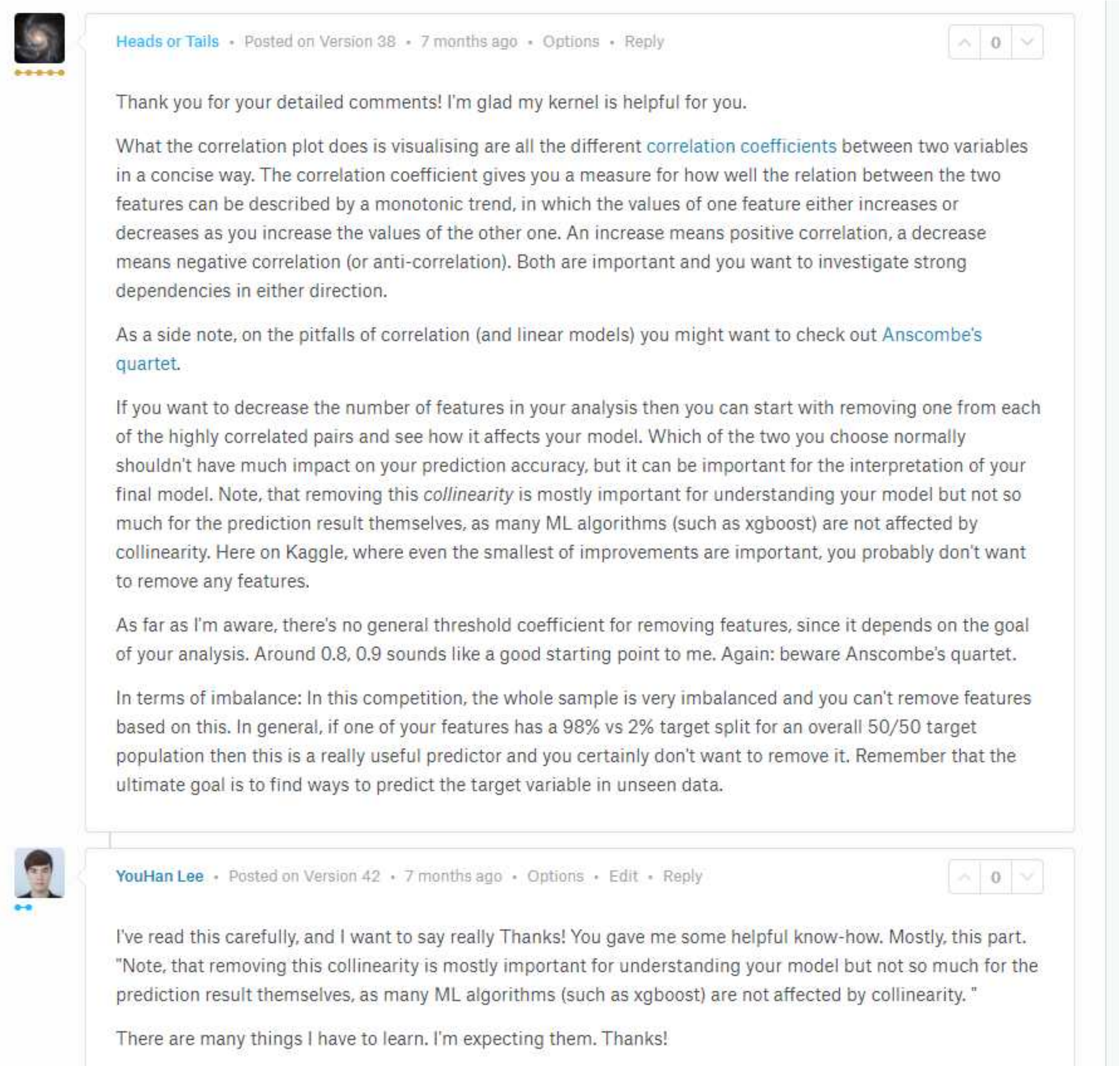
1. Selecting or removing features in correlation plot. Correlation plot is useful to see the dependency between one feature and other feature in 2D. Actually, plotting them is quite easy thanks to helpful packages. But, using properly is quite difficult. I got a correlation plot and see the 2-D correlation. But, I don't know the criterion to select or remove features from correlation plot. If either negative or positive correlation exists between two features, do I need to select both? or select only one feature because of some dependency? And, what value is the criterion of high dependency commonly, larger than 0.5? or 0.8?

Otherwise, If no correlation exists(the coefficient is about zero), do I need to select both?

1. Impact of imbalanced data I found that some feature has the imbalance on the quantitive amount of each class. For example, 1 class has a 98%, 0 class has a 2%. In this case, is correct that I need to delete the feature because this imbalance will cause some bias on the ML model?

Thanks!!

My first kaggle story – 범접할 수 없는 은하계 고수



Heads or Tails • Posted on Version 38 • 7 months ago • Options • Reply

Thank you for your detailed comments! I'm glad my kernel is helpful for you.

What the correlation plot does is visualising are all the different correlation coefficients between two variables in a concise way. The correlation coefficient gives you a measure for how well the relation between the two features can be described by a monotonic trend, in which the values of one feature either increases or decreases as you increase the values of the other one. An increase means positive correlation, a decrease means negative correlation (or anti-correlation). Both are important and you want to investigate strong dependencies in either direction.

As a side note, on the pitfalls of correlation (and linear models) you might want to check out Anscombe's quartet.

If you want to decrease the number of features in your analysis then you can start with removing one from each of the highly correlated pairs and see how it affects your model. Which of the two you choose normally shouldn't have much impact on your prediction accuracy, but it can be important for the interpretation of your final model. Note, that removing this *collinearity* is mostly important for understanding your model but not so much for the prediction result themselves, as many ML algorithms (such as xgboost) are not affected by collinearity. Here on Kaggle, where even the smallest of improvements are important, you probably don't want to remove any features.

As far as I'm aware, there's no general threshold coefficient for removing features, since it depends on the goal of your analysis. Around 0.8, 0.9 sounds like a good starting point to me. Again: beware Anscombe's quartet.

