Kaggle이란?

Kaggle 이란?

kaggle

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)

이터 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.

I Do Not Accept I Understand and Accept

Kaggle 에서 competition 을 주최한 단체, 기업들

여러 competition 들

Mercedes-Benz Greener Manufacturing — $25,000

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

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

Quora Question Pairs — $25,000

Can you identify question pairs that have the same intent?

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

Passenger Screening Algorithm Challenge — $1,500,000

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

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

Bosch Production Line Performance — $30,000

Reduce manufacturing failures

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

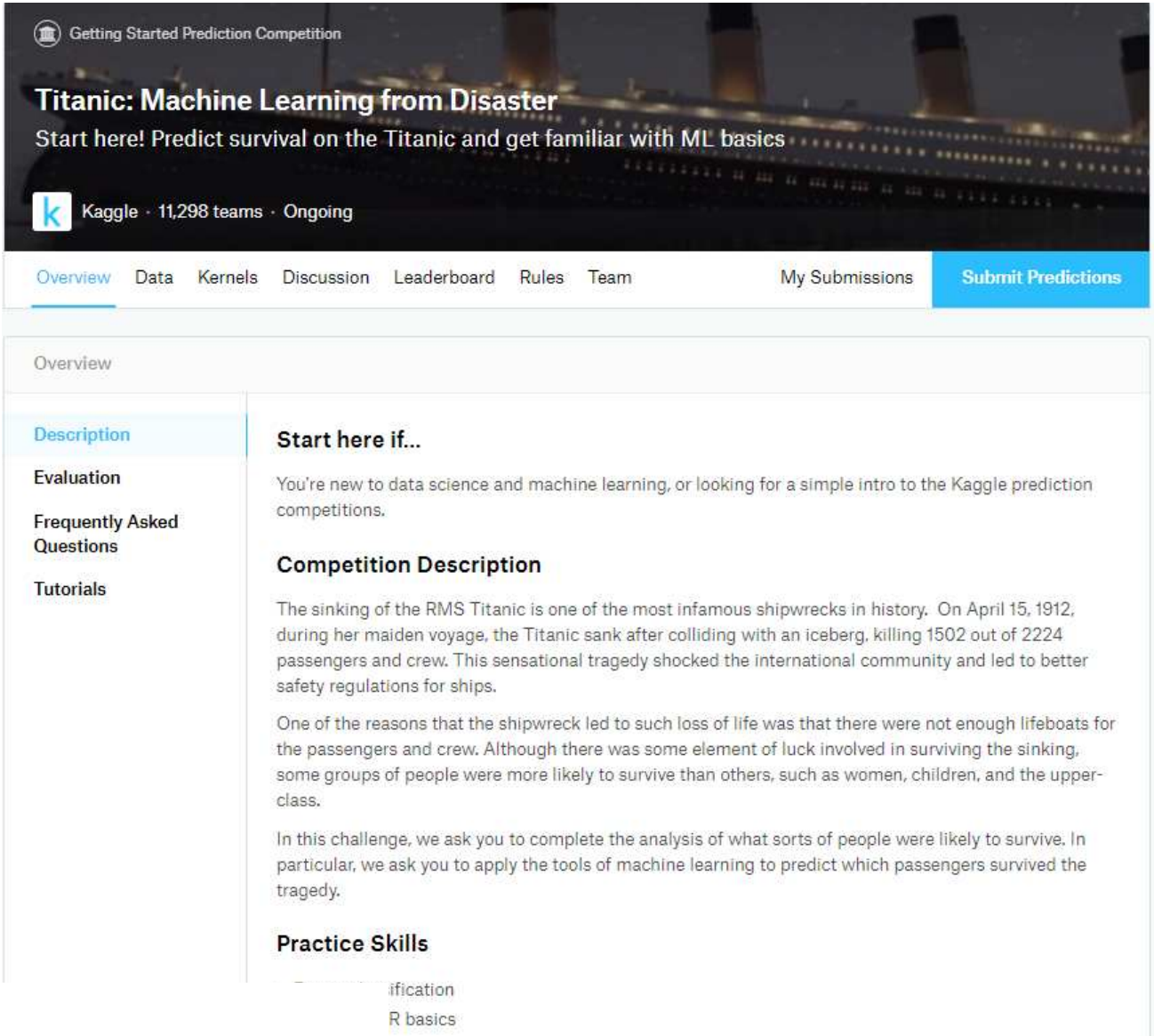
여지껏 다뤄본 것이
IRIS dataset, MNIST 뿐인데

저런 걸 어떻게
분석해야 하나?

