Machine Learning Algorithms and pipeline with kaggle competitions

이규영
mineatte@gmail.com

Kosscon 2018
2018.11.29

Kaggle 소개

Kaggle Break

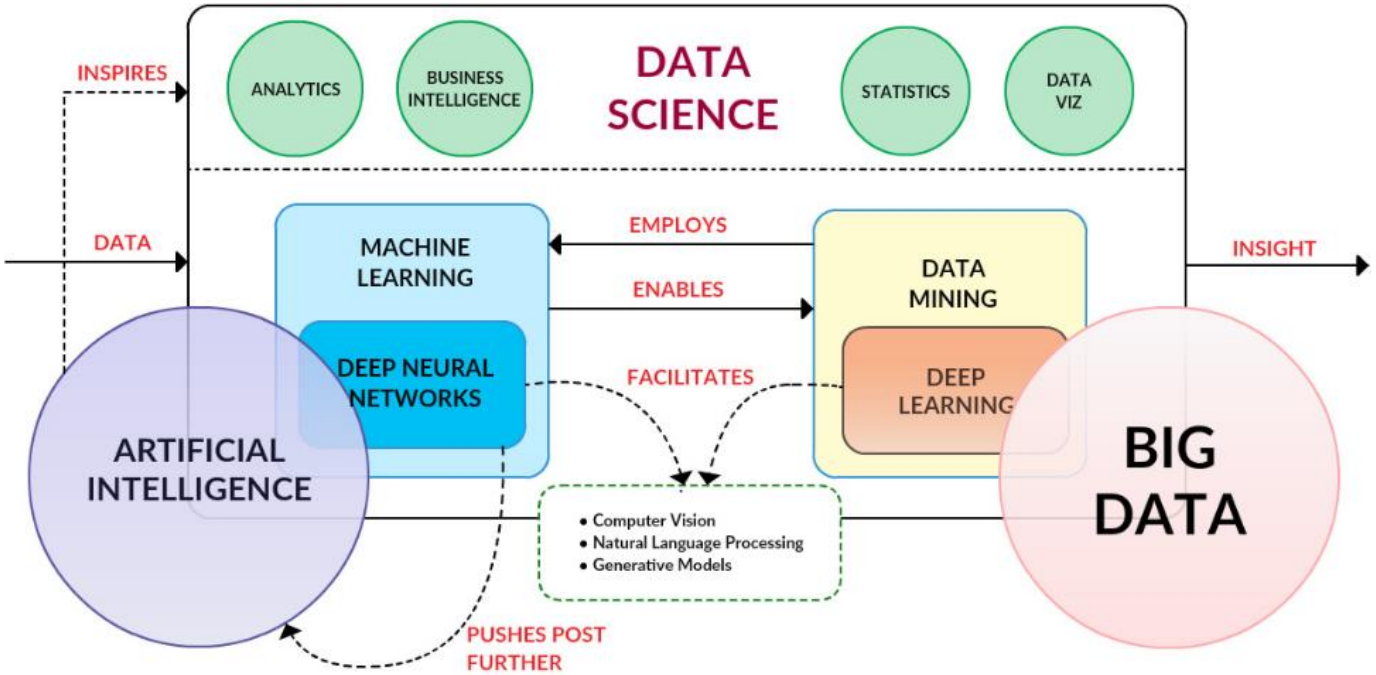
kaggle

2010년에 설립된 머신러닝 경진대회 플랫폼
2017년 구글에게 인수

- 다양한 기업, 단체, 개인들이 Dataset과 문제, 상금을 건 대회를 제시
- 전세계의 Data Scientist, Machine Learning Engineer들이 대회에 참가

