


Intro Kaggle & Competition

개요

- Data 이해 : 문제지 보기
 - Model 생성 : 풀 방법 정하기 & 풀기
 - Submission : 정답 제출하기
 - Evaluation : 체점하기
 - LeaderBoard : 등수로 줄세우자!!
-
- 요약 : DMSEL

1) Data


 Dataset

COVID-19 Open Research Dataset Challenge (CORD-19)

An AI challenge with AI2, CZI, MSR, Georgetown, NIH & The White House

 Allen Institute For AI and 7 collaborators • updated 3 days ago (Version 3)

[Data](#) [Tasks \(10\)](#) [Kernels \(17\)](#) [Discussion \(7\)](#) [Activity](#) [Metadata](#) [Do](#)

 Usability 9.4

 License Other (specified in description)

 Tags

Description

Dataset Description

In response to the COVID-19 pandemic, the White House and a coalition of leading research Open Research Dataset (CORD-19). CORD-19 is a resource of over 29,000 scholarly article about COVID-19, SARS-CoV-2, and related coronaviruses. This freely available dataset is p community to apply recent advances in natural language processing and other AI techniqu of the ongoing fight against this infectious disease. There is a growing urgency for these a acceleration in new coronavirus literature, making it difficult for the medical research comm

Data (2 GB)

Data Sources

2020-03-13

all_sources_metadata_2020-03-1... 14 columns

json_schema.txt

all_sources_metadata_2020-03-13.readme

biorexiv_medrxiv

biorexiv_medrxiv

0015023cc06b5362d332b3baf348d11567ca...

004f0f8bb66cf446678dc13cf2701feec4f36d...

004f0f8bb66cf446678dc13cf2701feec4f36d...

About this file

CORD-19 dataset (2020-03-13)

2020-03-13




all_sources_metadata_2020-03...

Size 46.93 MB



json_schema.txt

Size 2.84 KB



all_sources_metadata_2020-03...

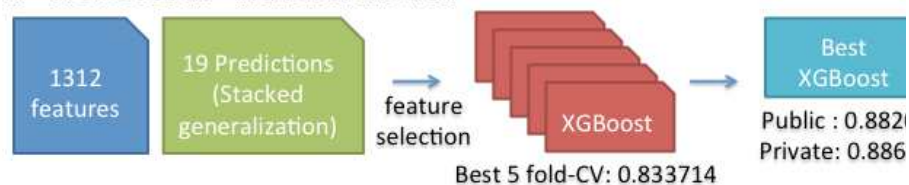
Size 1000 B

2) Model

2nd Place Solution(short summary)

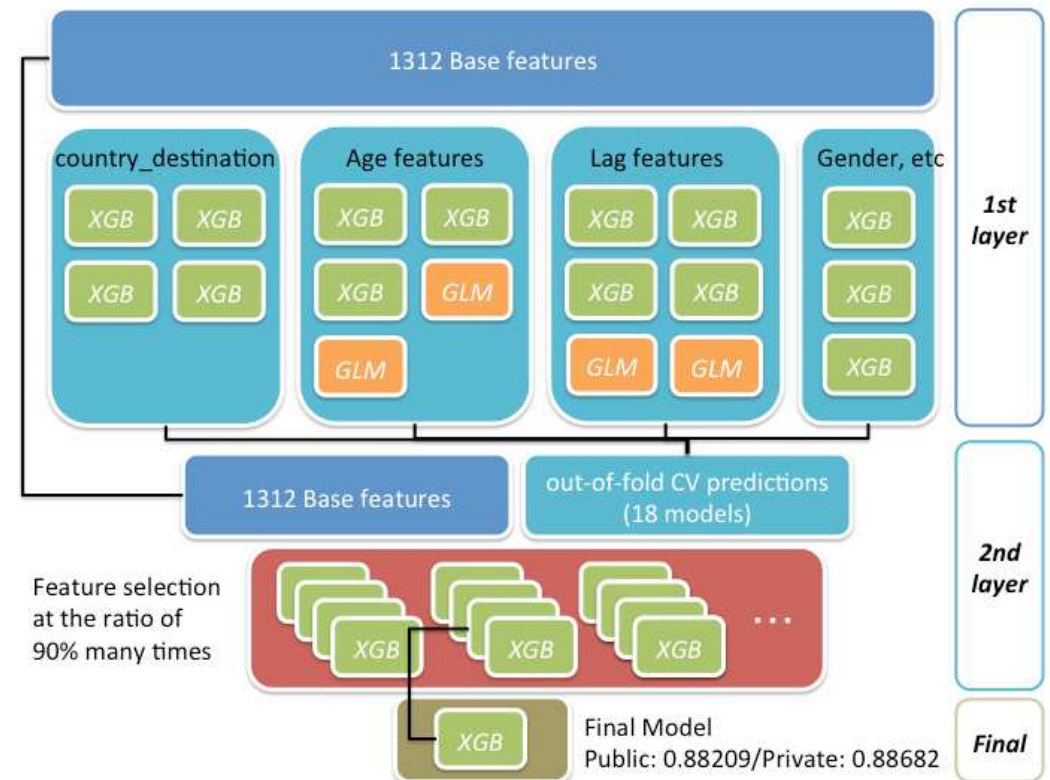
■ My final model was based in a 2-layer model as the following.

