本当に簡単なkaggleの始め方

@yukinagae

Agenda

1. kaggleとは？
2. データ分析のトレンドの変化
3. kaggleの仕組み
4. なぜkaggleをやるの？
5. やってみた(｀・ω・´)

1. kaggleとは？

世界最大の機械学習・データ分析のコンペを主催するプラットフォーム

kaggle

つまり

データサイエンティストの
世界最強を決める大会

kaggleの規模

- ユーザ数: 50万以上
- 国: 190カ国以上

らしい(`・ω・´)

2. データ分析トレンドの変化

理論（theory）

↓

実践（practice）

昔

- 理解するのが大事（theory）

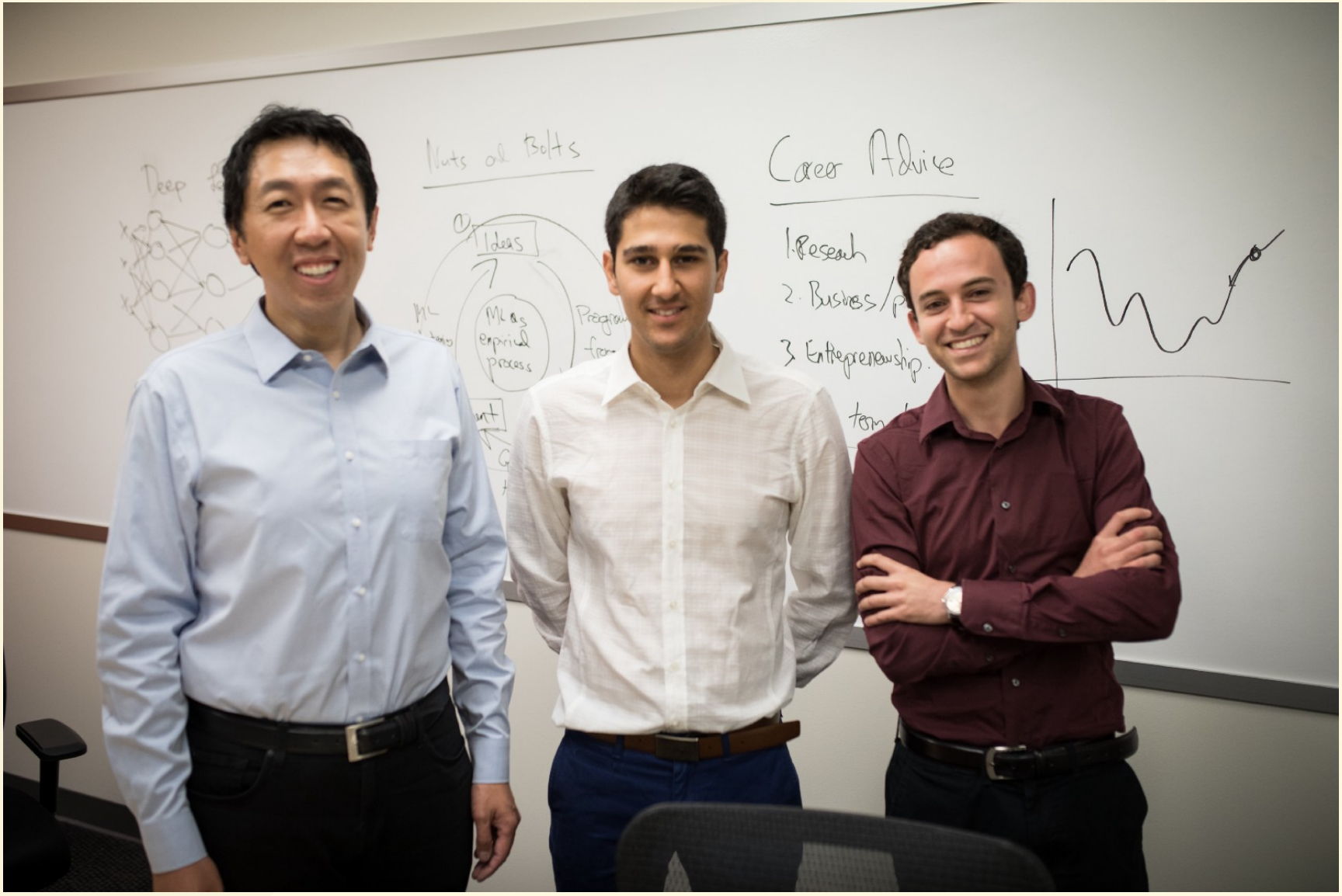
↓

今

- 実践・役立つのが大事（practice）

実践を重視している具体例

- kaggle
  - コンペ
  - ノウハウの共有（コード・ディスカッション）
  - データの共有・公開
- fast.ai
  - deep learning for coders（開発者の深層学習）
  - 理論より実践、SOA（state of art: 最先端）



From: [deeplearning.ai: Announcing new Deep Learning courses on Coursera]

AI社会による生活の向上

by Andrew Ng (AIや機械学習の有名な教授・教師)

> *I hope we can build an AI-powered society that gives everyone affordable healthcare, provides every child a personalized education, makes inexpensive self-driving cars available to all, and provides meaningful work for every man and woman. An AI-powered society that improves every person's life.*