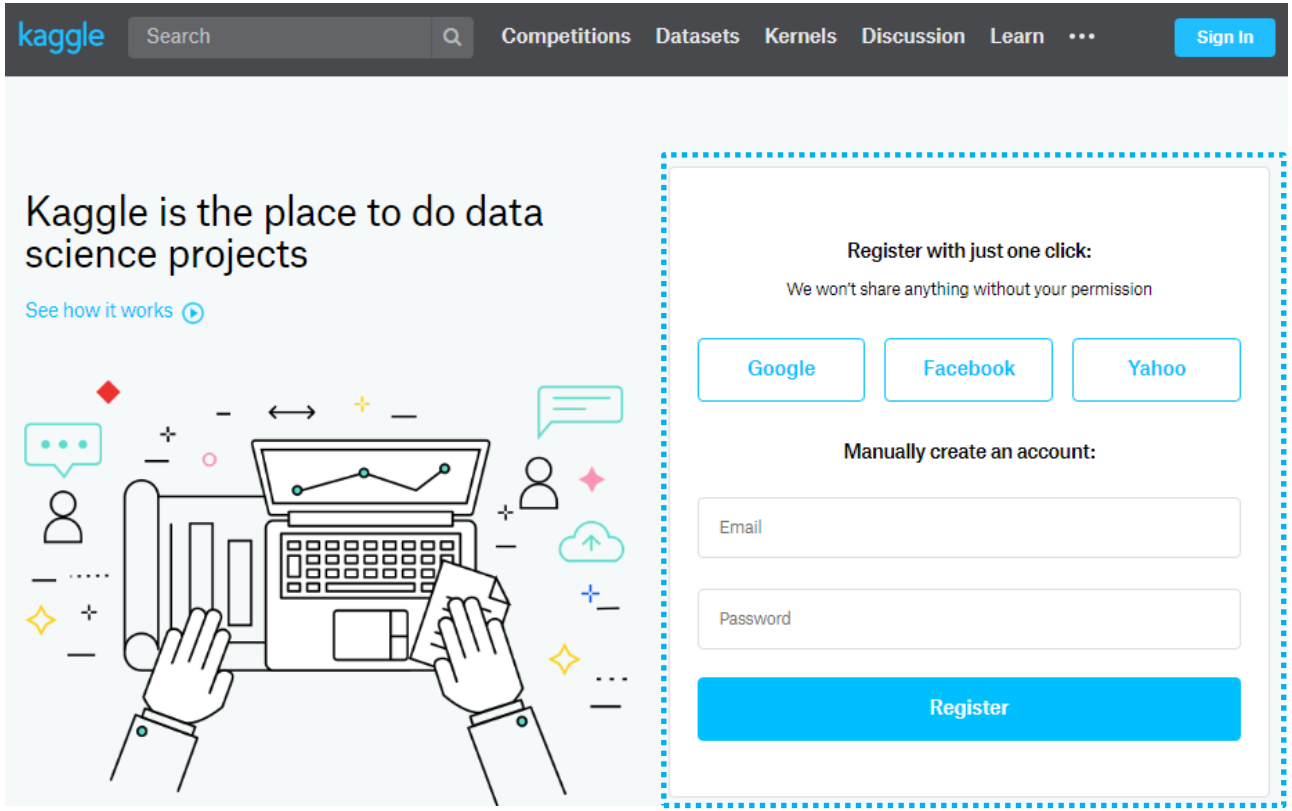
AI VS Data Science

Kaggle Break



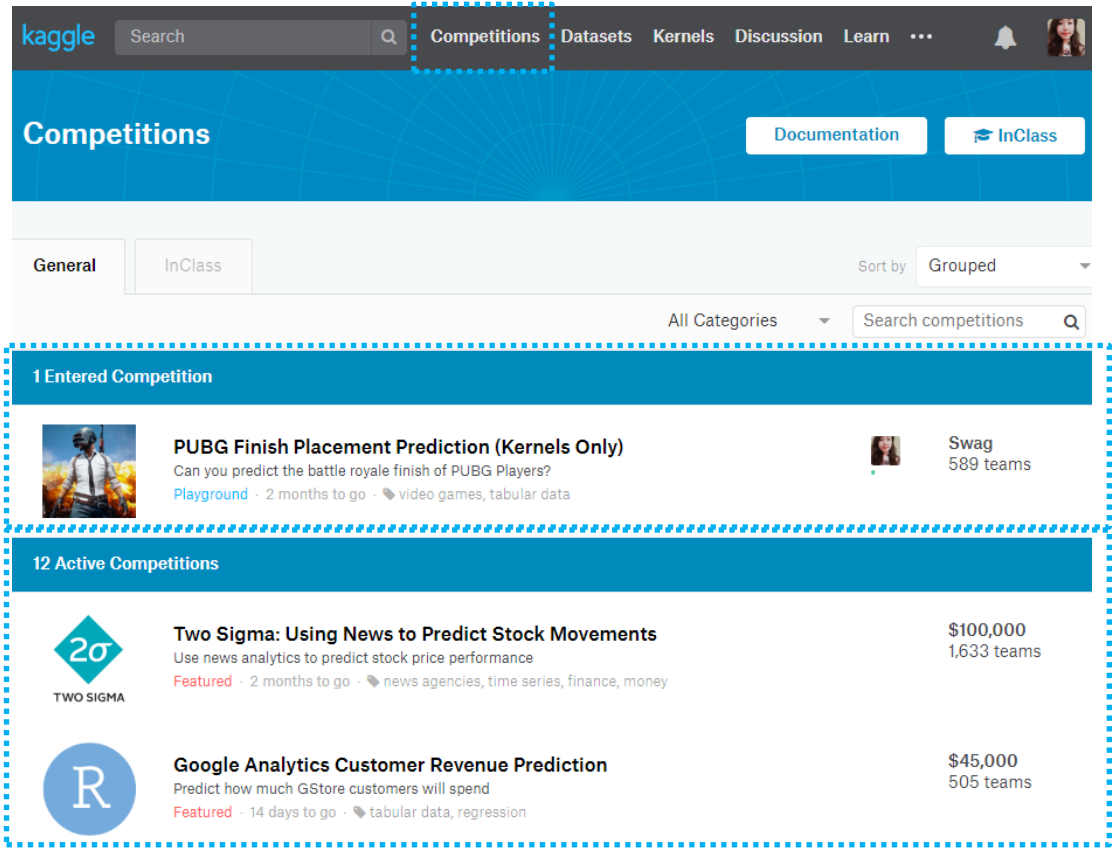
Kaggle 플랫폼 소개

Kaggle Break



1step. 가입하기

Kaggle 플랫폼 소개

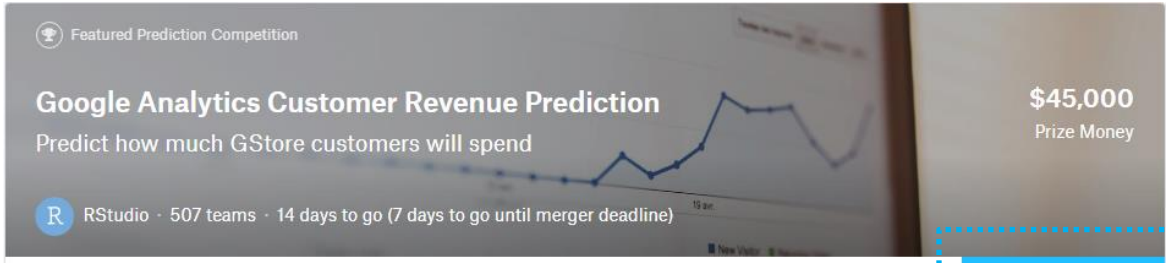


현재 참여 중인 대회

현재 진행 중인 대회

Kaggle 플랫폼 소개

Kaggle Break



Overview Data Kernels Discussion Leaderboard Rules

Join Competition

2step. 대회 선정 및 참가

Overview

대회 소개

Description

Evaluation

Prizes

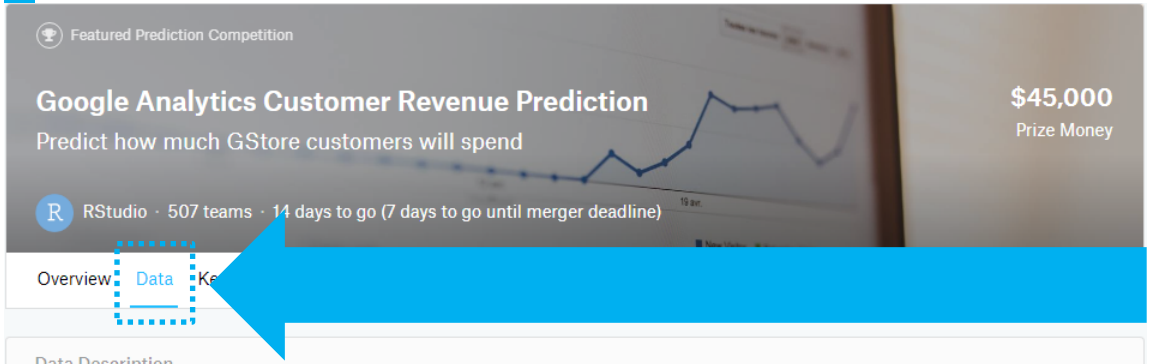
Timeline

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.