- 1312 features from original dataset.
 - ✓ Numerical features(cleared age, extracted date), Categorical features(OHE), joined age_gender_bkts and countries, summarized sessions.
- 19 models as meta features for the 2nd level.
 - ✓ Target: country_destination, age, cleared age, lag features (date_first_booking - timestamp_first_active and date_first_booking - first_affiliate_tracked), etc (Further details are provided my source code).
- Random feature selection(sampling rate = 90%).
- XGBoost as 2nd level model.

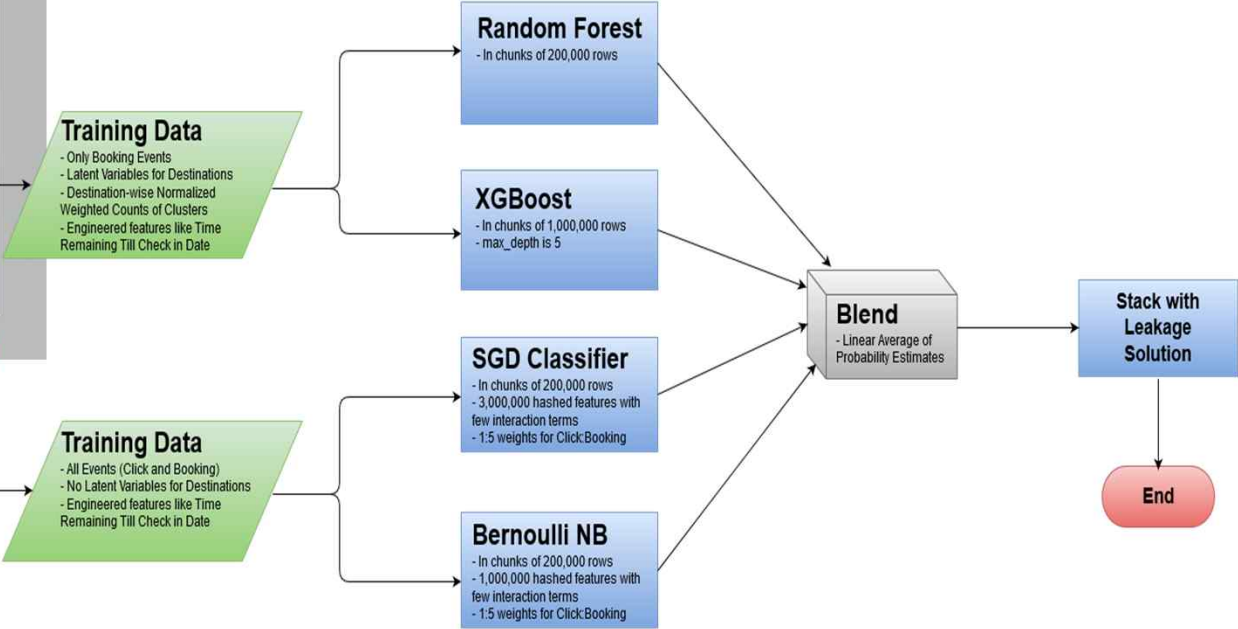
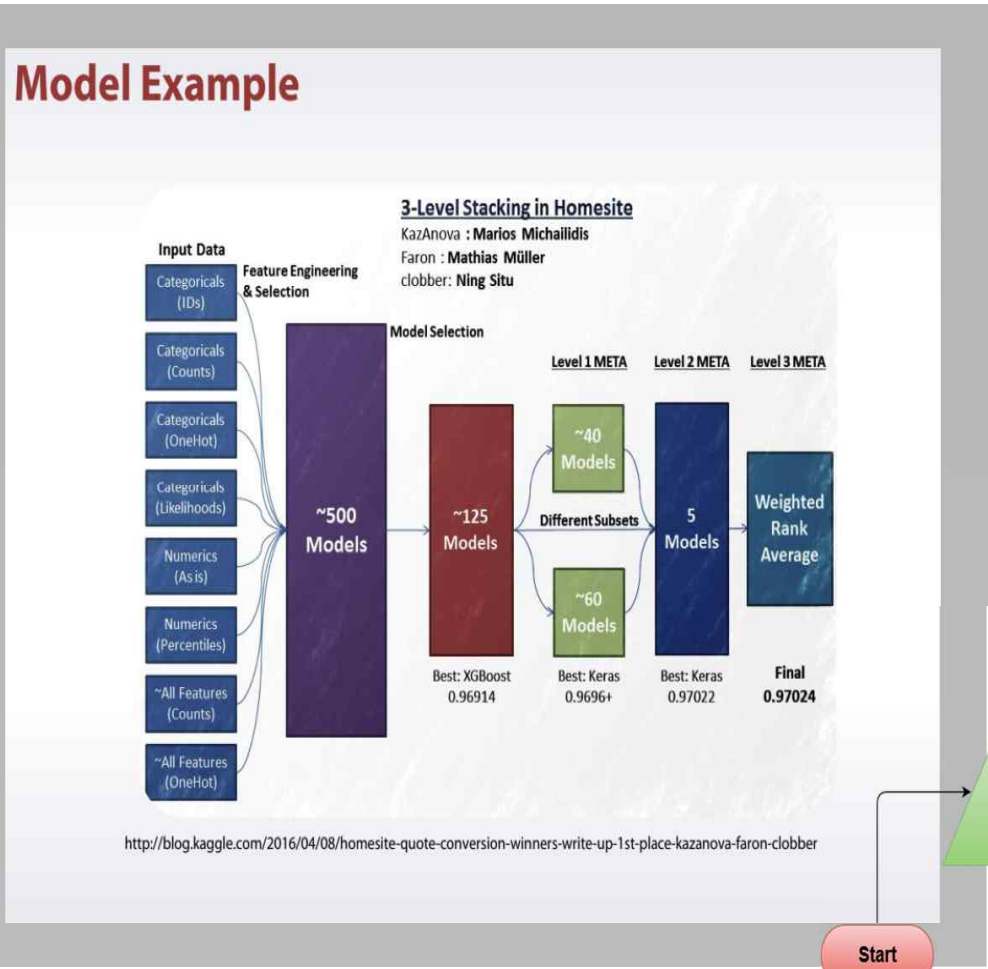


My Approach & Program Details.

Please see [Airbnb New User Bookings, Winner's Interview: 2nd place, Keiichi Kuroyanagi \(@Keiku\)](#) | no free hun



Model Example



3) Submission

• 정답지 제출

Overview

Data

Notebooks

Discussion

Leaderboard

Rules

Team

Description

Evaluation

Timeline

For example, if for a particular user the destination is FR, then

[FR] gives a $NDCG = \frac{2^1 - 1}{\log_2(1+1)} = 1.0$

[US, FR] gives a $DCG = \frac{2^0 - 1}{\log_2(1+1)} + \frac{2^1 - 1}{\log_2(2+1)} = \frac{1}{1.58496}$

Submission File

For every user in the dataset, submission files should contain destination country predictions must be ordered such that the first prediction goes first.