3. kaggleの仕組み

大まかな流れ

1. 主催者（企業など）がコンペを主催する
   a. データを準備
   b. 問題を定義する
2. 参加者は様々な手法を使ってベストなモデルを構築し、予測を提出する
   - スコアやランキングが分かる
3. 主催者は、精度が高い予測に賞金を払う

4. kaggleをなんでやるの？

参加者のメリット

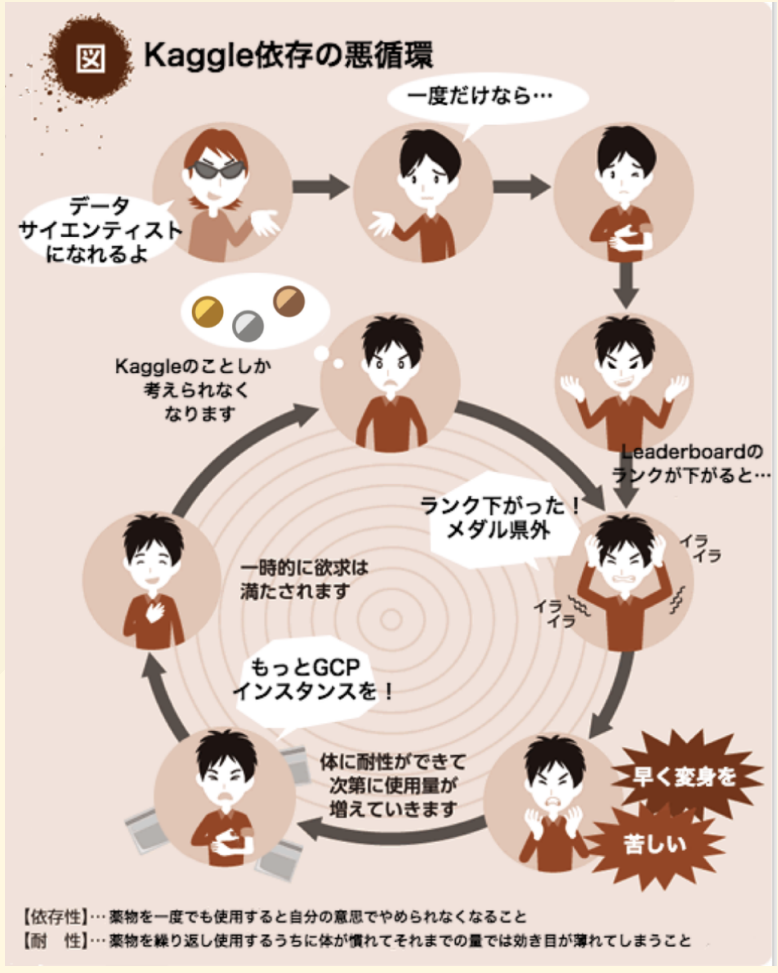
- 様々なデータに触れられる（企業が実データを提供してくれる。レアい）
- 他の参加者から学べる
- 入賞すれば賞金 + 良い仕事をGET！
- 楽しい
- ~~ギャンブル感覚~~

主催者側のメリット

- 世界中のデータサイエンティストが問題解決の手法を試行錯誤してくれる
- ブランディング・PR
- データサイエンティストの採用

が、しかし

kaggle依存の副作用
(kaggle is drug)