Kaggle 플랫폼 소개



Data Description

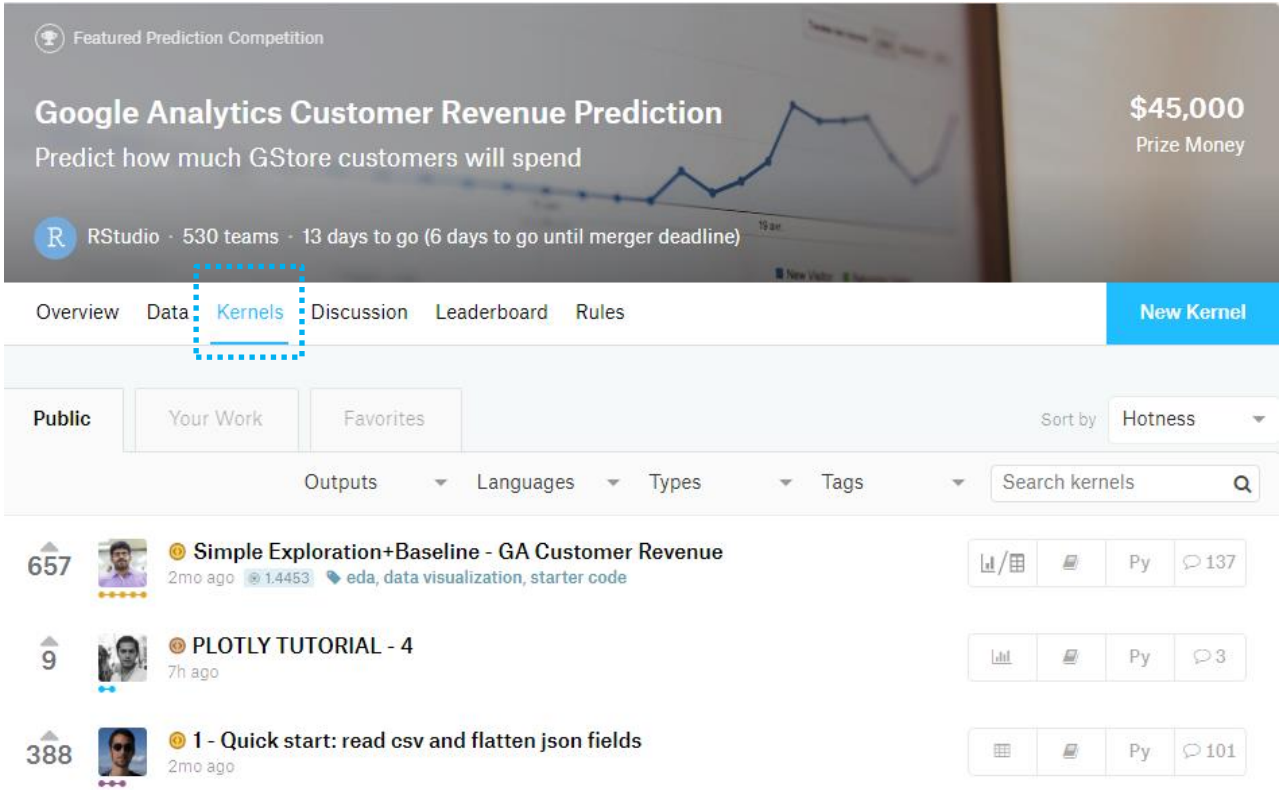
Data Fields

- ***fullVisitorId***- A unique identifier for each user of the Google Merchandise Store.
- ***channelGrouping*** - The channel via which the user came to the Store.
- ***date*** - The date on which the user visited the Store.
- ***device*** - The specifications for the device used to access the Store.
- ***geoNetwork*** - This section contains information about the geography of the user.
- ***socialEngagementType*** - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
- ***totals*** - This section contains aggregate values across the session.
- ***trafficSource*** - This section contains information about the Traffic Source from which the session originated.
- ***visitId*** - An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
- ***visitNumber*** - The session number for this user. If this is the first session, then this is set to 1.
- ***visitStartTime*** - The timestamp (expressed as POSIX time).
- ***hits*** - This row and nested fields are populated for any and all types of hits. Provides a record of all page visits.
- ***customDimensions*** - This section contains any user-level or session-level custom dimensions that are set for a session. This is a