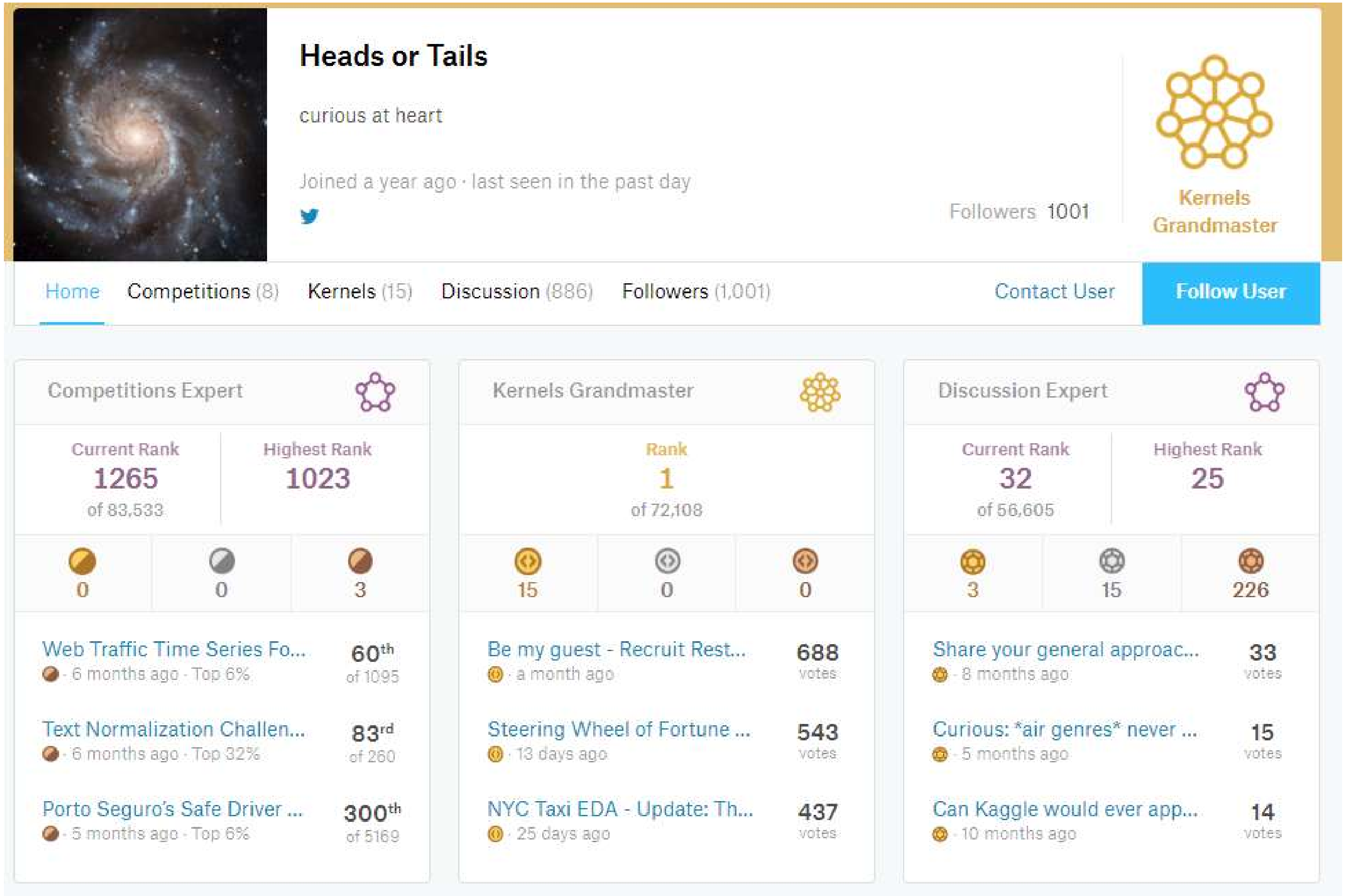
In terms of imbalance: In this competition, the whole sample is very imbalanced and you can't remove features based on this. In general, if one of your features has a 98% vs 2% target split for an overall 50/50 target population then this is a really useful predictor and you certainly don't want to remove it. Remember that the ultimate goal is to find ways to predict the target variable in unseen data.

YouHan Lee • Posted on Version 42 • 7 months ago • Options • Edit • Reply

I've read this carefully, and I want to say really Thanks! You gave me some helpful know-how. Mostly, this part. "Note, that removing this collinearity is mostly important for understanding your model but not so much for the prediction result themselves, as many ML algorithms (such as xgboost) are not affected by collinearity. "

There are many things I have to learn. I'm expecting them. Thanks!

My first kaggle story



Heads or Tails

curious at heart

Joined a year ago · last seen in the past day

Followers 1001

Kernels Grandmaster

Home | Competitions (8) | Kernels (15) | Discussion (886) | Followers (1,001) | Contact User | Follow User

Competitions Expert

Current Rank	Highest Rank
1265 of 83,533	1023

●	●	●
0	0	3

Web Traffic Time Series Fo... ● · 6 months ago · Top 6%	60th of 1095
Text Normalization Challen... ● · 6 months ago · Top 32%	83rd of 260
Porto Seguro's Safe Driver ... ● · 5 months ago · Top 6%	300th of 5169

Kernels Grandmaster

Rank
1 of 72,108

●	●	●
15	0	0

Be my guest - Recruit Rest... ● · a month ago	688 votes
Steering Wheel of Fortune ... ● · 13 days ago	543 votes
NYC Taxi EDA - Update: Th... ● · 25 days ago	437 votes

