

# 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)

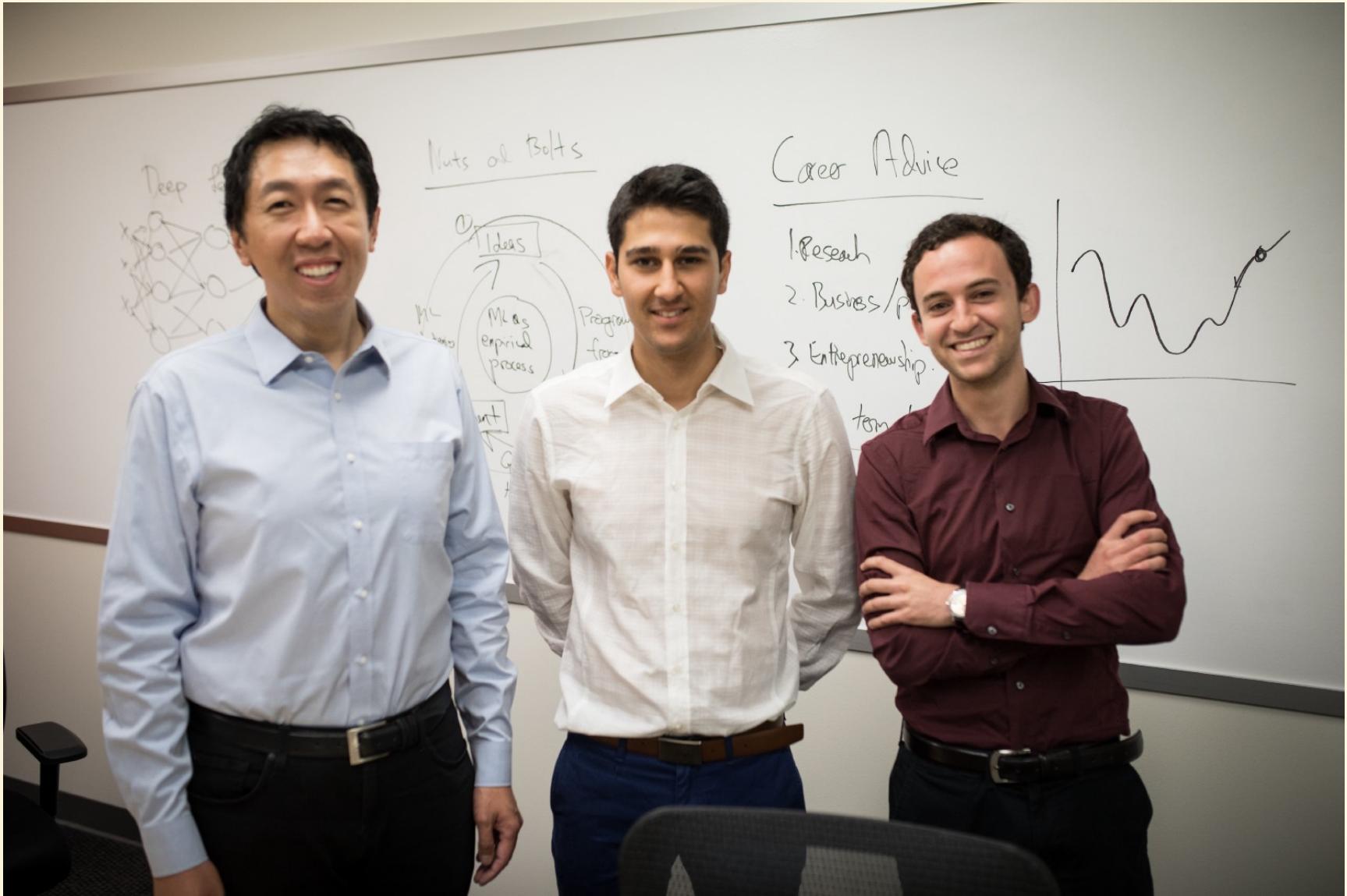


今

- 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](https://deeplearning.ai/Announcing%20new%20Deep%20Learning%20courses%20on%20Coursera)