3step. 데이터 다운로드

Kaggle 플랫폼 소개

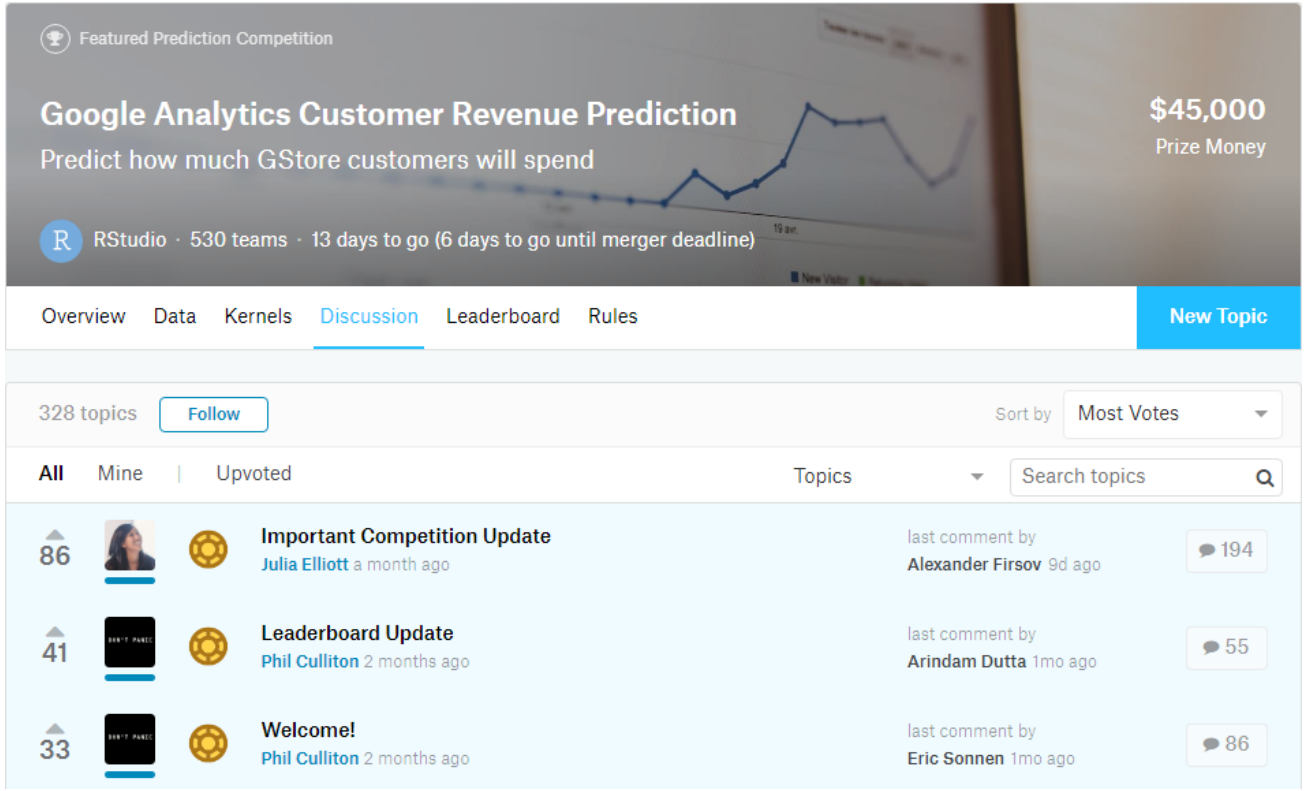
Kaggle Break



4step. 커널읽기

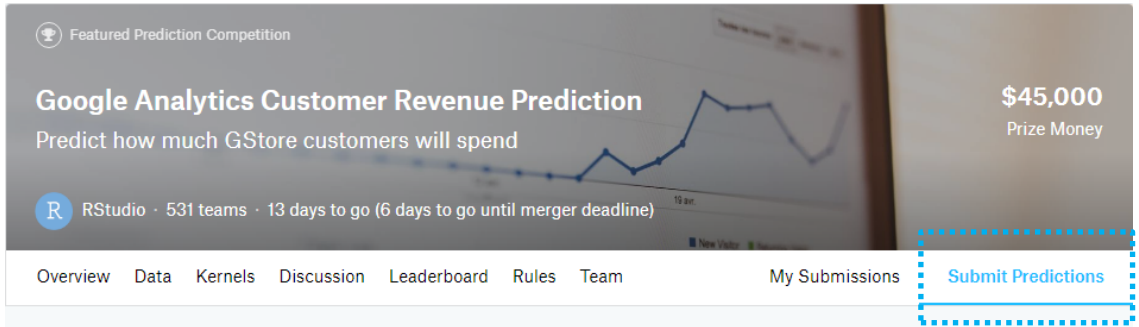
Kaggle 플랫폼 소개

Kaggle Break



5step. discussion

Kaggle 플랫폼 소개



Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Make a submission for KyuYoungLee

You have 5 submissions remaining today. This resets 18 hours from now (00: 00 UTC).

Step 1
Upload submission file

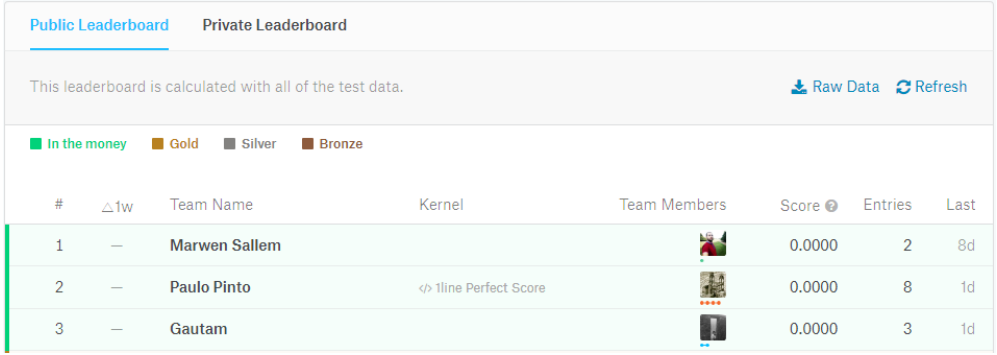
Upload Files

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 296530 prediction rows. This file should have a header row. Please see sample submission file on the data page.

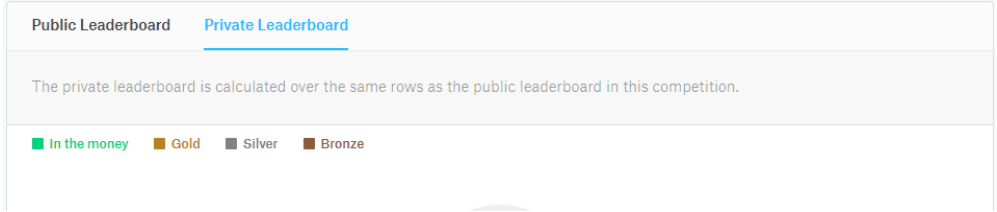
6step. Submit

리더보드 소개



Public Leaderboard
대회기간 동안 Test데이터의 일부 만 사용하여 Score 측정

Public 리더보드 vs Private 리더보드



Private Leaderboard
대회 기간이 끝나고 나머지 Test 데이터를 사용하여 Score 측정

Kaggle 대회 유형

Kaggle Break

분류	내용
Featured	외부 기업과 캐글이 연계해서 진행되는 일반적인 경진대회 (상금 O, 캐글포인트 O)
Getting Started	입문자를 위한 예제 기반 학습용 경진대회 (상금 X, 캐글포인트 X)
Research	연구목적으로 진행되는 경진대회
Playground	캐글에서 직접 주최하는 경진대회
Recruitment	채용을 목적으로 진행되는 캐글 경진대회 (상금 대신 채용, 캐글포인트 O)

Kaggle 메달

Competition Medals

Competition medals are awarded for top competition results. The number of medals awarded per competition varies depending on the size of the competition. Note that InClass, playground, and getting started competitions do not award medals.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
Bronze	Top 40%	Top 40%	Top 100	Top 10%
Silver	Top 20%	Top 20%	Top 50	Top 5%
Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

* (Top 10 + 0.2%) means that an extra gold medal will be awarded for every 500 additional teams in the competition. For example, a competition with 500 teams will award gold medals to the top 11 teams and a competition with 5000 teams will award gold medals to the top 20 teams.

Kaggle 등급

Kaggle Break

Novice

You've joined the community.

- [x] Register!

Contributor

You've completed your profile, engaged with the community, and fully explored Kaggle's platform.

- [ ] Add your bio
- [ ] Add your location
- [ ] Add your occupation
- [ ] Add your organization
- [x] SMS verify your account
- [ ] Run 1 script
- [x] Make 1 competition submission
- [ ] Make 1 comment
- [ ] Cast 1 upvote

Expert

You've completed a significant body of work on Kaggle in one or more categories of expertise. Once you've reached the expert tier for a category, you will be entered into the site wide Kaggle Ranking for that category.

Competitions	Kernels	Discussions
☐ 2 bronze medals	☐ 5 bronze medals	☐ 50 bronze medals

Kaggle 등급

Kaggle Break

Master

You've demonstrated excellence in one or more categories of expertise on Kaggle to reach this prestigious tier. Masters in the Competitions category are eligible for exclusive Master-Only competitions.