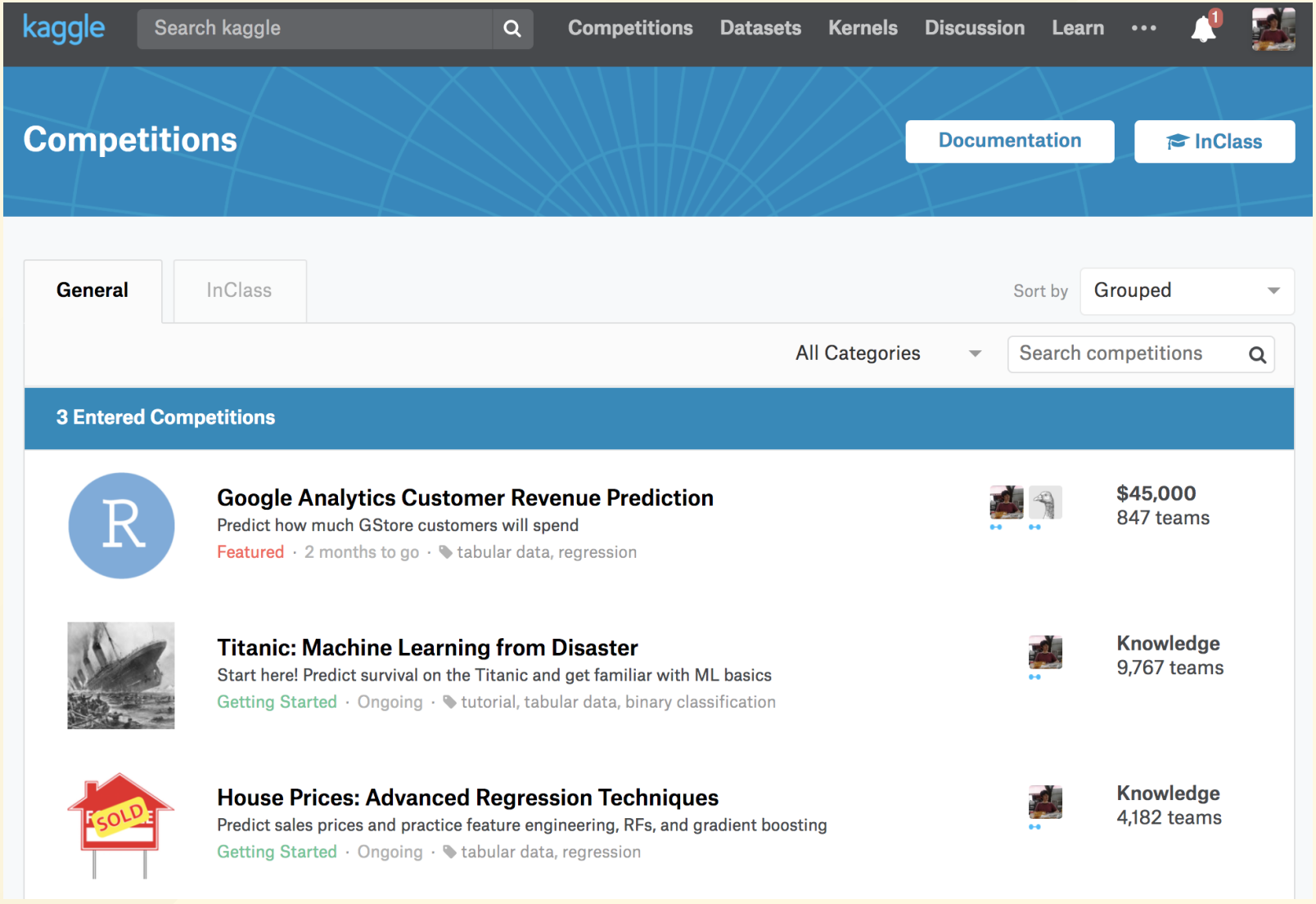


コンペが始まると仕事しなくなる人たち

ということで

5. 早速kaggleをやってみた
(｀・ω・´)

1. コンペを選ぶ


kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Learn
Competitions
Documentation
InClass
General
InClass
Sort by
Grouped
All Categories
Search competitions
3 Entered Competitions
Google Analytics Customer Revenue Prediction
Predict how much GStore customers will spend
Featured · 2 months to go · tabular data, regression
$45,000
847 teams
Titanic: Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started · Ongoing · tutorial, tabular data, binary classification
Knowledge
9,767 teams
House Prices: Advanced Regression Techniques
Predict sales prices and practice feature engineering, RFs, and gradient boosting
Getting Started · Ongoing · tabular data, regression
Knowledge
4,182 teams

最近始まったばかりのコンペ(｀・ω・´)

Featured Prediction Competition

Google Analytics Customer Revenue Prediction

Predict how much GStore customers will spend

$45,000
Prize Money

RStudio · 847 teams · 2 months to go (2 months to go until merger deadline)

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

Overview

- 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.

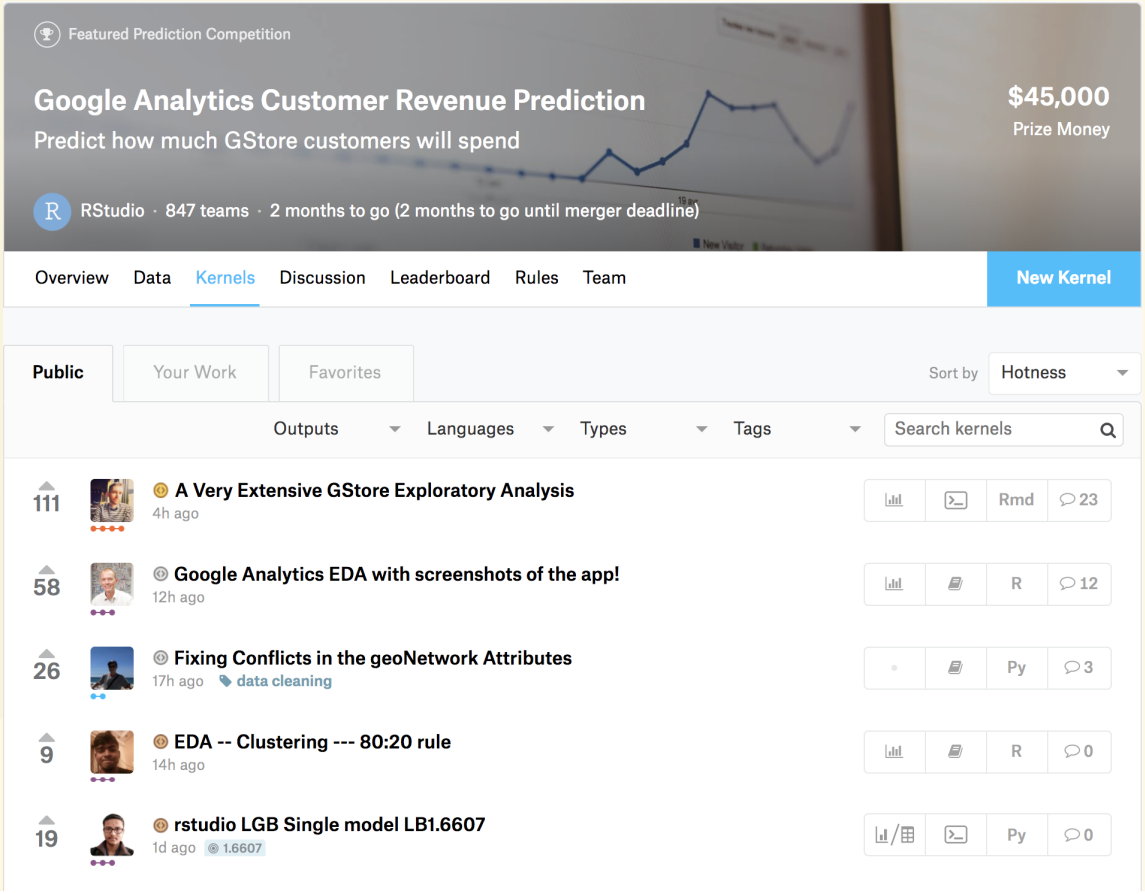
In this competition, you're challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies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