# 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~~

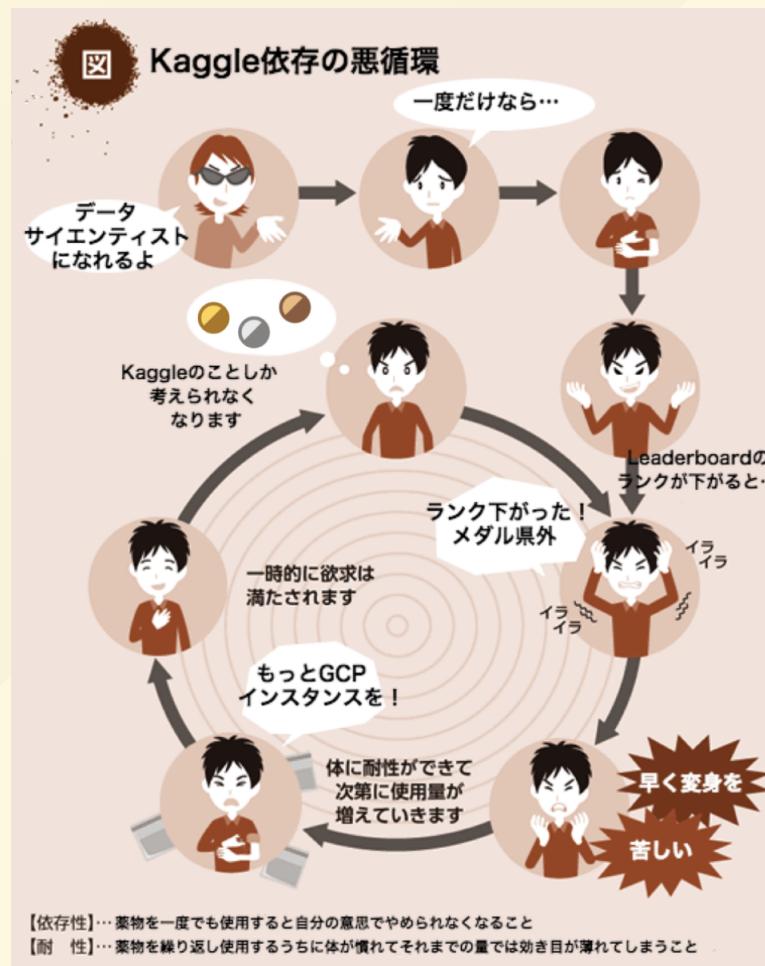
# 主催者側のメリット

- 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



onodera  
@Overfit

Following

新しいコンペが始まったので仕事できなくなつた

Translate Tweet

 Google Analytics Customer Revenue Prediction  
Predict how much GStore customers will spend  
[kaggle.com](https://kaggle.com)

6:32 AM - 14 Sep 2018

8 Retweets 48 Likes

RT

Comment 8 Like 48 Share

ということで

# 5. 早速kaggleをやってみた ( `・ω・' )

# 1. コンペを選ぶ

The screenshot shows the Kaggle website's 'Competitions' page. At the top, there is a navigation bar with links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', 'Learn', and a user profile icon. Below the navigation bar, the page title 'Competitions' is displayed, along with 'Documentation' and 'InClass' buttons. There are two filter buttons: 'General' and 'InClass', with 'General' being selected. A 'Sort by' dropdown is set to 'Grouped'. A search bar allows filtering by 'All Categories' or searching for specific competition names. The main content area displays three featured competitions:

- Google Analytics Customer Revenue Prediction**: Predict how much GStore customers will spend. Status: Featured · 2 months to go · tabular data, regression. Prizes: \$45,000. Teams: 847 teams.
- Titanic: Machine Learning from Disaster**: Start here! Predict survival on the Titanic and get familiar with ML basics. Status: Getting Started · Ongoing · tutorial, tabular data, binary classification. Prizes: Knowledge. Teams: 9,767 teams.
- House Prices: Advanced Regression Techniques**: Predict sales prices and practice feature engineering, RFs, and gradient boosting. Status: Getting Started · Ongoing · tabular data, regression. Prizes: Knowledge. Teams: 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

R 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** 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.