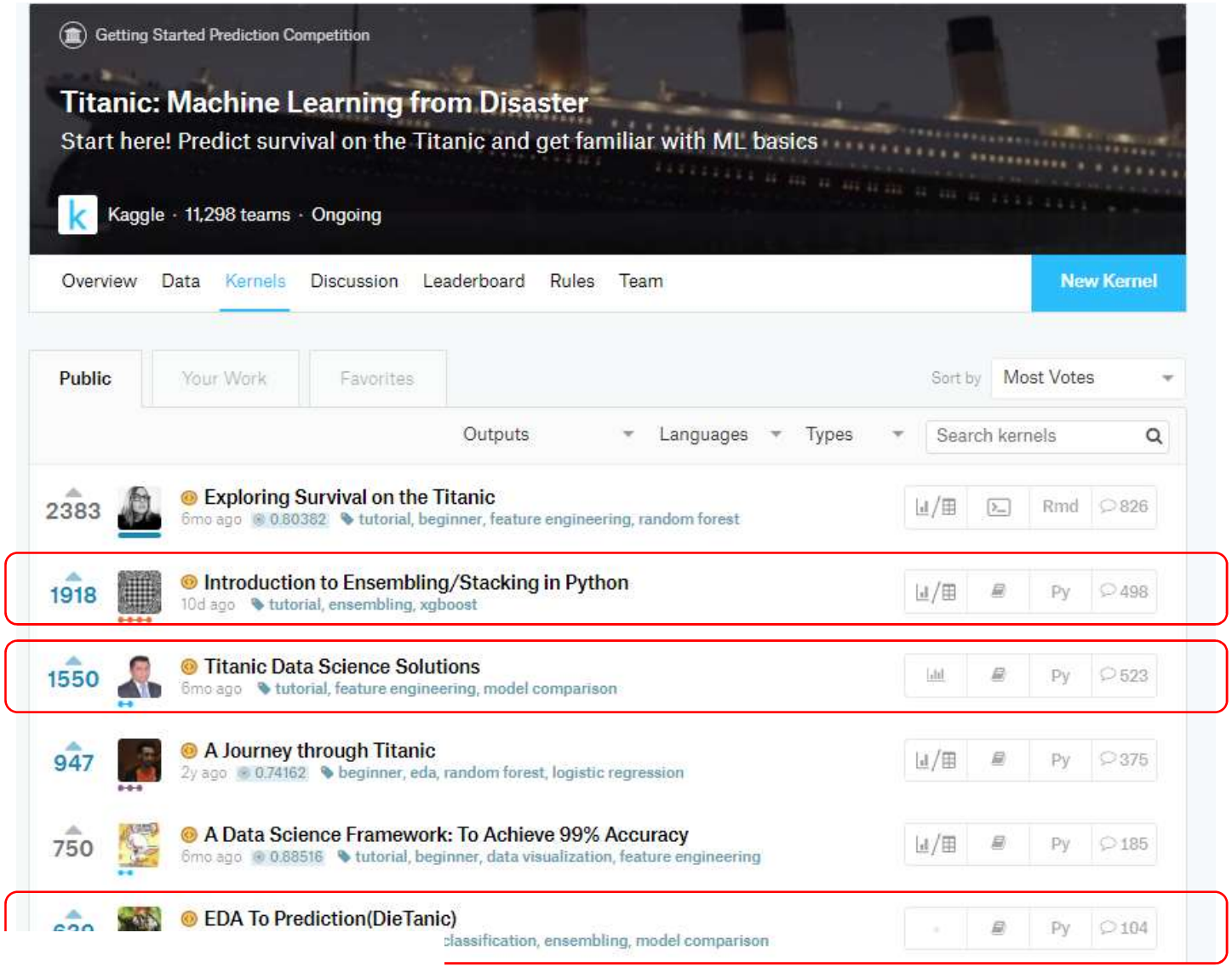
모방은 창조의 시작

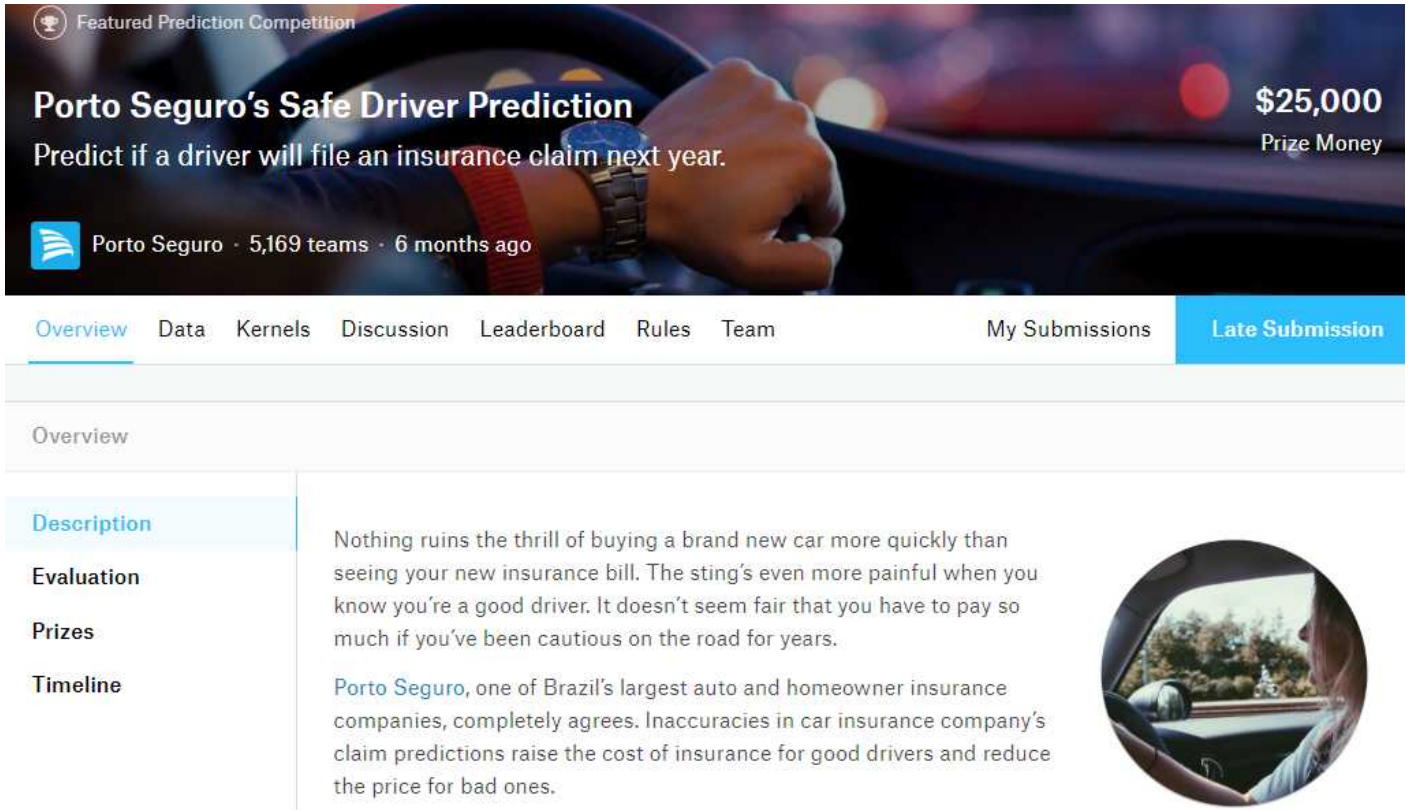
공부해서 함께 나누자! – 캐글 속 선순환

Titanic competition – Can you predict survival?



Titanic competition – Study with voted kernels!





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

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

115 Reconstruction of 'ps_reg_03'
8mo ago

94 Dimensionality reduction (PCA, tSNE)
8mo ago dimensionality reduction

80 Noise analysis of Porto Seguro's features
8mo ago

52 Bayesian Optimization of XGBoost Parameters
8mo ago

226 XGBoost CV (LB .284)