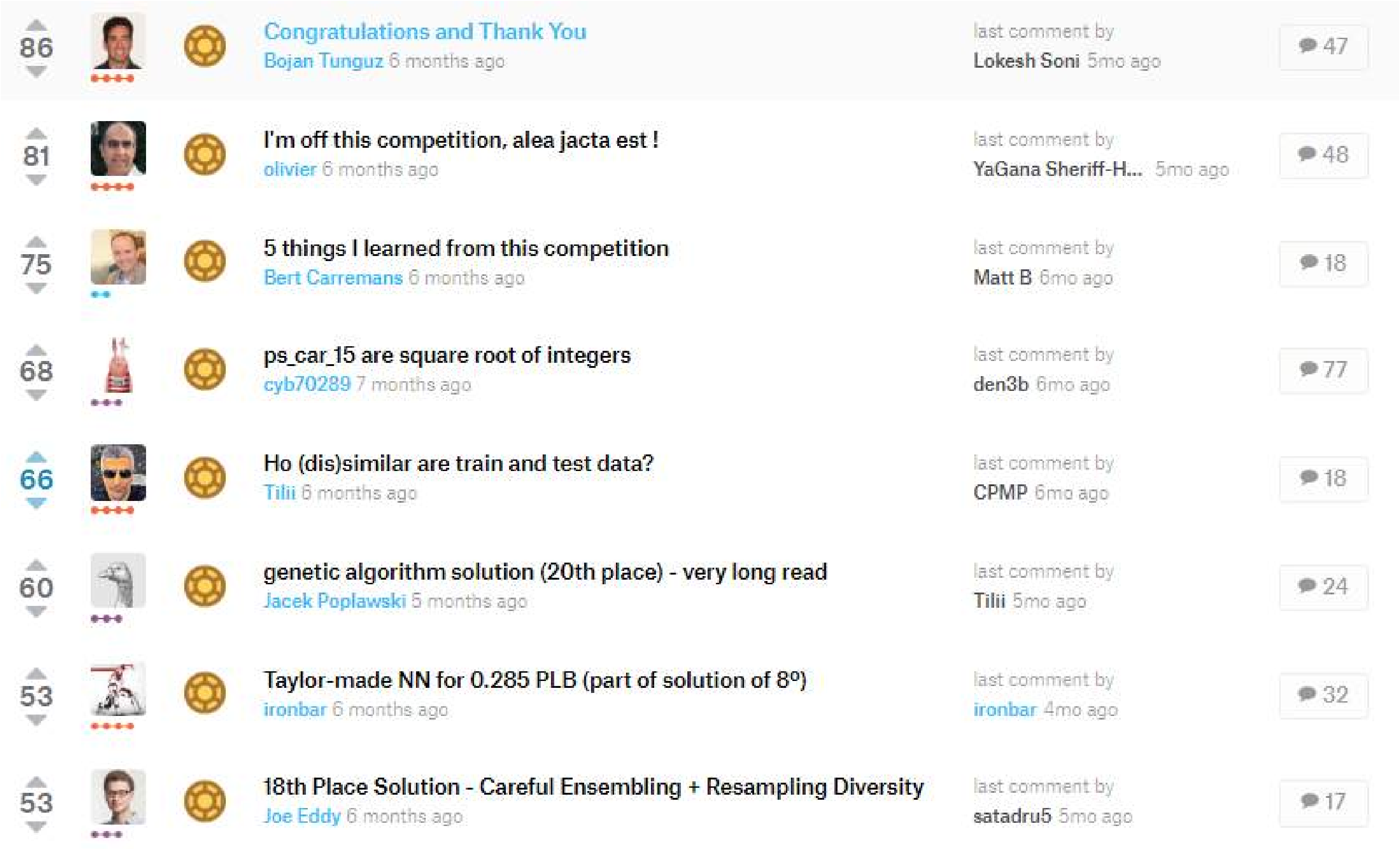
Discussion Expert

Current Rank	Highest Rank
32 of 56,605	25

●	●	●
3	15	226

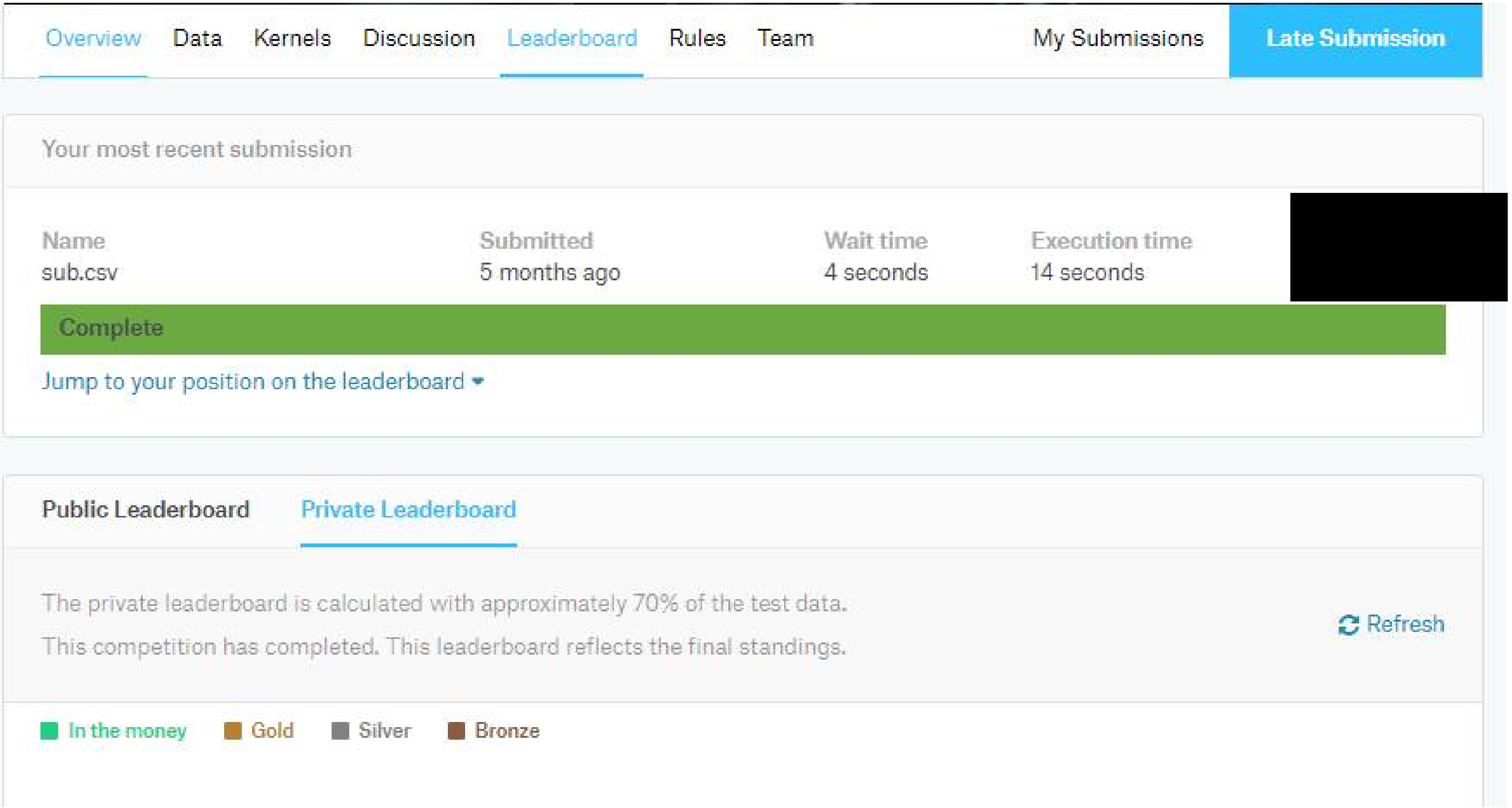
Share your general approac... ● · 8 months ago	33 votes
Curious: *air genres* never ... ● · 5 months ago	15 votes
Can Kaggle would ever app... ● · 10 months ago	14 votes