2. コンペの内容を読む

1. 概要: 大まかに把握
2. 評価指標: これが一番大事(`・ω・´)
3. 賞金: できればほしいよね
4. 期限: 時間厳守
5. データ: だいたいCSVファイル（BigQueryも）

3. 他の参加者から学ぶ

1. コード（kernel）
2. ディスカッション（discussion）



いろんな人がコードを載せてくれてるので助かる

In [2]:
```
def load_df(csv_path='../input/train.csv', nrows=None):
    JSON_COLUMNS = ['device', 'geoNetwork', 'totals', 'trafficSource']

    df = pd.read_csv(csv_path,
                     converters={column: json.loads for column in JSON_COLUMNS},
                     dtype={'fullVisitorId': 'str'}, # Important!!
                     nrows=nrows)

    for column in JSON_COLUMNS:
        column_as_df = json_normalize(df[column])
        column_as_df.columns = [f"{column}.{subcolumn}" for subcolumn in column_as_df.columns]
        df = df.drop(column, axis=1).merge(column_as_df, right_index=True, left_index=True)
    print(f"Loaded {os.path.basename(csv_path)}. Shape: {df.shape}")
    return df
```

In [3]:
```
%%time
train_df = load_df()
test_df = load_df("../input/test.csv")
```

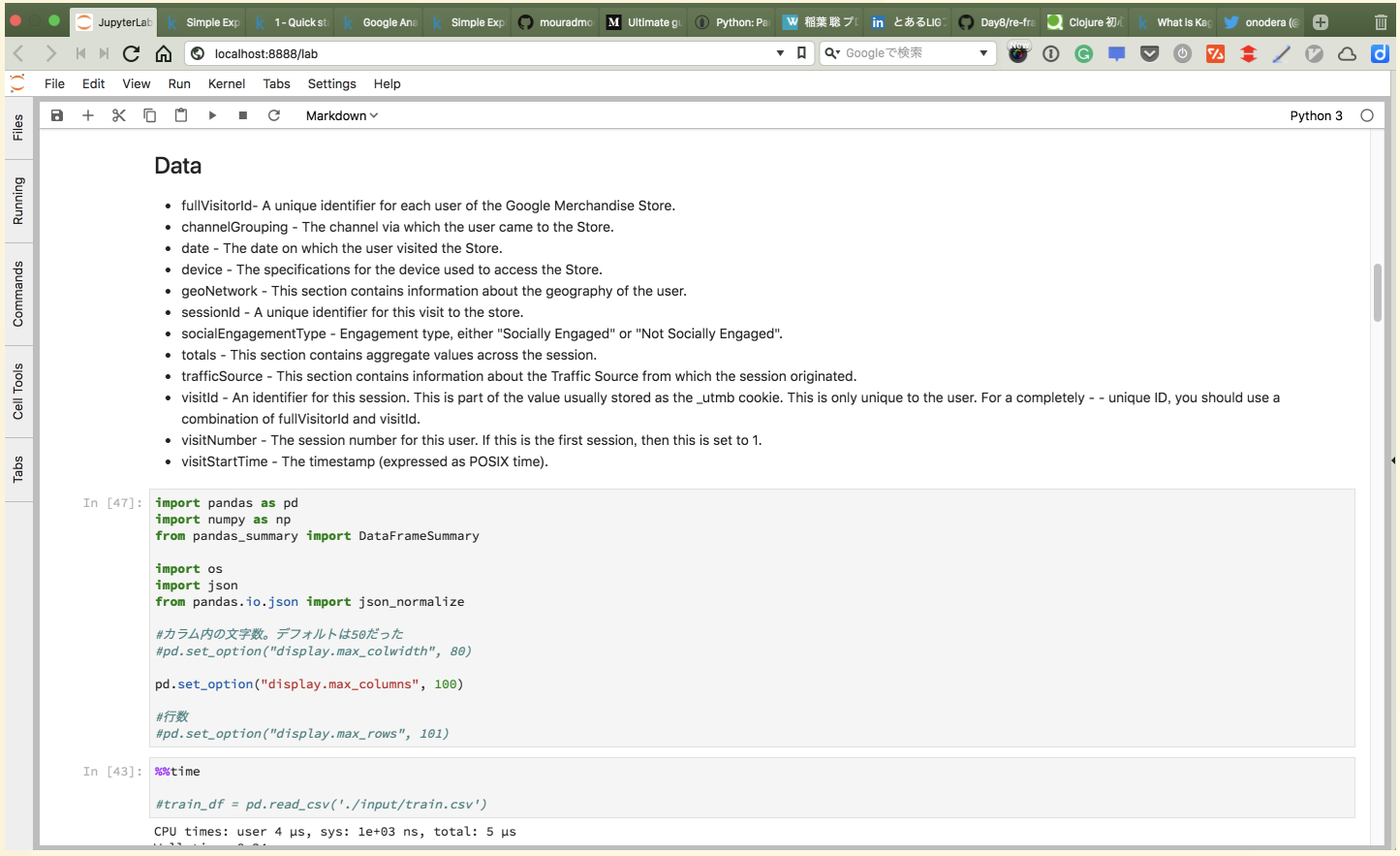
```
Loaded train.csv. Shape: (903653, 55)
Loaded test.csv. Shape: (804684, 53)
CPU times: user 5min 6s, sys: 12.9 s, total: 5min 19s
Wall time: 5min 19s
```