8mo ago 0.28456 gradient boosting, xgboost

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 ^^

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.

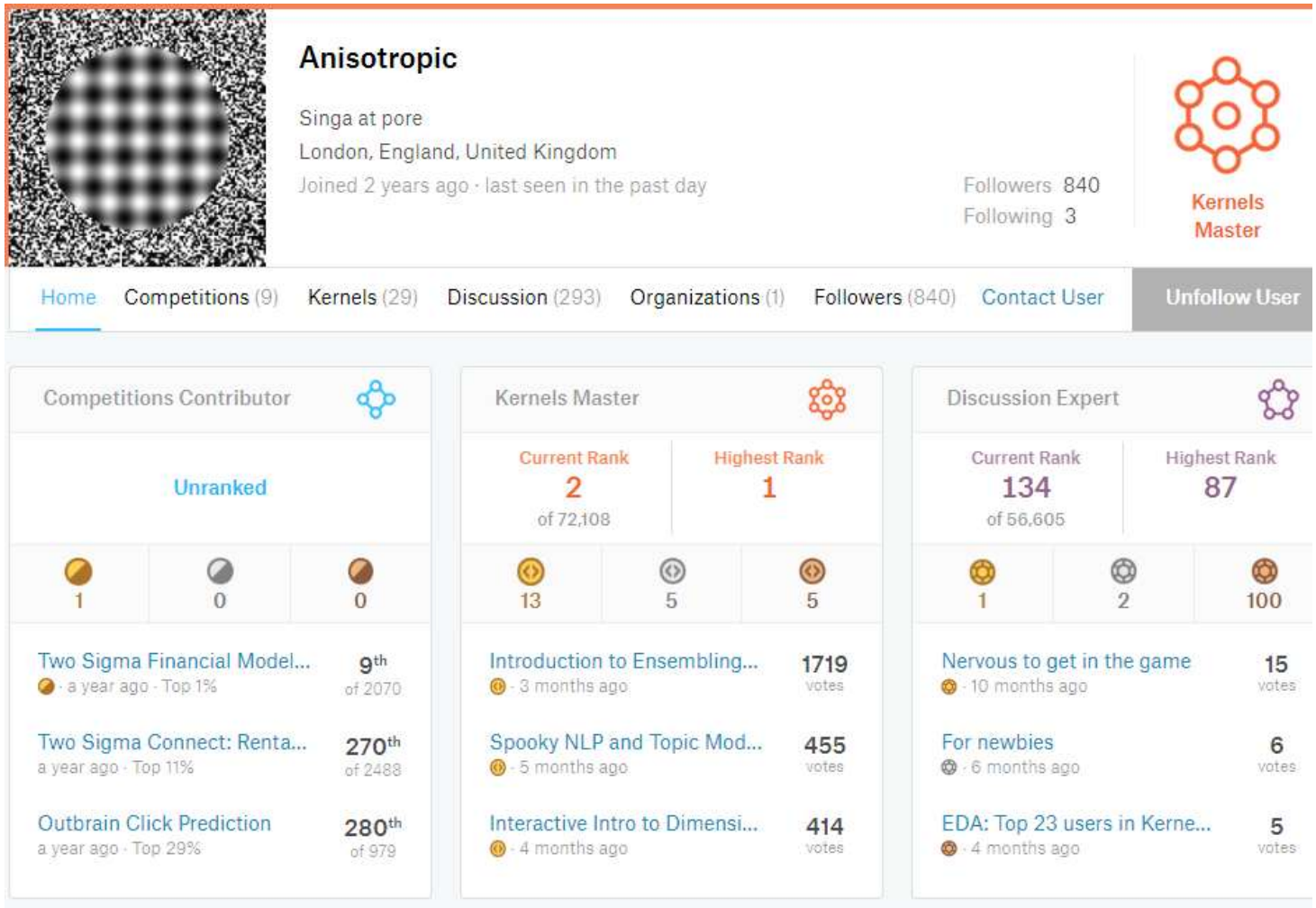
With regards to future work, I noted that you said you maybe thinking of using a PCA analysis. That would be interesting to note what comes out of it, especially so as we have noted that there are quite a lack of linearly correlated features. If time permits, it would also be interesting to try out nonlinear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