My first kaggle story – Get insight from discussion



86	Congratulations and Thank You Bojan Tunguz 6 months ago	last comment by Lokesh Soni 5mo ago	47
81	I'm off this competition, alea jacta est ! olivier 6 months ago	last comment by YaGana Sheriff-H... 5mo ago	48
75	5 things I learned from this competition Bert Carremans 6 months ago	last comment by Matt B 6mo ago	18
68	ps_car_15 are square root of integers cyb70289 7 months ago	last comment by den3b 6mo ago	77
66	Ho (dis)similar are train and test data? Tilii 6 months ago	last comment by CPMP 6mo ago	18
60	genetic algorithm solution (20th place) - very long read Jacek Poplawski 5 months ago	last comment by Tilii 5mo ago	24
53	Taylor-made NN for 0.285 PLB (part of solution of 8º) ironbar 6 months ago	last comment by ironbar 4mo ago	32
53	18th Place Solution - Careful Ensembling + Resampling Diversity Joe Eddy 6 months ago	last comment by satadru5 5mo ago	17

My first kaggle story – Submission



Your most recent submission

Name	Submitted	Wait time	Execution time	
sub.csv	5 months ago	4 seconds	14 seconds	[redacted]

Complete

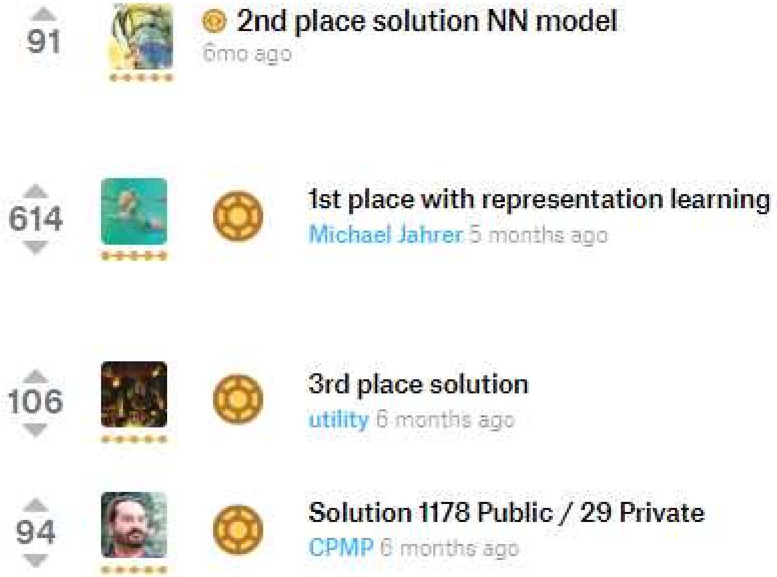
Jump to your position on the leaderboard ▾

The private leaderboard is calculated with approximately 70% of the test data.

This competition has completed. This leaderboard reflects the final standings.