The file should contain a header and have the following form:

```
id, country
000am9932b, NDF
000am9932b, US
000am9932b, IT
01wi37r0hw, FR
etc.
```

Overview

Data

Notebooks

Discussion

Leaderboard

Rules

Team

My Submissions

Late Submission

You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.

Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.

You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

>_

kaggle competitions submit -c airbnb-recruiting-new-user-bookings -f submission.csv -m "Message"

?

0 submissions for MinkyuKwon

Sort by

Most recent

All

Successful

Selected

Submission and Description

Private Score

Public Score

Use for Final Score

No submissions to show

4) Evaluation

- Examples)
 - Accuracy
 - Logistic Loss
 - AUC
 - RMSE / MAE
 - Cross Entropy
 - Etc...

[Overview](#) [Data](#) [Notebooks](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Late Submission](#)

Description	The evaluation metric for this competition is NDCG (Normalized discounted cumulative gain) @k where k=5. NDCG is calculated as:
Evaluation	$DCG_k = \sum_{i=1}^k \frac{2^{rel_i} - 1}{\log_2(i + 1)},$ $nDCG_k = \frac{DCG_k}{IDCG_k},$
Timeline	where rel_i is the relevance of the result at position i . $IDCG_k$ is the maximum possible (ideal) DCG for a given set of queries. All NDCG calculations are relative values on the interval 0.0 to 1.0. For each new user, you are to make a maximum of 5 predictions on the country of the first booking. The ground truth country is marked with relevance = 1, while the rest have relevance = 0. For example, if for a particular user the destination is FR, then the predictions become: [FR] gives a $NDCG = \frac{2^1 - 1}{\log_2(1+1)} = 1.0$ [US, FR] gives a $DCG = \frac{2^0 - 1}{\log_2(1+1)} + \frac{2^1 - 1}{\log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

5) LeaderBoard

[Overview](#) [Data](#) [Notebooks](#) [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