stion! :)


Anisotropic
Singa at pore
London, England, United Kingdom
Joined 2 years ago · last seen in the past day
Followers 840
Following 3
Kernels Master
Home
Competitions (9)
Kernels (29)
Discussion (293)
Organizations (1)
Followers (840)
Contact User
Unfollow User
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 1st kaggle race – 은하계 고수의 가르침

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

0

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!!

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

0

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

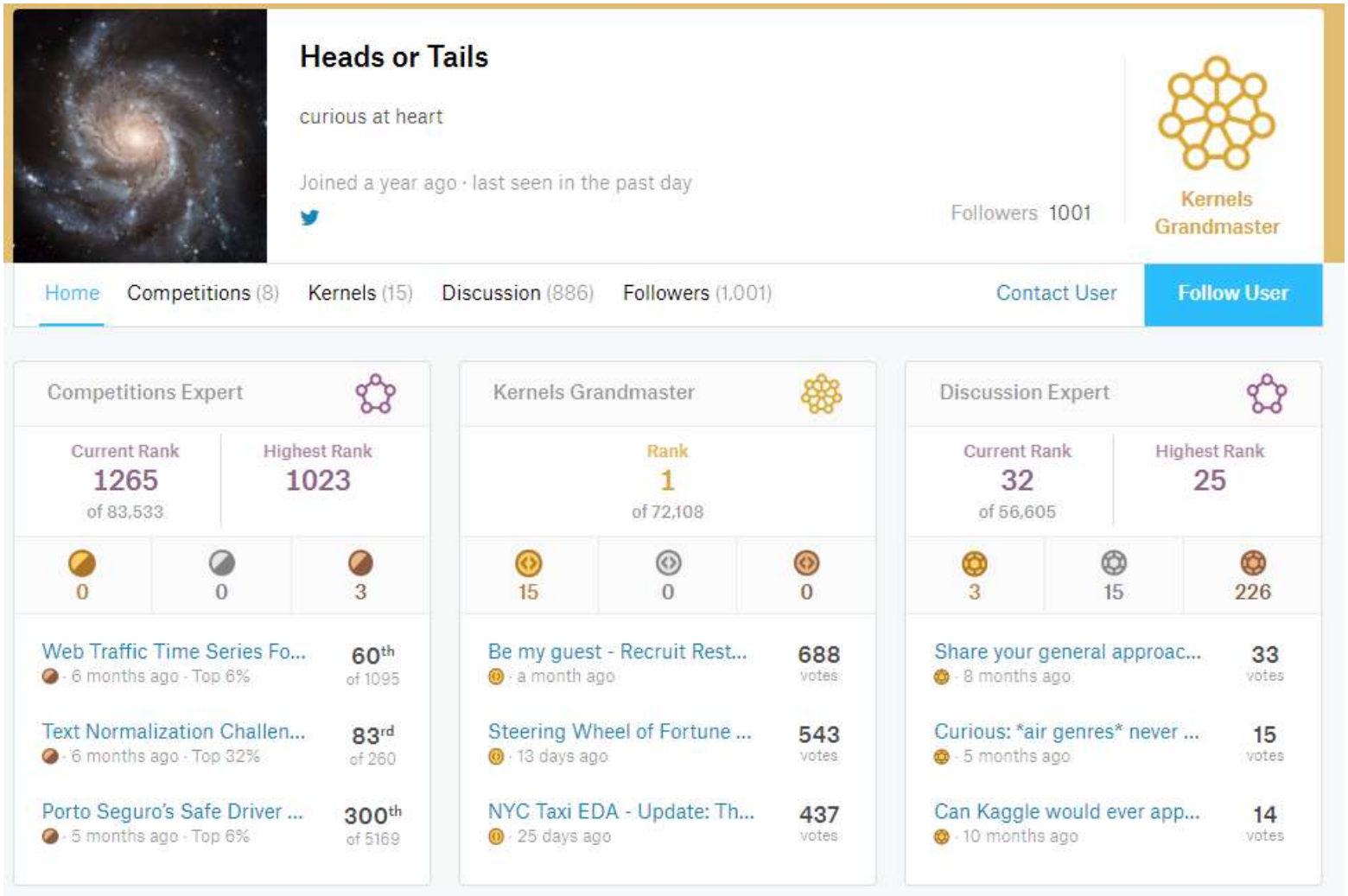
0