🟩 In the money 🟧 Gold ⬜ Silver 🟫 Bronze

My first kaggle story – After competition



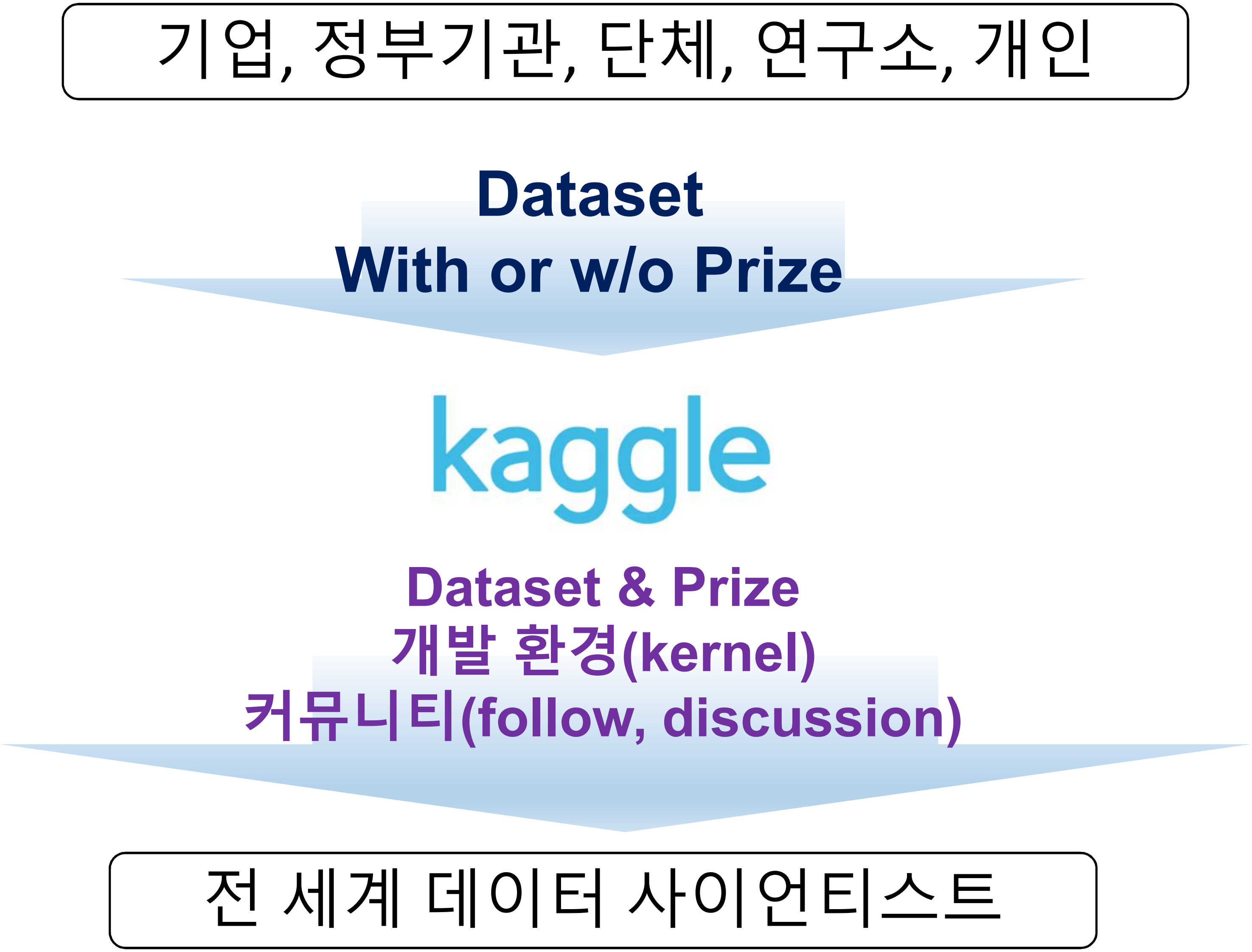
91 2nd place solution NN model
6mo ago

614 1st place with representation learning
Michael Jahrer 5 months ago

106 3rd place solution
utility 6 months ago

94 Solution 1178 Public / 29 Private
CPMP 6 months ago

Data Playground for 데이터 과학자!