In [4]:
```
train_df.head()
```

see: [Simple Exploration+Baseline - GA Customer Revenue | Kaggle]

4. 他の参加者の方法を真似てみる

コードをパクってローカルPCで実行するだけの簡単なお仕事(｀・ω・´)



5. 助け合う <= New!

ちょうどライブラリのバージョンで上手く動作しなかったので、上手くいった方法を教え合う(｀・ω・´)

Eiji Sakai • Posted on Latest Version • a day ago • Options • Reply

1

Thank you for your suggestion. It could be even simpler like this:

```
column_as_df = df[column].apply(lambda x: pd.Series(x))
```

YukiNagae • Posted on Latest Version • 15 hours ago • Options • Edit • Reply

0

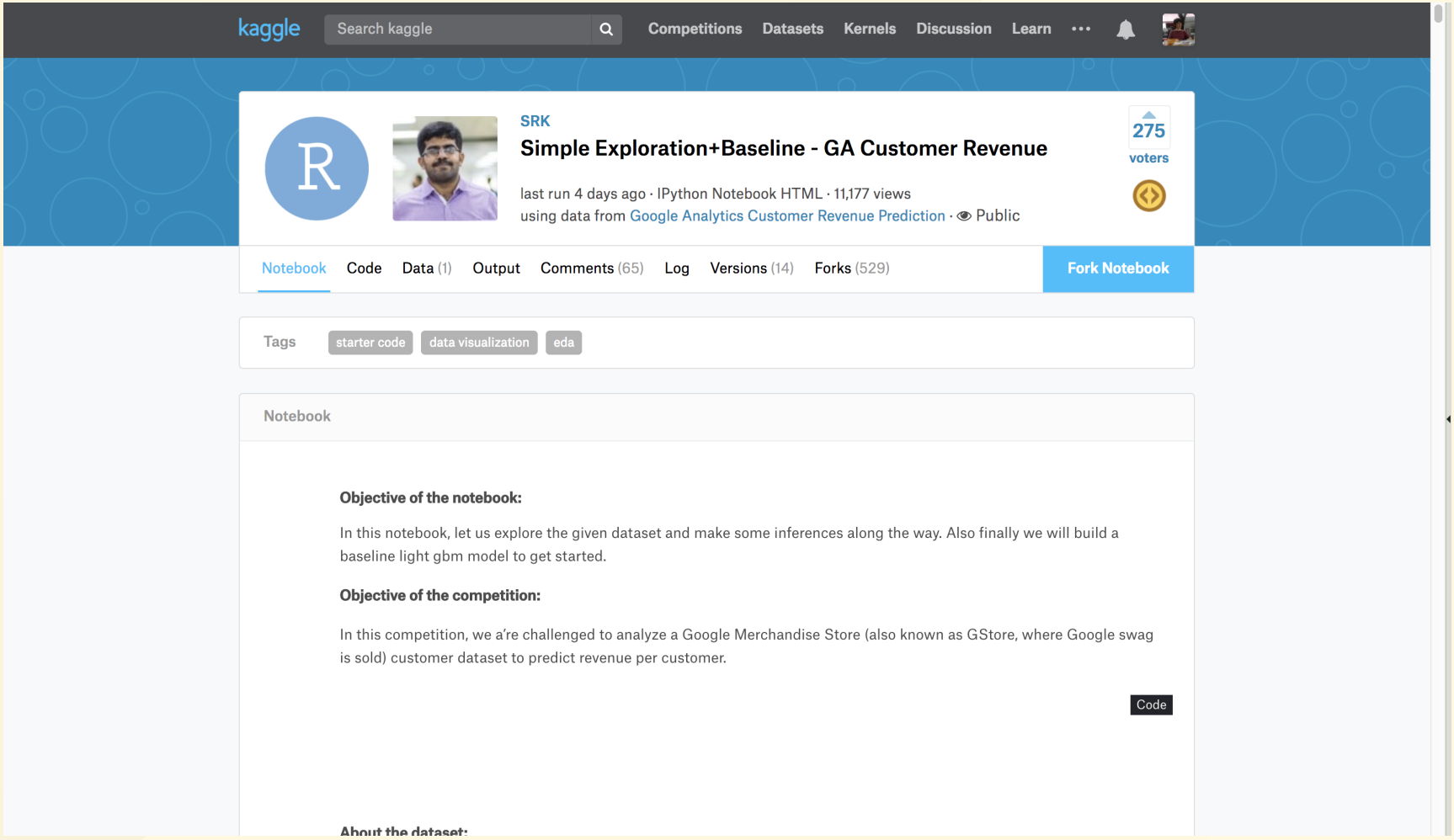
It looks like upgrading `pandas` is also a solution.