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

many ML algorithms (such as xgboost) are not affected by collinearity. "

learn. I'm expecting them. Thanks!

My 1st kaggle race – 1st rank grandmaster!



YouHan Lee • Posted on Latest Version • 6 months ago • Options • Edit • Reply 2

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 2

@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.

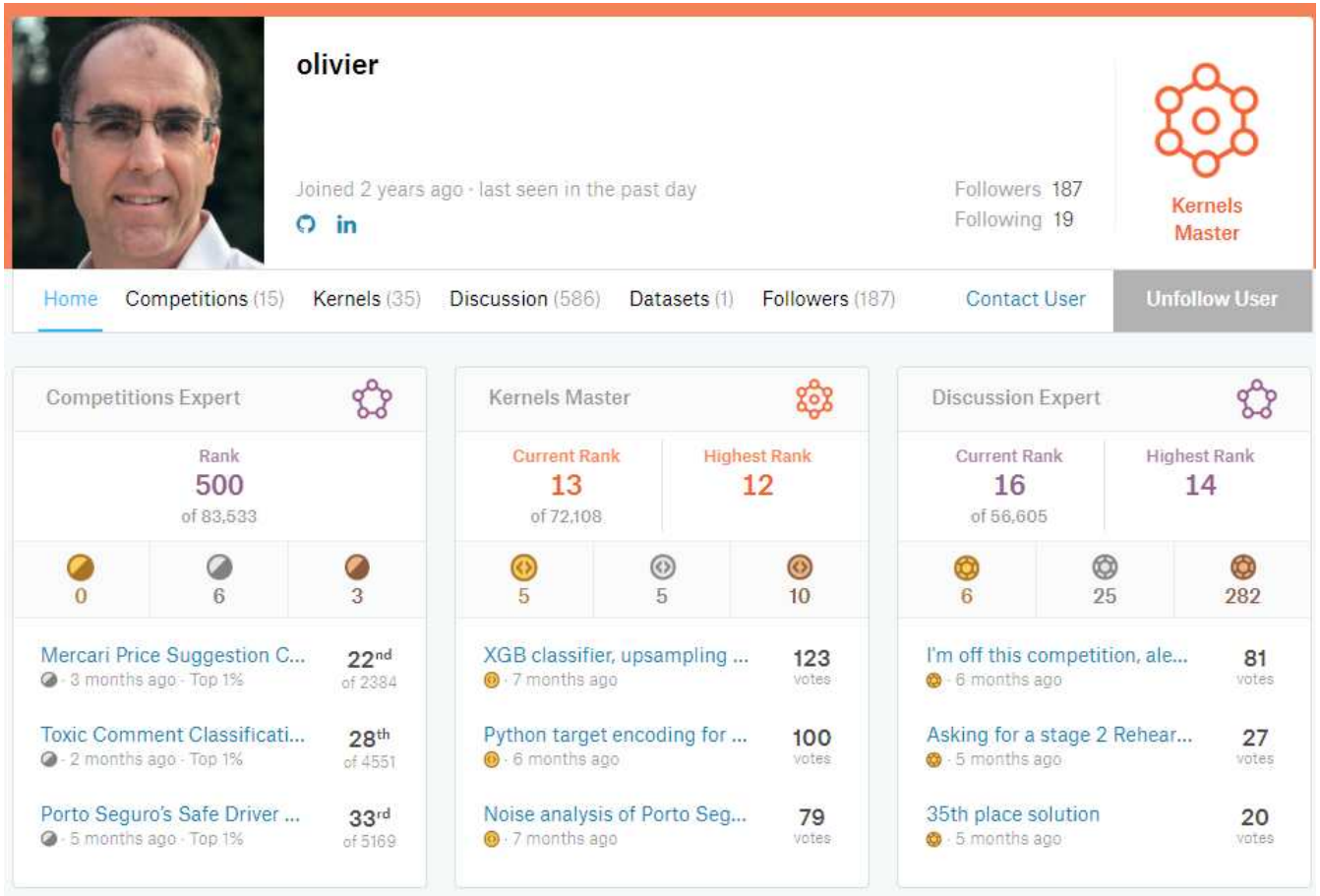
YouHan Lee • Posted on Latest Version • 6 months ago • Options • Edit • Reply 1

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 1

@YouHan Lee, thank you for your kind words.