**Evaluation** 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.

**Prizes**

**Timeline**

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)

The screenshot shows a competition page for "Google Analytics Customer Revenue Prediction". The top banner features a chart showing revenue trends over time and a "\$45,000 Prize Money" badge. Below the banner, the competition title is "Google Analytics Customer Revenue Prediction" and the subtitle is "Predict how much GStore customers will spend". A note indicates "RStudio · 847 teams · 2 months to go (2 months to go until merger deadline)". The navigation bar includes links for Overview, Data, Kernels (which is underlined), Discussion, Leaderboard, Rules, and Team, along with a "New Kernel" button. The main content area displays a list of kernels sorted by "Hotness". The list includes:

- 111 A Very Extensive GStore Exploratory Analysis (4h ago) - Rmd, Py, 23 comments
- 58 Google Analytics EDA with screenshots of the app! (12h ago) - R, Py, 12 comments
- 26 Fixing Conflicts in the geoNetwork Attributes (17h ago) - Py, 3 comments
- 9 EDA -- Clustering --- 80:20 rule (14h ago) - R, 0 comments
- 19 rstudio LGB Single model LB1.6607 (1d ago) - R, Py, 0 comments

Below the list are filters for Outputs, Languages, Types, Tags, and a search bar for "Search kernels".

# いろんな人がコードを載せてくれるので助かる

```
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")
```

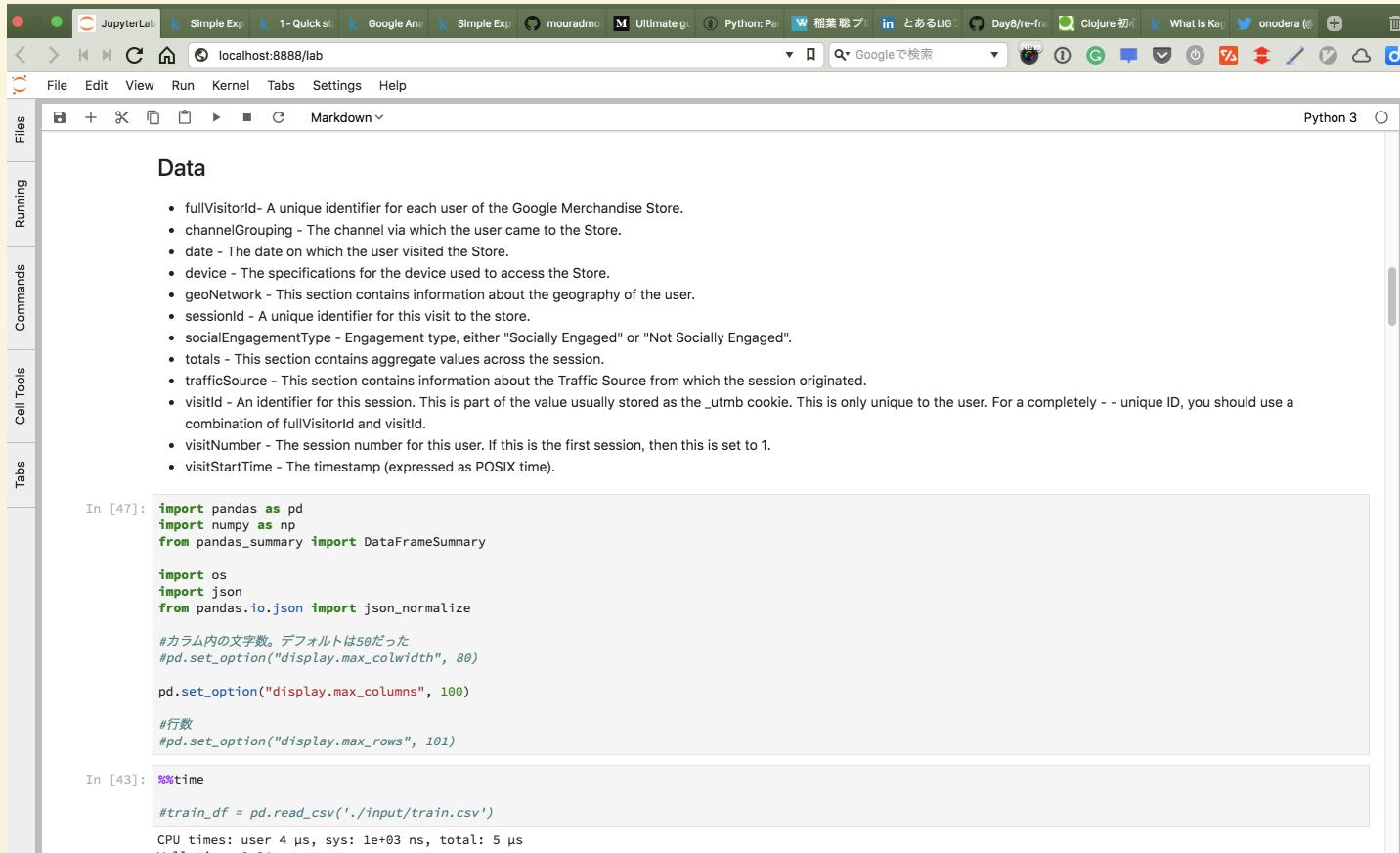
```
Loaded train.csv. Shape: (903653, 55)  
Loaded test.csv. Shape: (804684, 53)  
CPU times: user 5min 6s, sys: 12.9 s, total: 5min 19s  
Wall time: 5min 19s
```

```
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・` )



The screenshot shows a Jupyter Notebook interface running in a browser window. The left sidebar has tabs for 'Files', 'Running', 'Commands', 'Cell Tools', and 'Tabs'. The main area has a 'Markdown' tab selected. Below it, a 'Data' section contains a bulleted list of store-related 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.
- sessionId - A unique identifier for this visit to the store.
- 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).

In [47]:

```
import pandas as pd
import numpy as np
from pandas_summary import DataFrameSummary

import os
import json
from pandas.io.json import json_normalize

#カラム内の文字数。デフォルトは50だった
#pd.set_option("display.max_colwidth", 80)

pd.set_option("display.max_columns", 100)

#行数
#pd.set_option("display.max_rows", 101)
```

In [43]:

```
%time

#train_df = pd.read_csv('./input/train.csv')

CPU times: user 4 µs, sys: 1e+03 ns, total: 5 µs
```

# 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

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

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

The screenshot shows a Kaggle notebook page. At the top, there's a navigation bar with links for Competitions, Datasets, Kernels, Discussion, Learn, and a user icon. Below the navigation is a search bar and a main content area.

The main content area features a profile picture of a man with glasses and a purple shirt, next to a blue circular icon containing a white 'R'. The title of the notebook is "Simple Exploration+Baseline - GA Customer Revenue" by "SRK". It was last run 4 days ago, is an IPython Notebook HTML file, has 11,177 views, and uses data from "Google Analytics Customer Revenue Prediction". It is marked as "Public". There are 275 voters and a gold badge icon.

Below the title, there are tabs for Notebook (which is underlined), Code, Data (1), Output, Comments (65), Log, Versions (14), Forks (529), and a prominent blue "Fork Notebook" button.

Under the tabs, there's a section for "Tags" with buttons for "starter code", "data visualization", and "eda".

The main content area contains two sections: "Objective of the notebook:" and "Objective of the competition:". The "Objective of the notebook:" section describes the goal of exploring the dataset and building a baseline light gbm model. The "Objective of the competition:" section describes the challenge of analyzing a Google Merchandise Store dataset to predict customer revenue.

At the bottom right of the main content area, there's a "Code" button.

At the very bottom of the page, there's a small link "About the dataset".

# forkしたコードを実行するだけ( `・ω・` )

The screenshot shows a Jupyter Notebook interface with the following details:

- Title:** Simple Exploration+Baseline - GA Customer Revenue
- Language:** Python
- Session Information:** Interactive Session (52m:17s / 6h), CPU 0%, RAM 9.5GB/17.2GB, GPU Off, Disk 299.2MB/5.2GB
- Versions:** 1 uncommitted draft (YukiNagae's draft, based on V1) and 2 committed versions (V2 12m, V1 36m)
- Draft Environment:** Add Data, input (read-only), Google Analytics Customer Revenue
- Settings:** Sharing (Private, 0 collaborators), Language (Python), Docker (Latest available), GPU BETA (GPU off), Internet BETA (Internet blocked), Packages (No custom packages)
- Code Cell:** [ ]:

```
import os
import json
import numpy as np
import pandas as pd
from pandas.io.json import json_normalize
import matplotlib.pyplot as plt
import seaborn as sns
color = sns.color_palette()