I've solved the problem by upgrading `pandas` from 0.20.1 to 0.23.4.

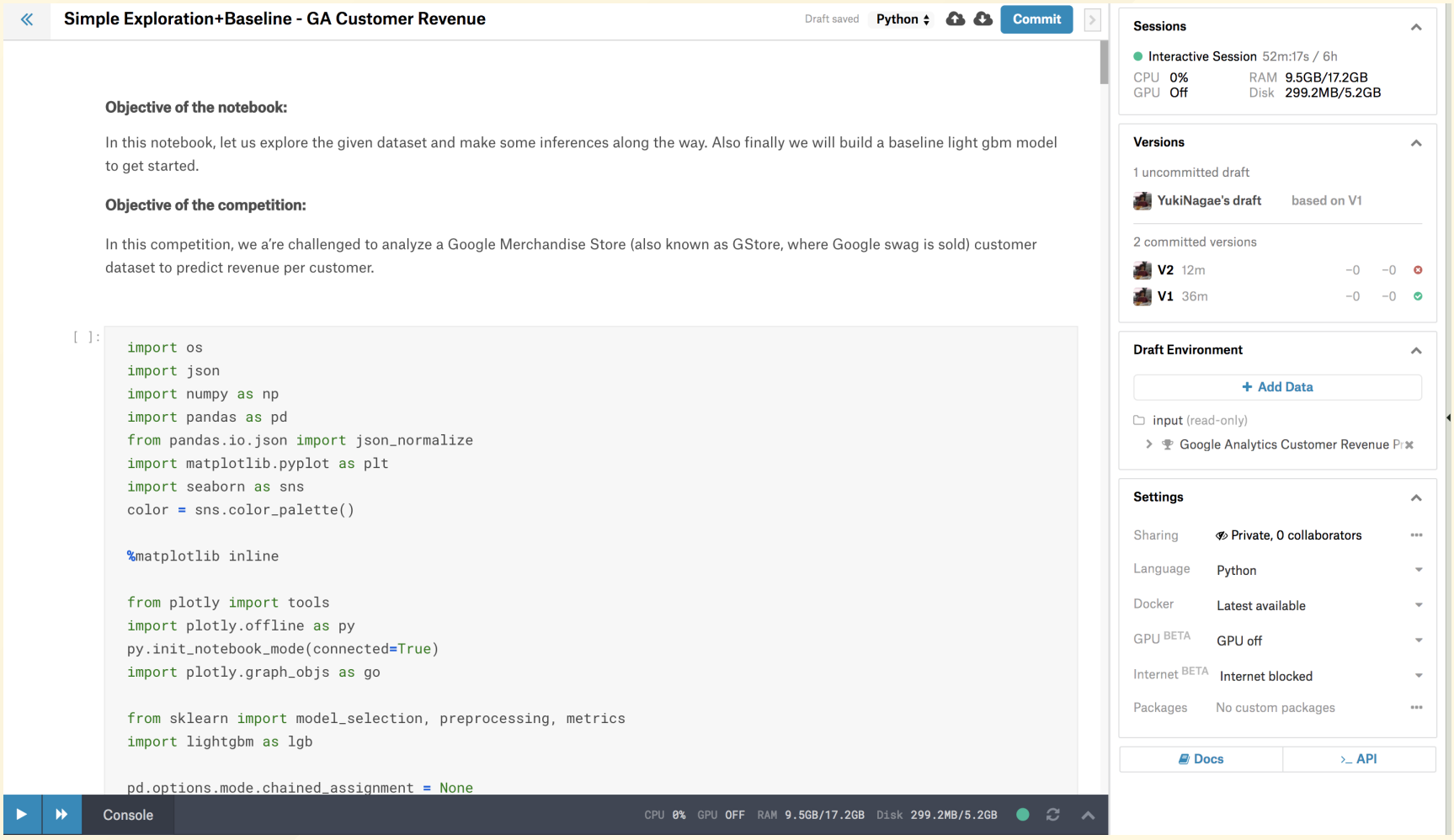
You can try `pip install pandas --upgrade` to upgrade the version of pandas.

As mentioned in the below comment, `0.23.x` of pandas is necessary.
https://www.kaggle.com/julian3833/1-quick-start-read-csv-and-flatten-json-fields/notebook#390062