My 1st kaggle race – 친절한 올리비에 아저씨



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

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Late 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

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

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

91 2nd place solution NN model
6mo ago

106 3rd place solution
utility 6 months ago

94 Solution 1178 Public / 29 Private
CPMP 6 months ago

- After_bayesian_LGB.ipynb
- bayesian_random_forest.ipynb
- ensemble.ipynb
- ensemble_2_gbc-Copy1.ipynb
- ensemble_2_gbc.ipynb
- ensemble_2_lgb_original.ipynb
- ensemble_2_lgb_polynomial-Copy1.ipynb
- ensemble_2_lgb_polynomial.ipynb
- ensemble_3_model.ipynb
- GBM_pc2.ipynb
- Gini_Coefficient.ipynb
- Interactive_Porto_Insights_A_plot_ly_tutorials.ipynb
- Keras_test.ipynb
- Kinetic_and_transforms_with_last_features.ipynb
- Kinetic_and_transforms_with_last_features_ensemble.ipynb
- Kinetic_and_transforms_with_last_features_ensemble_RF_tuning-Real.ipynb
- Kinetic_and_transforms_with_last_features_ensemble_RF_tuning_backup.ipynb
- My_analysis.ipynb
- My_analysis_ver_2_with_median.ipynb
- My_analysis_ver_2_with_median_and_stratified_without_calc-Copy
- My_analysis_ver_2_with_median_and_stratified_without_calc.ipynl
- My_analysis_ver_2_without_null_data.ipynb
- My_analysis_ver_3_CORR_drop_DATA.ipynb
- My_analysis_ver_3_CORR_drop_DATA_with_param_optim.ipynb
- My_analysis_ver_3_NULL_with_ML.ipynb
- My_analysis_ver_4_Probability_more_feature.ipynb
- My_analysis_ver_4_Probability_more_feature_ensemble.ipynb
- My_analysis_ver_4_Probability_without_calc.ipynb
- My_anaysis_5.ipynb
- My_anaysis_5_ensemble.ipynb
- New_feature_pc2_lgb_baysian.ipynb
- New_feature_pc2_rf.ipynb
- py2_lgb_by.ipynb
- Simple_Safe_Driver_Prediction_EDA.ipynb
- Simple_XGBoost_BTB.ipynb
- Study_onehot_encoding.ipynb
- Untitled.ipynb
- Untitled1.ipynb
- Untitled2.ipynb
- xgboost_tutorial_first.ipynb
- xgboost_tutorial_second.ipynb

41 개 주피터 노트북 생성!!!!

배울수 있는 것들

- 데이터 분석에서 머신러닝 모델 생성 및 예측 까지 이어지는 프로세스 경험
- 각종 데이터 분석 라이브러리 사용법 습득
  - Visualization
    - Matplotlib, seaborn, plotly
  - Data analysis
    - Pandas
    - Numpy
  - Machine learning
    - Sklearn
- 머신 러닝 모델 습득
  - Sklearn 내장 알고리즘 들
  - Randomforest
  - Xgboost
  - Lightgbm
- Hyper parameter tunning 방법
  - Gridsearch
  - Randomsearch
  - Baysian optimization
- 머신 러닝 노하우
  - 학습 방법
    - Stratified, shuffle
  - Ensembling
    - Voting, average
- 모델 평가 방법
  - Precision, recall, f1-score, accuracy, AUC
- 영어공부
  - 커널 쓰기, 질문, 응답하며 writing 공부

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!

Sung Kim님이 링크를 공유했습니다.
관리자 · 2017년 11월 22일

[Tensorflow 음성인식 챌린지 같이 참여해요!]
https://www.kaggle.com/c/tensorflow-speech-recognition-chal...

Tensorflow/딥러닝의 활성화와 AI저변확대를 위해 TF-KR에서는 인심 좋은 기업체 후원을 받아 TensorFlow 음성인식 챌린지에 참여할 팀들을 후원하고 멘토링을 지원합니다. (누구나 참여 가능!!). 회의 및 운영비 지원 뿐 아니라 무엇보다 업계 최고 능력자분들의 멘토링을 받으면서 재미있게 챌린지에 참여할 수 있는 절호의 찬스!! 무엇을 망설이시나요? 바로 고고!