기업, 정부기관, 단체, 연구소, 개인

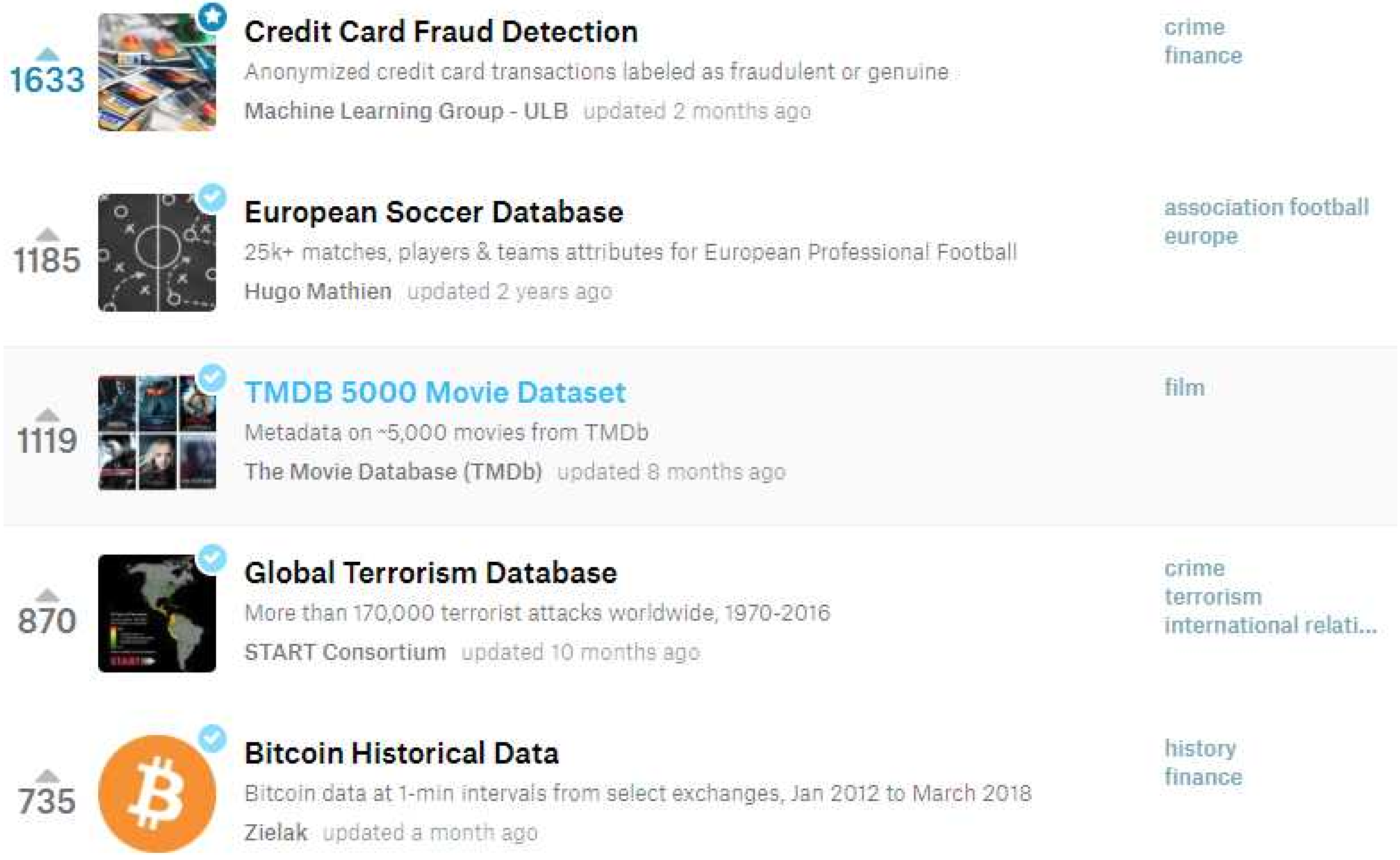
Dataset
With or w/o Prize

kaggle

Dataset & Prize
개발 환경(kernel)
커뮤니티(follow, discussion)

전 세계 데이터 사이언티스트

Competition Example – Tensorflow competition



1633 **Credit Card Fraud Detection**
Anonymized credit card transactions labeled as fraudulent or genuine
Machine Learning Group - ULB updated 2 months ago
crime
finance

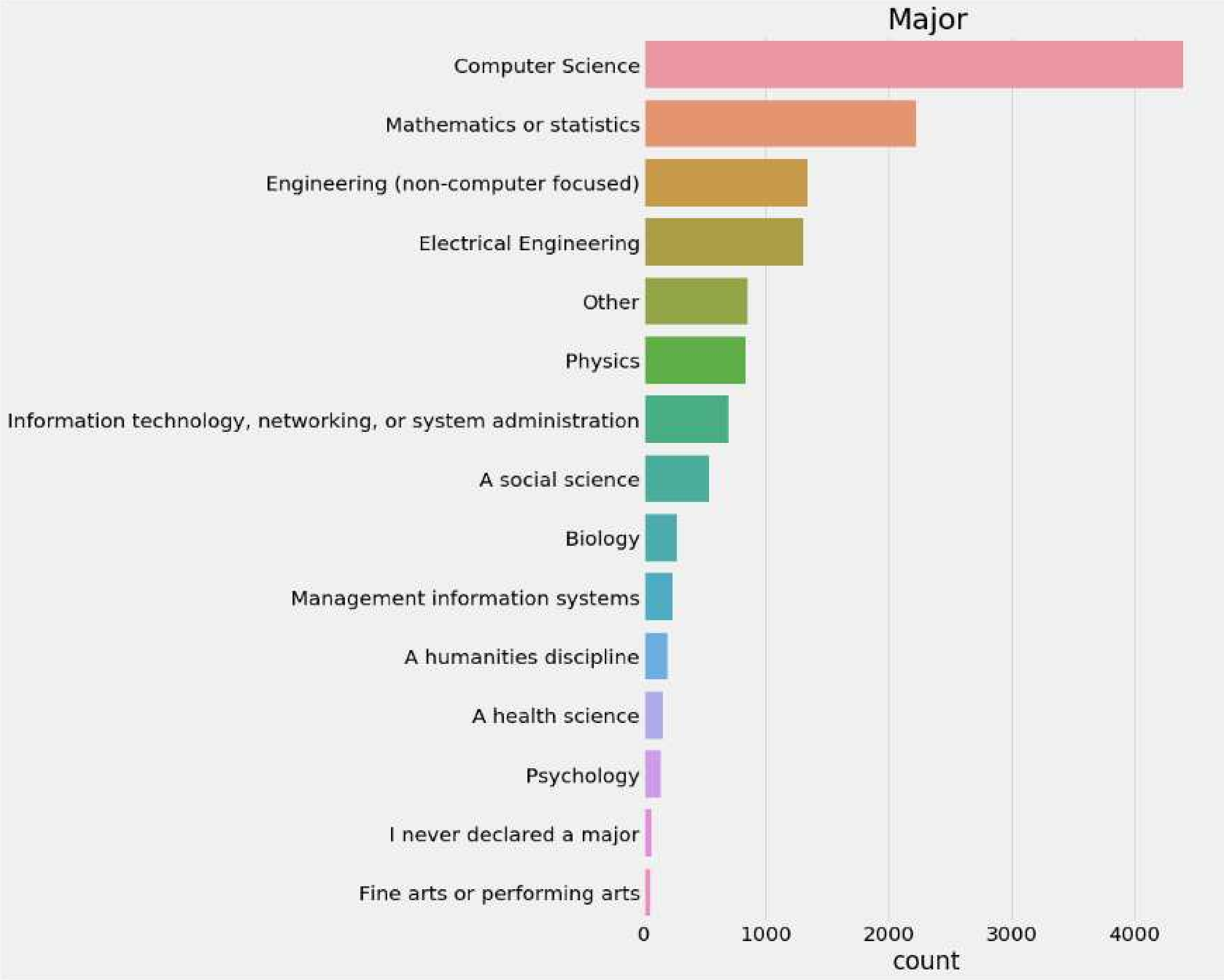
1185 **European Soccer Database**
25k+ matches, players & teams attributes for European Professional Football
Hugo Mathien updated 2 years ago
association football
europe

1119 **TMDB 5000 Movie Dataset**
Metadata on ~5,000 movies from TMDb
The Movie Database (TMDb) updated 8 months ago
film