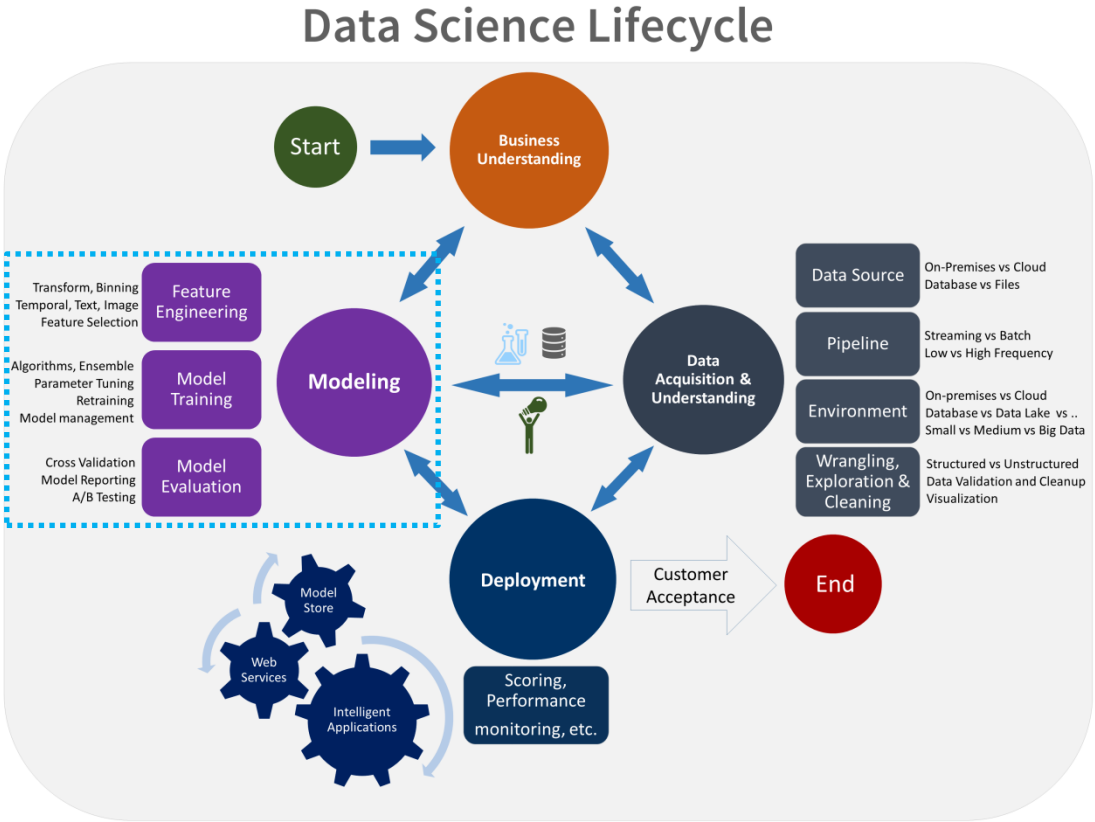
Competitions	Kernels	Discussions
☐ 1 gold medal	☐ 10 silver medals	☐ 50 silver medals
☐ 2 silver medals		☐ 200 medals in total

Grandmaster

You've consistently demonstrated outstanding performance in one or more categories of expertise on Kaggle to reach this pinnacle tier. You're the best of the best.

Competitions	Kernels	Discussions
☐ 5 gold medals	☐ 15 gold medals	☐ 50 gold medals
☐ Solo gold medal		☐ 500 medals in total

Kaggle을 하는 이유



정제된 데이터

실력자들의 공유

스펙 쌓기

Kaggle에서 배울 수 없는 것

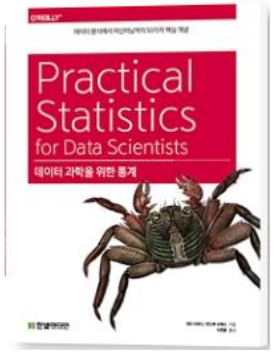
Kaggle Break

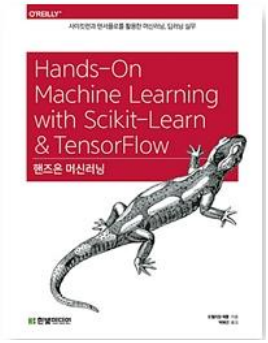
그러나 캐글에서는 배울 수 없는 것들…

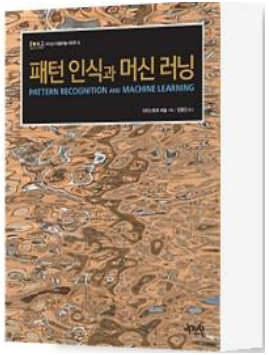
1. 문제 정의
2. 평가 지표 정의
3. 수학적/프로그래밍적인 기초
4. 현업에서 부딪히는 다른 문제들..

coursera

Data Science Academy





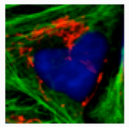


Kaggle Problems

Kaggle Break

Computer Vision

ImageNet Object Localization Challenge
Identify the objects in images
Research · 11 years to go · image data, object detection
Knowledge
30 teams



Human Protein Atlas Image Classification
Classify subcellular protein patterns in human cells
Featured · 2 months to go · image data, classification
$37,000
1,000 teams

Kaggle Problems

Kaggle Break

Audio Data

Freesound General-Purpose Audio Tagging Challenge
Can you automatically recognize sounds from a wide range of real-world environments?
Research · 4 months ago · sound technology

Knowledge
558 teams

TensorFlow Speech Recognition Challenge
Can you build an algorithm that understands simple speech commands?
Featured · 10 months ago

$25,000
1,315 teams

Kaggle Problems

Kaggle Break

Natural Language Processing

Toxic Comment Classification Challenge
Identify and classify toxic online comments
Featured · 8 months ago · arguments, text data

$35,000
4,551 teams

En

Text Normalization Challenge - English Language
Convert English text from written expressions into spoken forms
Research · a year ago · text data, languages, linguistics

$25,000
260 teams

Ru

Text Normalization Challenge - Russian Language
Convert Russian text from written expressions into spoken forms
Research · a year ago · linguistics, languages, text data

$25,000
162 teams

Kaggle Problems

Kaggle Break

Tabular data of diverse domain

제조

Mercedes-Benz Greener Manufacturing
Can you cut the time a Mercedes-Benz spends on the test bench?
Featured · a year ago · regression, automobiles, tabular data

$25,000
3,835 teams

스포츠

March Machine Learning Mania 2017
Predict the 2017 NCAA Basketball Tournament
Playground · 2 years ago · sports, future prediction, basketball

Swag
442 teams

광고

TalkingData AdTracking Fraud Detection Challenge
Can you detect fraudulent click traffic for mobile app ads?
Featured · 6 months ago

$25,000
1951/3951

Machine Learning Pipeline

Kaggle Break

Data이해 → **평가척도 이해** → **Cross-validation** → **Features Engineering** → **Model tuning** → **Ensemble**

GrandMaster Pipeline - KazAnova

Machine Learning Pipeline

Kaggle Break

Data의 이해

Zillow Prize: Zillow's Home Value Prediction (Zestimate)
Can you improve the algorithm that changed the world of real estate?
Featured · 10 months ago · housing, real estate

- Target value에 대한 이해

$$logerror = log(Zestimate) - log(SalePrice)$$

- 주어진 데이터에 대한 이해

- 결측값 수준 확인
- 데이터 컬럼들의 의미 확인

Machine Learning Pipeline

Kaggle Break

평가 척도의 이해

- 문제의 의도를 파악
- 어떤 예측값이 패널티를 크게 받고, 어떤 예측값이 덜 받는지를 이해

Santander Product Recommendation

Can you pair products with people?