[언제]
- 캐글 종료 시간: Mon Jan 16 2018 16:00:00 GMT-0800 (PST)
- 한국 시간: Tue Jan 17 2018 09:00

[무엇]
캐글 TF SR 챌린지에 참여하실 분들을 후원합니다. (간단한 음성을 문자로 바꾸는 문제). 팀을 구성하여 위 캐글 대회 참여를 독려하는 행사입니다.

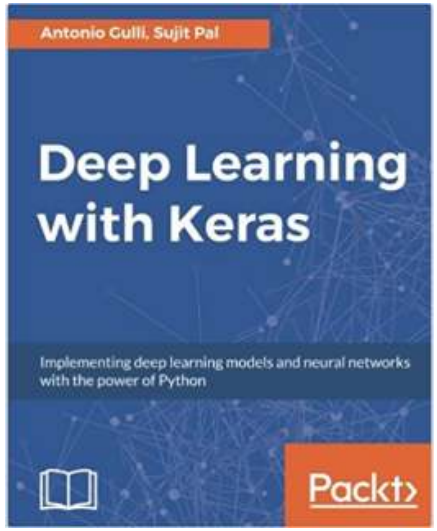
[어떻게]
1. 3인 이상으로 팀을 구성하여 팀원 소개 + 계획 + 기본 알고리즘 계획을 12월 8일까지 제출: http://bit.ly/tfkr_voice
- 팀원 남녀노소제한 없음. 한국국적 1명이상 포함

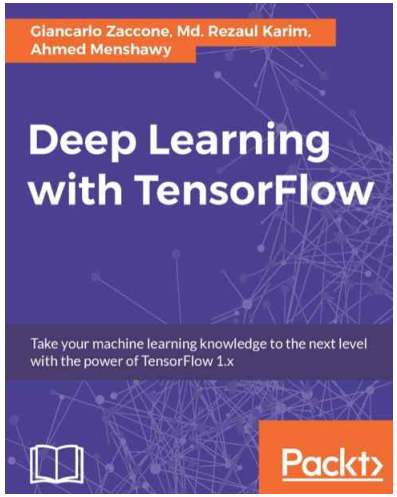
2. 접수팀중 10 팀을 선발하여 회의/운영비, 클라우드 크래딧, 멘토링 지원

[선발된 10팀 지원 내역]
1. 10팀 선발후 팀별로 50만원 (KRW) 후원 (12월 15일 이전 선발)
2. 성공적으로 캐글 챌린지 결과 제출 성공 + 챌린지 후기 모임 발표한 경우 100만원 (KRW) 추가 후원
챌린지 종료후 TF-KR주최의 "챌린지 후기모임" 개최 (TBA)
3. 챌린지 기간중 멘토링 지원: 김태훈 (데브시스터즈), 이찬규 (NAVER Clova Speech), 김훈(카카오), 홍석진 (SKT 음성인식 기술팀), AWS (윤석찬), TF-KR 운영진
4. 클라우드 크레딧 지원 (AWS 예정 - 팀당 USD1,000)
5. 전체 캐글 챌린지 1위 할경우 2,500만원 (KRW) 상금
6. 전체 캐글 챌린지 10위안에 들면 해당팀 1,000만원 (KRW) 상금

국내 모 기업에서 후원하여
+ prize 추가 됨

친한 사람들 3명과
팀을 맺고 시작





그 외 여러
깃허브들!
stackoverflow

620 Speech representation and data exploration
6mo ago beginner, data visualization

178 End-to-end baseline TF Estimator LB 0.72
7mo ago

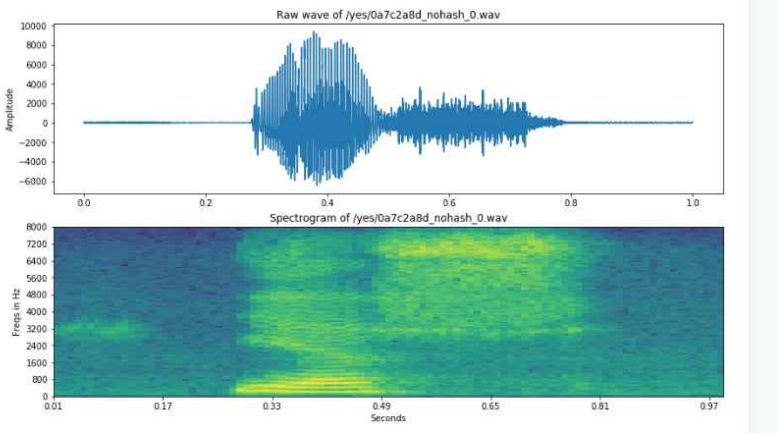
5 Sound Augmentation Librosa
6mo ago sound technology