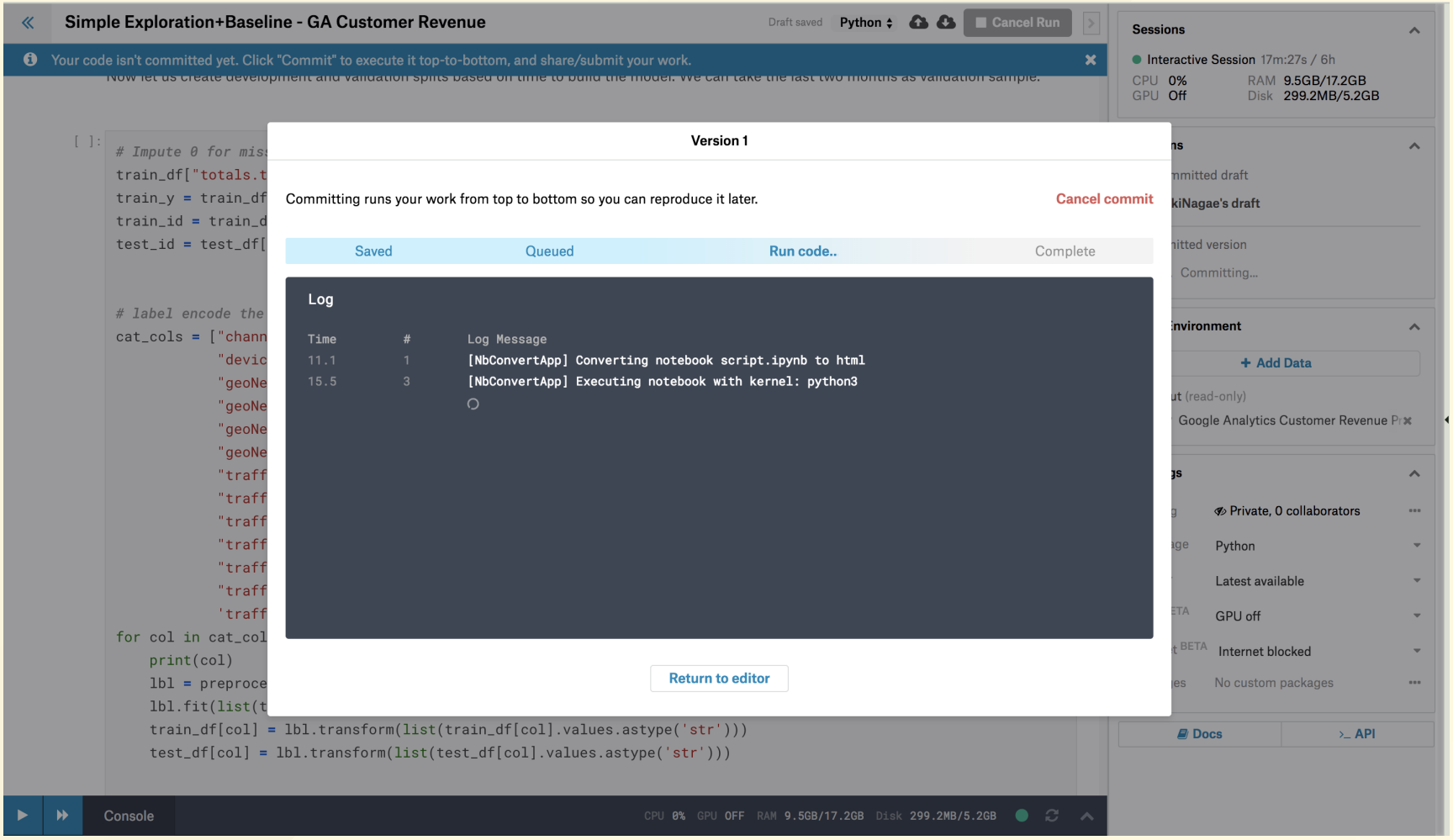
6. めんどくさいので人のコードをforkする



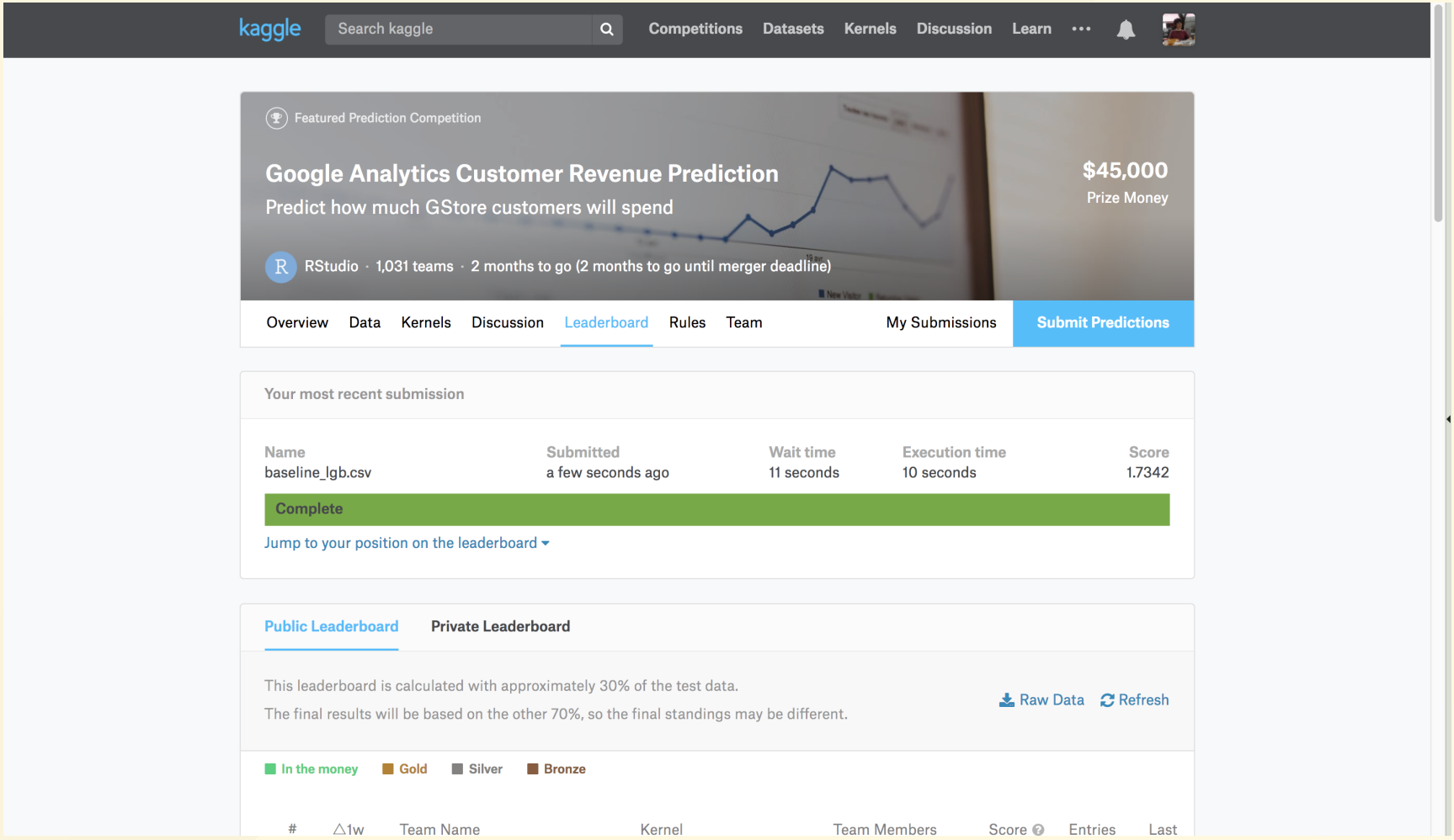
forkしたコードを実行するだけ(｀・ω・´)