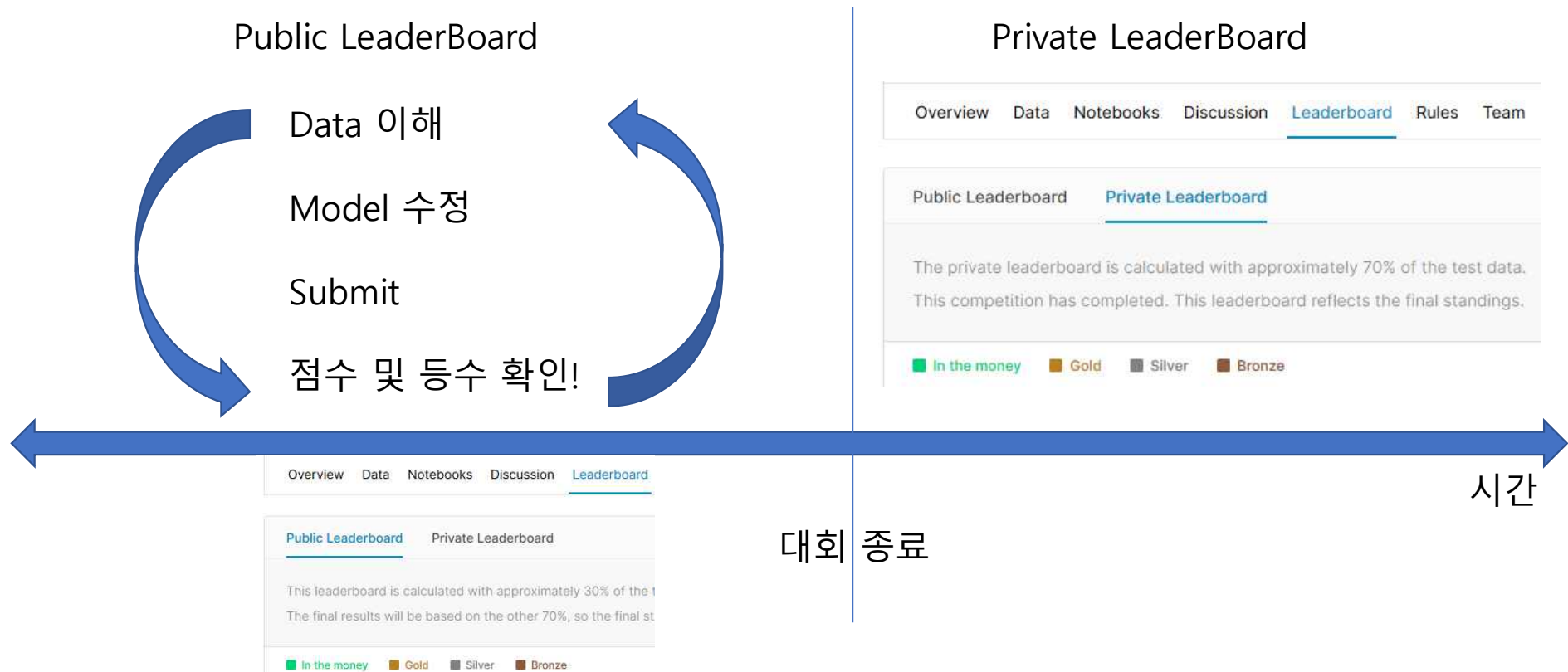
Silver

Bronze

#	△pub	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	▲6	Anupam Pandey		 ★★★★	0.88697	16	4y
2	—	Keiku		 ★★★★	0.88682	16	4y
3	▲1	Sandro		 ★★★★	0.88670	49	4y

- Public Test : 대회 기간 중에 업로드 하는 곳
- Private Test : 대회 종료 후에도 사람들이 올리고 평가를 해볼 수 있는 공간!!!
- → 크게 2가지 공간으로 구분이 되어 있음!!!!
- → 따라서 대회에 출전을 한다면 꼭 "Public"으로 제출을 해야함!!!!!!!!

6) Total Process



기타 플랫폼

- Kaggle
- DrivenData
- CodaLab
- DataScienceChallenge.net

한국

- 데이콘

고려사항	실제 업무	경진대회
문제의 정의 및 정형화	문제부터 정의하고 정형화가 문제임;;	주어져 있음! → 고민이 필요 없음
최종 평가 방식	고를 수 있으며, 변경 가능함.	주어져 있음! → 고민이 필요 없음! → 유사 대안 제시 없음!
돌발 상황	늘 언제든지 돌발상황 존재 → 고객 변심, 계약 변경 등	고민 대상이 아님!!
데이터 수집	제일 처음 단계부터 어떻게 할지가 큰 문제임!! → 살 것인지, 수집할 것인지, 가능한 한 지 등	주어져 있음!! → 고민이 필요 없음 (단, 상황에 따라서 외부 데이터 등을 활 용할 수는 있음!!)
성능에 대한 기준	사전에 계획은 하지만 상황에 따라서 변경도 됨.	정해져 있어, 그 방식으로 제일 잘하면 됨!

* 현실에 대한 문제는 좀 더 복잡하고, 정의하는 것 부터가 쉽지 않음!!

그럼에도 불구하고 왜?

- 쉽게 오해하는 부분은 이러한 경진대회는 "알고리즘"에 있는 것이 아니다.
 - 왜? 누구나 다 알고있는 알고리즘이기에...
 - 다만, 그것을 어떻게 어디에 활용할지는 다 개별적임
 - 그리고 특징에 대해서도 어떻게 하는지에 따라서 상당히 차이가 발생 함.
- 경우에 따라서는 ML이 아닌 경우도 있음!!!
- 주어진 알고리즘들에 대해서 잘 사용하고, 이들을 결합하여 새로운 알고리즘을 제시하고, 데이터에 대한 특징에 대한 변경을 하는 등 "데이터 이해 + 창의적인 방식 " 이 중요한 부분임!!!