18 Simple Keras Model with data generator
7mo ago

15 High Resolution Mel Spectrograms
6mo ago data visualization

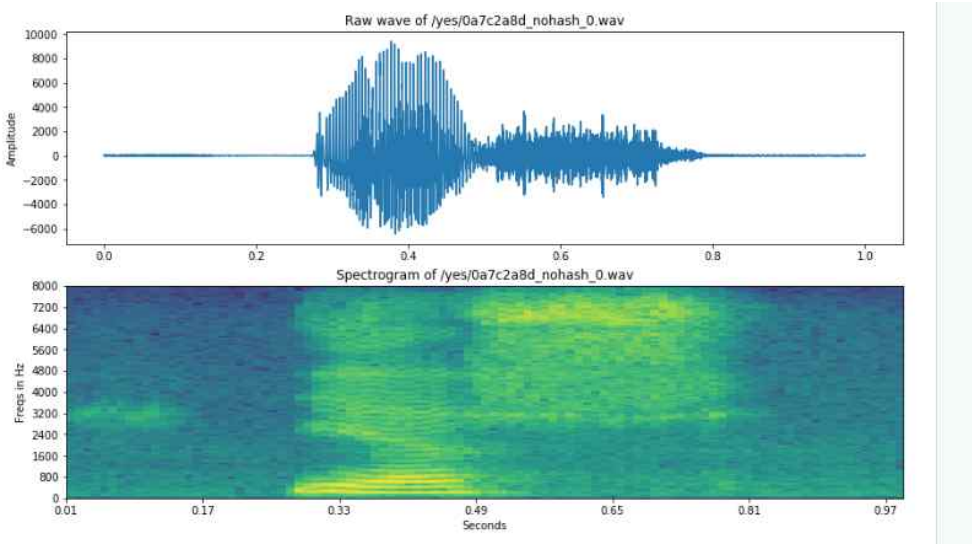
13 Keras Directory Iterator - (LB 0.72)
7mo ago

기본 3번, 내 것으로 될 때 까지 반복



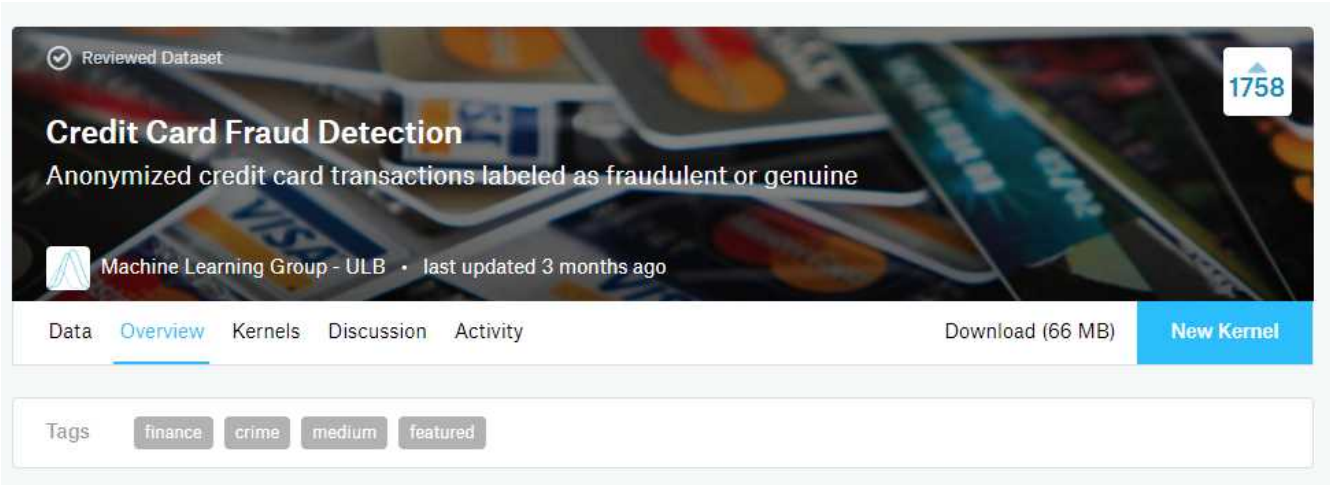
- Audio processing
  - Spectogram
- Deep learning
  - Convolutional neural network(CNN)
    - 1D, 2D
  - Recurrent neural network
    - LSTM
    - GRU
- Deep learning tools
  - Tensorflow
  - Keras
- Deep learning technique
  - Data augmentation
  - Parameter tunning
  - tensorboard

- Time series data 에 특정 signal(outlier)를 판별하는 neural net 을 만들어 보자!



Tensorflow competition 에서 배운 spectrogram + 2D CNN 을 사용해보자!

Anomaly detection 문제로 끌어 가볼까?



Credit card transaction data 에 있는 Fraud(outlier) detection

Time series 에 있는 Outlier detection

커널 공부 시작

Autoencoder 를 활용한 비지도 학습

Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei*
Tsinghua University

Yang Feng, Jie Chen, Zhaogang Wang, Honglin Qiao
Alibaba Group

정상 데이터만 Autoencoder 에 학습 시킴

학습된 neural network 에 비정상 데이터 를 주기

Error(reconstruction error) 가 나옴.
- **How far** an abnormal is from the normal regions

정상 데이터와 비정상 데이터가 잘 구분되는 **threshold** 선택