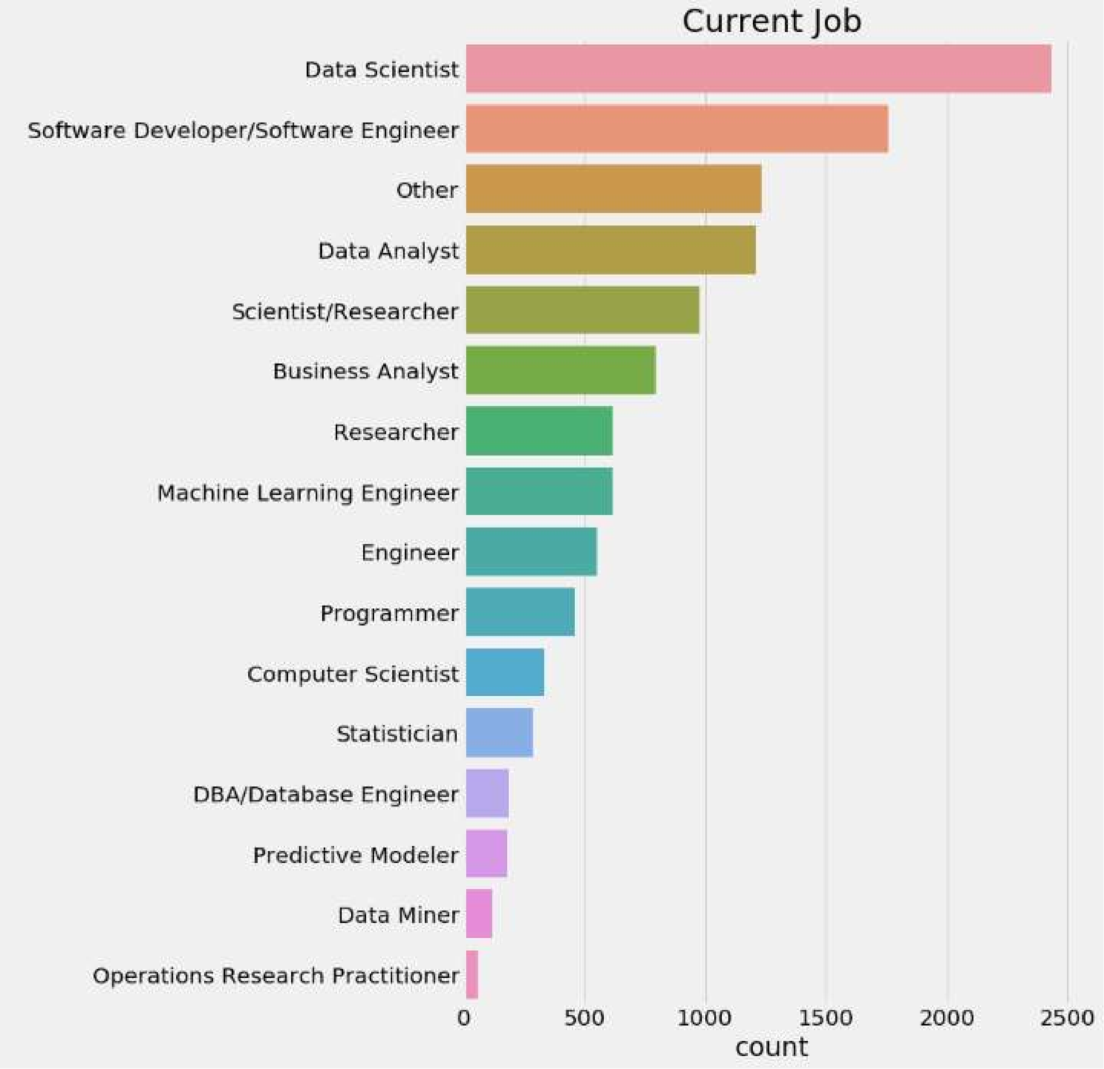
870 **Global Terrorism Database**
More than 170,000 terrorist attacks worldwide, 1970-2016
START Consortium updated 10 months ago
crime
terrorism
international relati...

735 **Bitcoin Historical Data**
Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to March 2018
Zielak updated a month ago
history
finance