Featured · 2 years ago · tabular data, banking, multiclass classification

$60,000
1,787 teams

Submissions are evaluated according to the Mean Average Precision @ 7 (MAP@7):

$$MAP@7 = \frac{1}{|U|}\sum_{u=1}^{|U|}\frac{1}{min(m,7)}\sum_{k=1}^{min(n,7)}P(k)$$

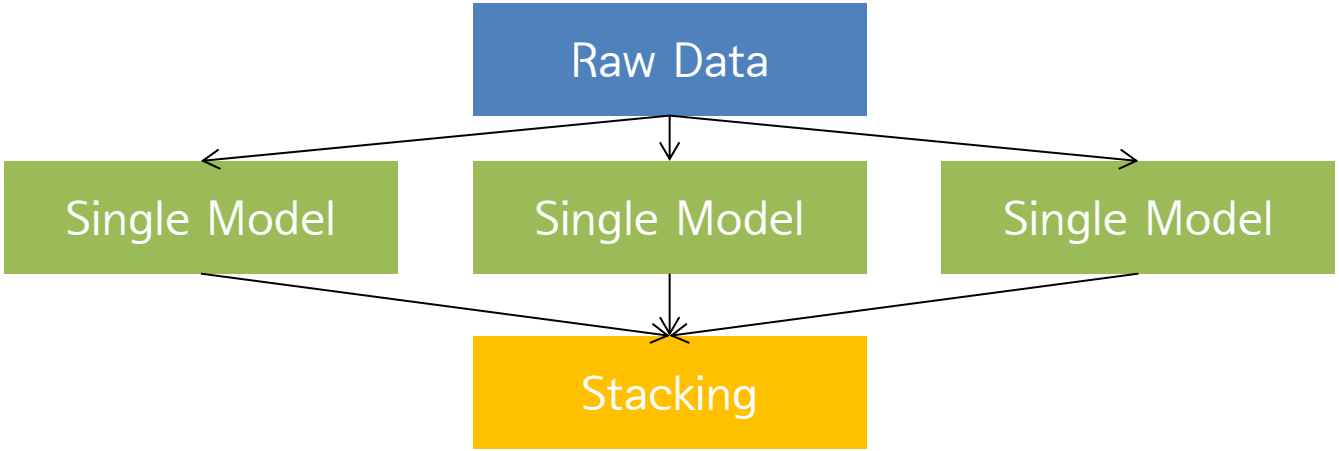
where |U| is the number of rows (users in two time points), P(k) is the precision at cutoff k, n is the number of predicted products, and m is the number of added products for the given user at that time point. If m = 0, the precision is defined to be 0.

Machine Learning Pipeline

Kaggle Break

Model Tuning - Baseline 모델

- 최소한의 성능을 보이는 기본 머신러닝 파이프라인

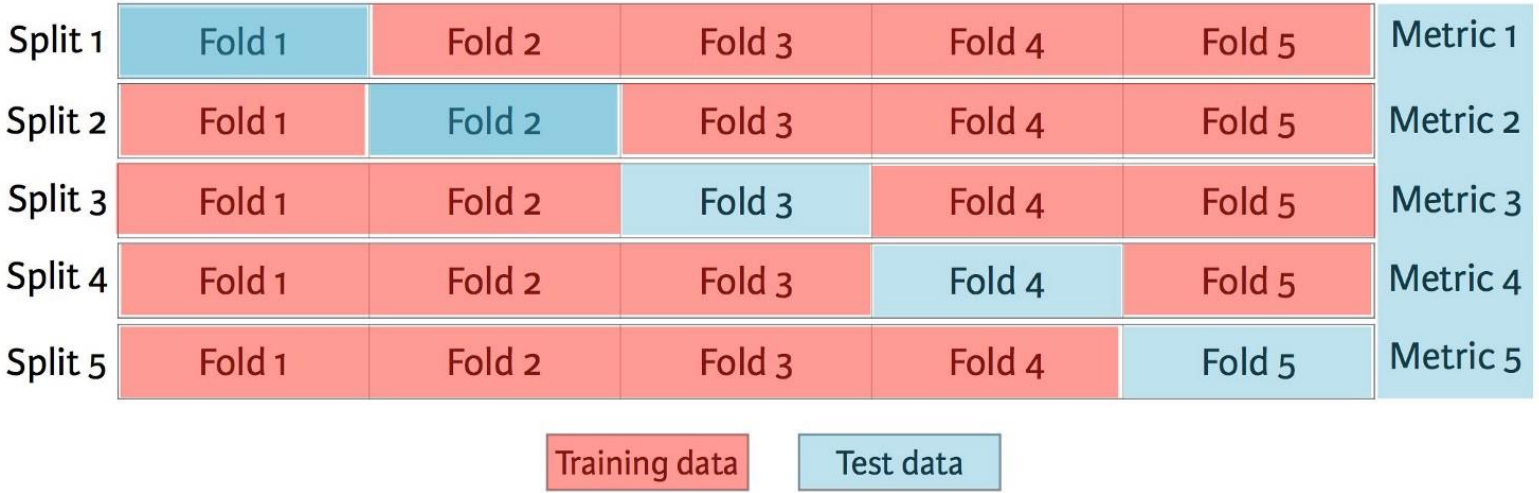


- 최소한의 Cleansing 된 Raw data, Single model들을 Stacking하는 구조에 넣어 Feature Engineering과 모델링 성능 평가를 위한 Baseline 모델 생성

Machine Learning Pipeline

Kaggle Break

Model Tuning - Cross-validation



- Stable한 validation system을 Feature Engineering 전에 구축
- Competition에서 가장 중요한 것 중에 하나는 Cross Validation Score가 Public Leaderboard 점수와 동일하게 따라가야 되는 것

Machine Learning Pipeline

Kaggle Break

변수별 Feature Engineering

Feature engineering

- The type of problem defines the feature engineering.
- **Image classification**: Scaling, shifting, rotations, CNNs. Suggestion previous data science bowls.
- **Sound classifications**: Fourier , Mfcc, specgrams, scaling . Tenso flow speech recognition
- **Text classification**: Tf-idf, svd, stemming, spell checking, stop words' removal, x-grams. StumbleUpon Evergreen Classification.
- **Time series**: Lags, weighted averaging, exponential smoothing . Walmart recruitment.
- **Categorical** : Target enc, freq, one-hot, ordinal, label encoding. Amazon employee
- **Numerical** : Scaling , binning, derivatives ,outlier removals, dimensionality reduction. Africa soil.
- **Interactions**: multiplications, divisions, group-by features . Concatenations. Homesite.
- **Recommenders**: Features on transactional history. Item popularity, frequency of purchase. Acquire Valued Shoppers.
- This process **can be automated** using selection with cross validation.