実行中


Simple Exploration+Baseline - GA Customer Revenue
Draft saved
Python
Cancel Run
Your code isn't committed yet. Click "Commit" to execute it top-to-bottom, and share/submit your work.
Sessions
Interactive Session 17m:27s / 6h
CPU 0%
RAM 9.5GB/17.2GB
GPU Off
Disk 299.2MB/5.2GB
Version 1
Committing runs your work from top to bottom so you can reproduce it later.
Cancel commit
Saved
Queued
Run code..
Complete
Log
Time # Log Message
11.1 1 [NbConvertApp] Converting notebook script.ipynb to html
15.5 3 [NbConvertApp] Executing notebook with kernel: python3
Return to editor
Add Data
Google Analytics Customer Revenue
Private, 0 collaborators
Python
Latest available
GPU off
Internet blocked
No custom packages
Docs
API
Console
CPU 0% GPU OFF RAM 9.5GB/17.2GB Disk 299.2MB/5.2GB

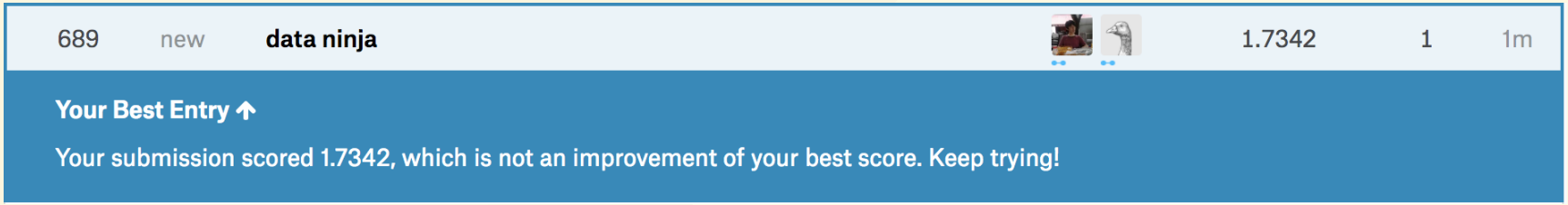
7. 予測を提出する


kaggle
Search kaggle
Competitions
Datasets
Kernels
Discussion
Learn
Featured Prediction Competition
Google Analytics Customer Revenue Prediction
Predict how much GStore customers will spend
$45,000
Prize Money
RStudio · 1,031 teams · 2 months to go (2 months to go until merger deadline)
Overview
Data
Kernels
Discussion
Leaderboard
Rules
Team
My Submissions
Submit Predictions
Your most recent submission
Name
baseline_lgb.csv
Submitted
a few seconds ago
Wait time
11 seconds
Execution time
10 seconds
Score
1.7342
Complete
Jump to your position on the leaderboard
Public Leaderboard
Private Leaderboard
This leaderboard is calculated with approximately 30% of the test data.
The final results will be based on the other 70%, so the final standings may be different.
Raw Data
Refresh
In the money
Gold
Silver
Bronze
#
△1w
Team Name
Kernel
Team Members
Score
Entries
Last

8. スコアとランクを確認

689位（全1,031チーム）

ちーん(｀・ω・´)



結局言いたいのは

パクった後が勝負

まとめ

- kaggleはデータサイエンティストのNo.1を決める大会
- 理論より実践のトレンド
- とりあえず人のコードをパクって頑張る
- kaggleは沼(｀・ω・´)

参考資料

- Kaggle - Wikipedia
- What is Kaggle, Why I Participate, What is the Impact?
- fast.ai · Making neural nets uncool again
- deeplearning.ai: Announcing new Deep Learning courses on Coursera

おわり(｀・ω・´)

ようこそkaggle沼へ