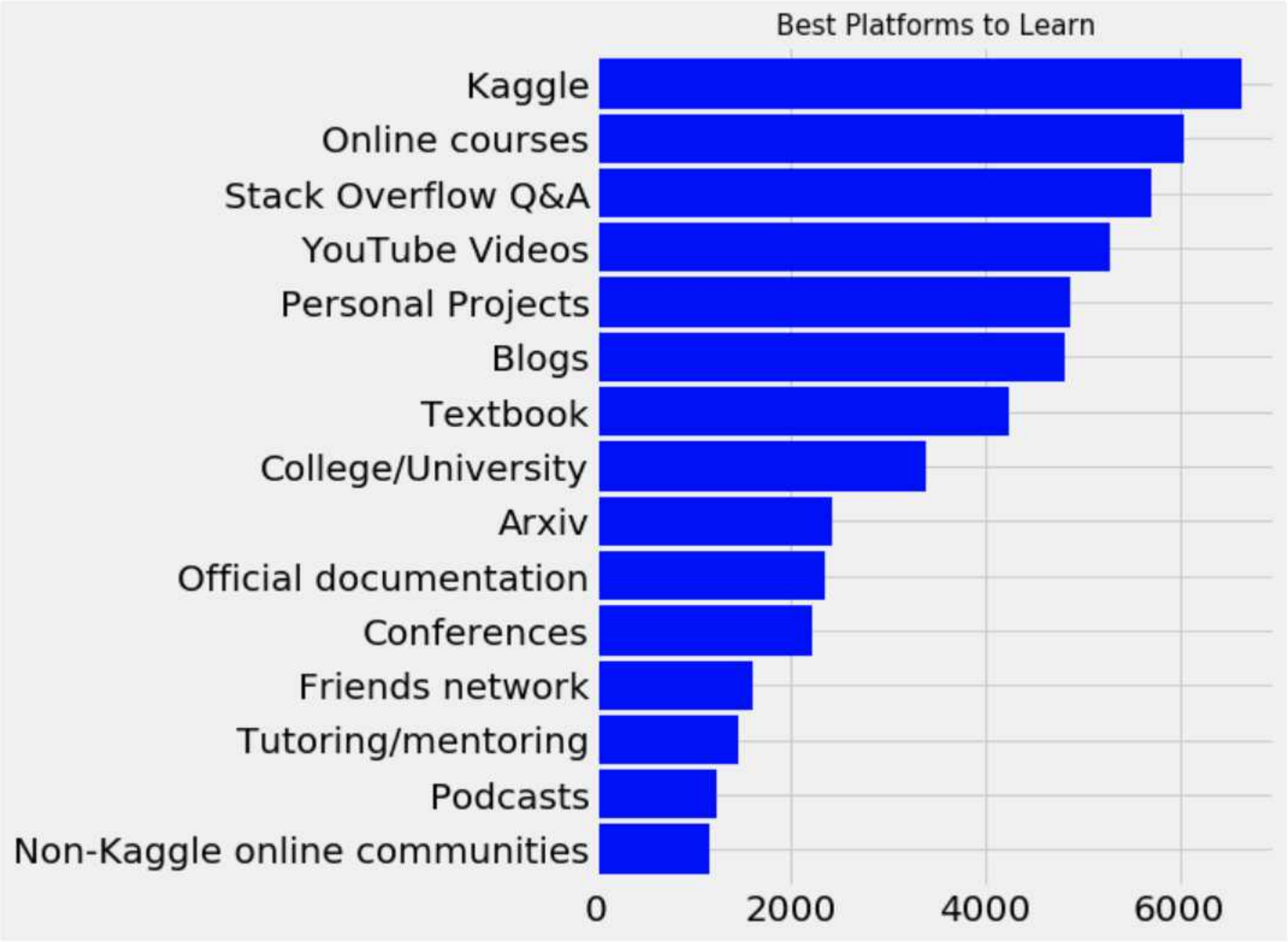
2017 Kaggle survey – 캐글러 전공



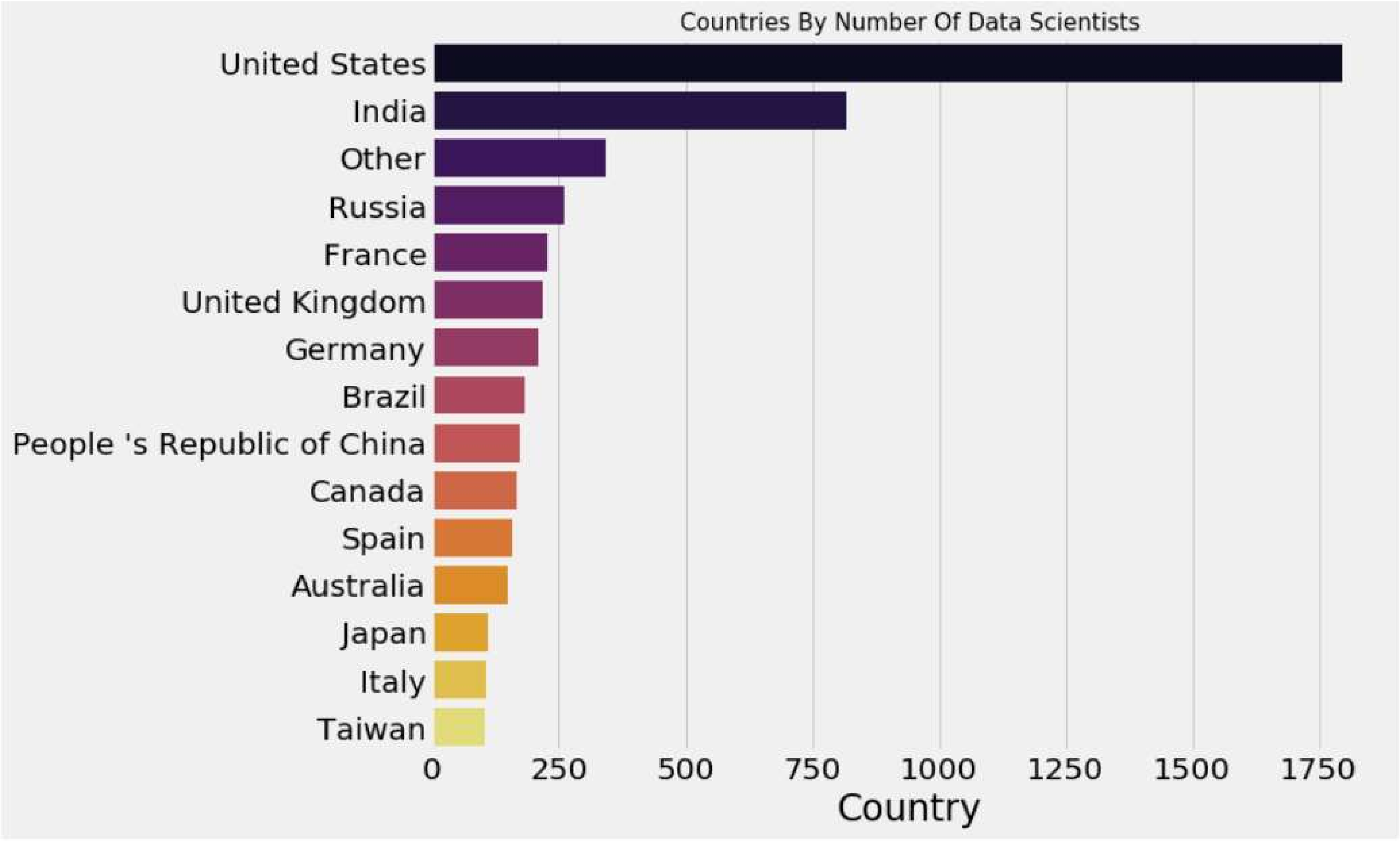
2017 Kaggle survey – 캐글러 현재 직업



2017 Kaggle survey – 캐글러는 어디서 배우나?



2017 Kaggle survey – 캐글러들의 국적?



캐글에서 뭘 얻을 수 있나?

1. 현업에 바로 적용 가능한 데이터 분석 기술 습득
2. 현업에 바로 적용 가능한 머신러닝, 딥러닝 알고리즘 적용 방법 습득
3. 인적 네트워크 형성

캐글 합시다!