Machine Learning Pipeline

Kaggle Break

Feature Engineering - Data Transformation

Numerical Feature	Categorical Feature	Text Feature
Standard Scaler	Label Encoding	Bag-of-Words
MinMax Scaler	Frequency Encoding	TF-IDF
Winsorization	One-Hot Encoding	N-gram
Rank Transform	**Mean Encoding**	Character-n-gram
Log & Box-Cox Transform	그 외 1	K-skip-n-gram
그 외…	그 외 2	

Machine Learning Pipeline

Kaggle Break

Feature Engineering -Missing Value & Data Cleansing

Zillow Data Competition 참가

- 삭제한 컬럼 - 완전히 동일한 컬럼들은 삭제
- Missing Values 처리 및 데이터 정제
  - 컬럼 명세나 컬럼명을 통해서 빈 값을 유추
  - 일부는 상관관계와 의미를 파악해가면서 데이터를 유추
  - 일부는 log-log 모형 선형 보간(price 관련 데이터들…)
  - KNN 보간
  - 진짜 Outlier 또한 정제(Zillow 외 에어비앤비 연령 데이터)

Machine Learning Algorithms

Kaggle Break

Modeling

- The type of problem defines the feature engineering.
- **Image classification**: CNNs (Resnet, VGG, densenet…)
- **Sound classifications**: CNNs(CRNN), LSTM
- **Text classification**: GBMs, Linear, DL, Naïve bayes, KNNs, LibFM, LIBFFM
- **Time series**: Autoregressive models, ARIMA, linear, GBMs, DL, LSTMs
- **Categorical features**: GBMs, Linear models, DL , LibFM, libFFm
- **Numerical Features**: GBMs, Linear models, DL, SVMs
- **Interactions**: GBMs, Linear models, DL
- **Recommenders**: CF, DL, LibFM, LIBFFM, GBMs

Machine Learning Algorithms

Kaggle Break

Classification	Algorithms	Tool
Tree	Gradient Boosting Machine	XGBoost, LightGBM, Catboost
	Random Forests	Scikit-Learn, randomForest
Deep Learning	Neural Networks/ Deep Learning	Keras, MXNet, PyTorch, CNTK
FM-FTRL	FTRL	Vowpal Wabbit
	Factorization Machine	libFM, fastFM
	Field-aware Factorization Machine	libFFM

Machine Learning Algorithms

Kaggle Break

	Gradient Boosting Machine	Deep Learning
Base algorithm	Decision Tree	Perceptron
Use cases	Structured, categorical data	Image, speech, natural language data
Crucial step	Feature engineering	Architecture design. Finding pre-trained models
Tools	LightGBM, XGBoost, CatBoost, H2O	Keras, PyTorch, Tensorflow, CNTK, MXNet, Caffe