%matplotlib inline

from plotly import tools
import plotly.offline as py
py.init_notebook_mode(connected=True)
import plotly.graph_objs as go

from sklearn import model_selection, preprocessing, metrics
import lightgbm as lgb

pd.options.mode.chained_assignment = None
```
- Console:** Shows the output area where the code would be run.

# 実行中

The screenshot shows a Jupyter Notebook interface with the following details:

- Title Bar:** Simple Exploration+Baseline - GA Customer Revenue
- Header Buttons:** Draft saved, Python, Cancel Run, Session dropdown.
- Session Panel:** Shows an Interactive Session (17m:27s / 6h) with resource usage: CPU 0%, RAM 9.5GB/17.2GB, GPU Off, Disk 299.2MB/5.2GB.
- Code Cell:** Contains Python code for data processing, including handling missing values and categorical encoding.
- Commit Dialog (Version 1):** A modal window titled "Version 1" with the message "Committing runs your work from top to bottom so you can reproduce it later." It includes a "Cancel commit" button and a "Run code.." button.
- Log Panel:** Displays log messages from the execution process.
- Environment Panel:** Shows the environment configuration, including Python 3.6, latest available, GPU off, and Internet blocked.
- Bottom Navigation:** Includes a "Console" tab, a progress bar, and a status bar showing CPU 0%, GPU OFF, RAM 9.5GB/17.2GB, Disk 299.2MB/5.2GB.

# 7. 予測を提出する

The screenshot shows the Kaggle interface for a competition titled "Google Analytics Customer Revenue Prediction". The competition details include:

- Featured Prediction Competition
- Google Analytics Customer Revenue Prediction
- Predict how much GStore customers will spend
- RStudio · 1,031 teams · 2 months to go (2 months to go until merger deadline)
- \$45,000 Prize Money

The navigation bar at the top includes links for Overview, Data, Kernels, Discussion, Leaderboard (which is underlined), Rules, Team, My Submissions, and Submit Predictions.

The "Your most recent submission" section displays the following information for a file named "baseline\_lgb.csv":

Name	Submitted	Wait time	Execution time	Score
baseline_lgb.csv	a few seconds ago	11 seconds	10 seconds	1.7342

A green button labeled "Complete" is visible next to the submission details. A link "Jump to your position on the leaderboard ▾" is also present.

The "Public Leaderboard" tab is selected, showing the following message: "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." Below this, there are download and refresh buttons.

Legend for the leaderboard: █ In the money █ Gold █ Silver █ Bronze

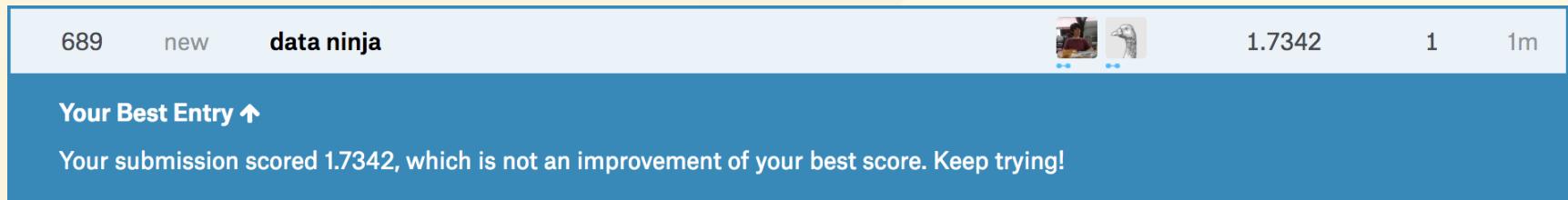
The table below lists the top entries:

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1							

# 8. スコアとランクを確認

689位 (全1,031チーム)

ちーん( `・ω・` )



A screenshot of a competition interface. At the top, it shows the rank (689), status (new), and team name (data ninja). To the right are two small profile pictures and a penguin icon. The score is listed as 1.7342 with a '1' and '1m' next to it. Below this, a blue banner displays the message "Your Best Entry ↑" and "Your submission scored 1.7342, which is not an improvement of your best score. Keep trying!"

**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・')

# 參考資料

- [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