그 외…

Machine Learning Pipeline

Kaggle Break

Hyper parameter tuning

Grid Search

- Range 설정
- Range 내에서 전체를 탐색

Random Search

- Range 내에서 Random 탐색

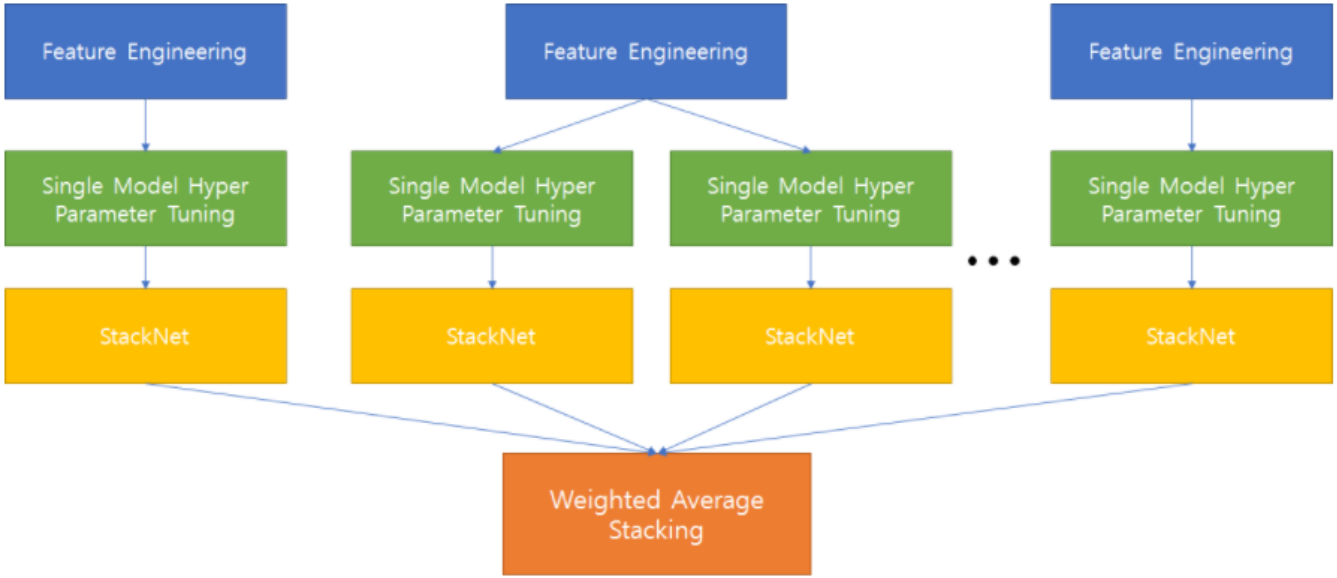
Bayesian Optimization

- Parameter를 함수로 가정하여 형태를 추정하면서 optimal search

KAGGLE PIPELINE

Kaggle Break

Ensemble

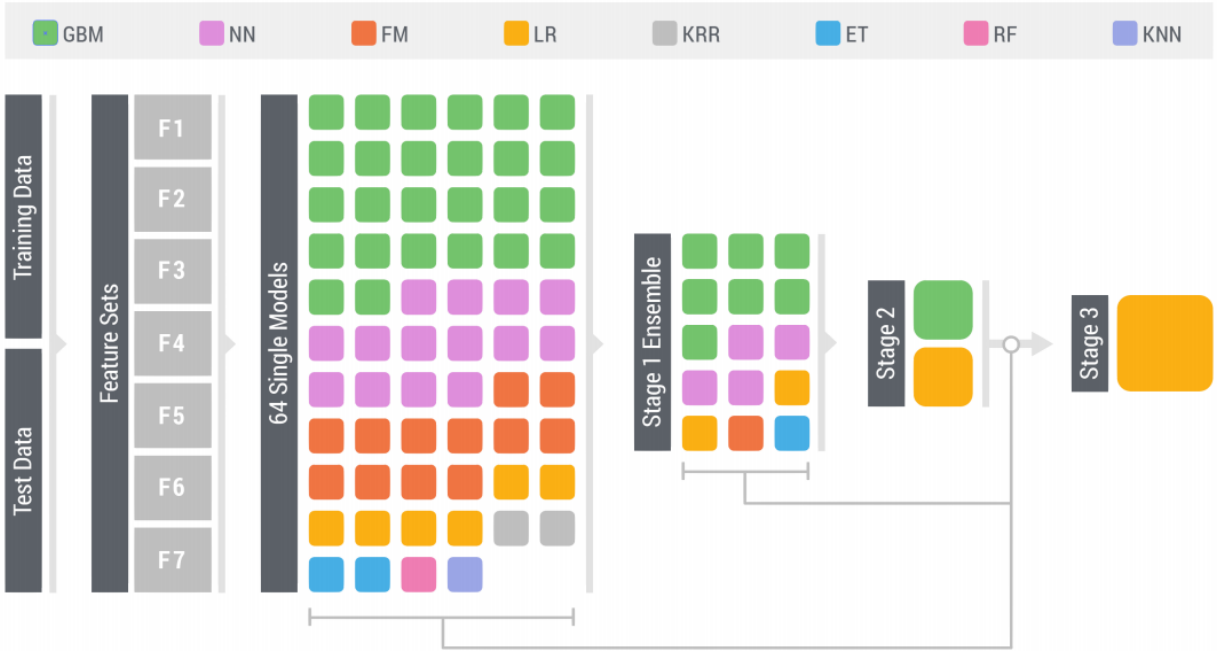


Stacknet more information

KAGGLE PIPELINE

Kaggle Break

Ensemble - KDD Cup 2015 Solution



- 한번 구축해놓은 PIPELINE은 경진대회에서 재활용 가능(Microsoft 이정윤님 자료)

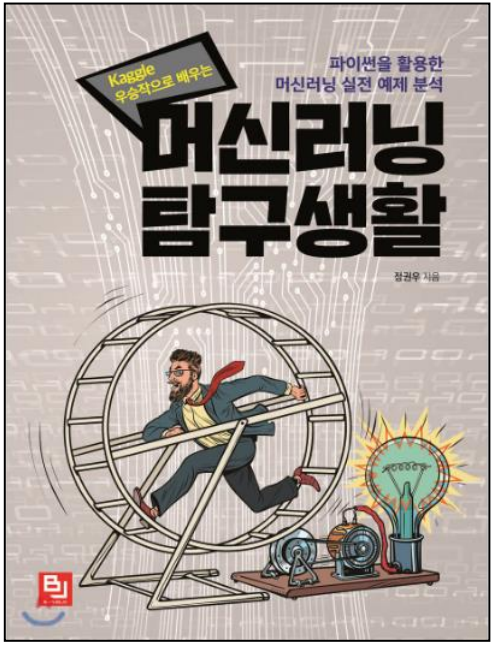
캐글을 공부하기 위해 필요한 자료들

Kaggle Break

Kaggle-knowhow

- Kaggle-Knowhow(Korean Ver)
- 한국분들을 위한 Kaggle 자료 모음입니다
- Kaggle Intro와 Kaggle Flow은 직접 작성하였으며, 주관이 들어가 있기 때문에 실제와 다를 수 있습니다!
- Pull Request 환영합니다!

- 작성자 : SeongYun Byeon
- 최근 수정일 : 18.11.14



Kaggle Break 커뮤니티

2015년에 개설된 캐글(kaggle.com) 플랫폼 대회를 참여하는 스터디

워킹캐글 Part4

캐글뽀개기 주말반(워킹캐글) 스터디 자료모음

- 파트 1 : 2017/4/8 ~ 2017/5/13
- 파트 2 : 2017/08/26 ~ 2017/10/28
- 파트 3 : 2018/03/10 ~ 2018/05/12
- 파트 4 : 2018/05/26 ~
- 운영그룹 캐글뽀개기 (https://www.facebook.com/groups/kagglebreak/)
- 스터디 구글드라이브 폴더 (https://drive.google.com/drive/folders/0B2l0iH28o85xSG83OTVfMzhhNFE)
- 장소 : 토즈타워점 강남
- 격주 토요일 오전
- 스터디 KossLab(공개SW 개발자센터)에서 장소를 지원하고 있습니다.
- 파트 4 주제

요일	주제	발표자	발표자료
2018.05.26(토)	스터디 소개	이상열님	발표자료
2018.06.09(토)	Azure pass 사용법 안내	이상열님	

함수산책 (이전 이름 캐글즐기기)

함수산책 (캐글뽀개기) 파트5 평일반 스터디 자료모음

- https://www.facebook.com/groups/kagglebreak/
- https://drive.google.com/drive/folders/0B2l0iH28o85xcHJRNWNUc1FvbEk
- 장소 : 토즈 강남점
- 파트5는 격주 수요일
- 스터디 KossLab(공개SW 개발자센터)에서 장소를 지원하고 있습니다.

교재

- 수리통계, 주교재 최신 수리통계학 출판사 경문사 저자 안승철,이재원,최원
- 선형대수학, 주교재 프로그래머를 위한 선형대수
- 네트워크 분석, 주교재 Python for Graph and Network Analysis

Kaggle Break 행사

Pycon 튜토리얼 : 미운 우리 캐글

https://www.slideshare.net/yeonminkim/pycon-korea-2018-kaggle-tutorialkaggle-break

Databreak 2018 : Hello, kaggler!

http://kagglebreak.com/databreak2018/

Kaggle Break 참가

캐글뽀개기 커뮤니티
http://kagglebreak.com/

Github
https://github.com/KaggleBreak

캐글뽀개기 페이스북
https://www.facebook.com/groups/kagglebreak/

감사합니다

질문은 mineatte@gmail.com 으로
커뮤니티 관련 질문은 admin@kagglebreak.com 으로

Q&A