

# kaggleの始め方

# Agenda

- kaggleとは
- なんでやるの？
- どんな感じの流れ？
- やってみた

# 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

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・ ^ )

**なぜこんなことをやるの？**

# 理論から実践へのトレンドの 変化

昔

- 理解する

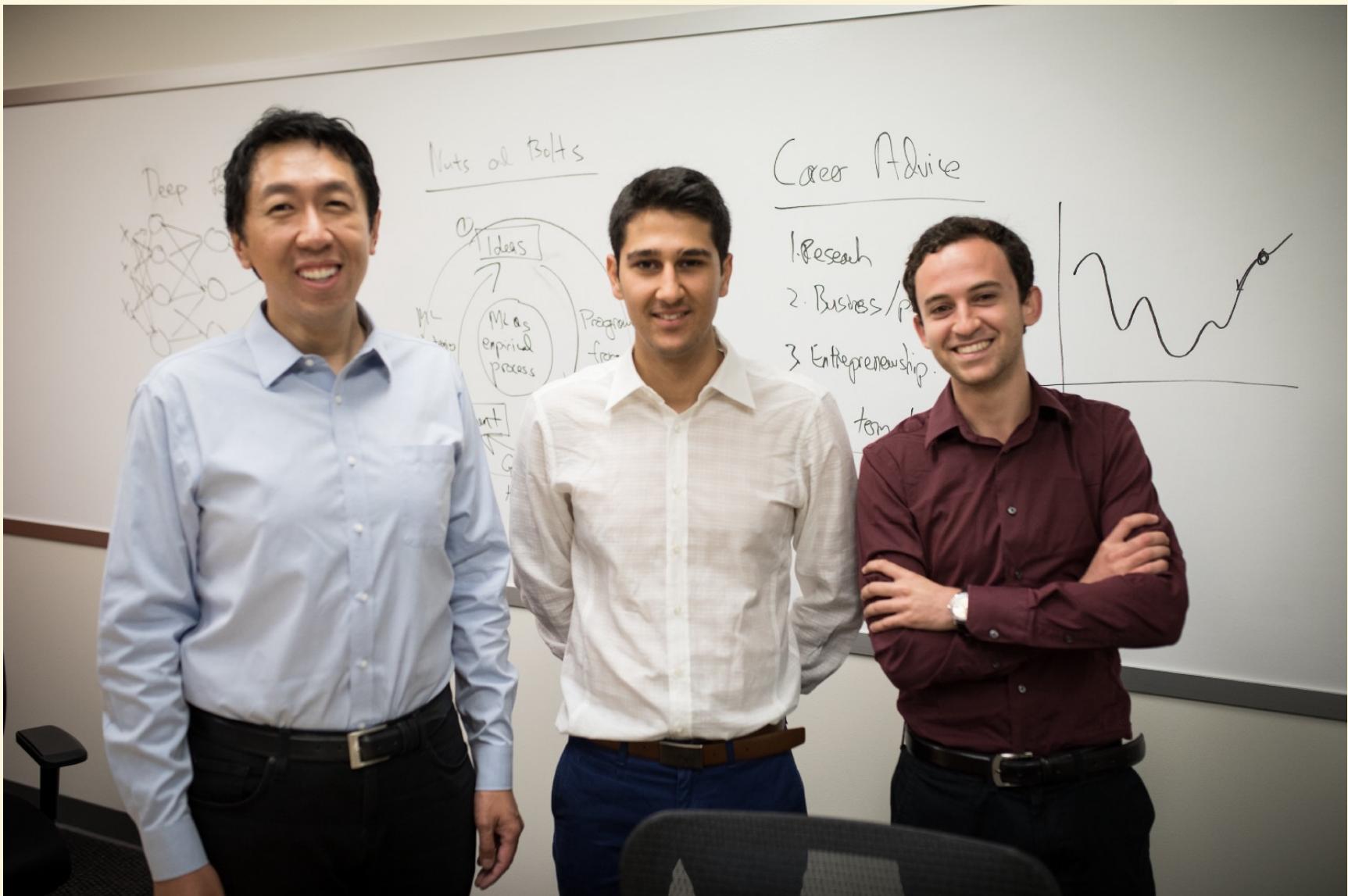


今

- 実践・役立つ

# 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-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.* ”

# kaggleの大まかな流れ

1. ホスト（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. ホストは精度が高い予測を提出した入賞者に賞金を払う

# kaggleをなんでやるのか？

# 参加者のメリット

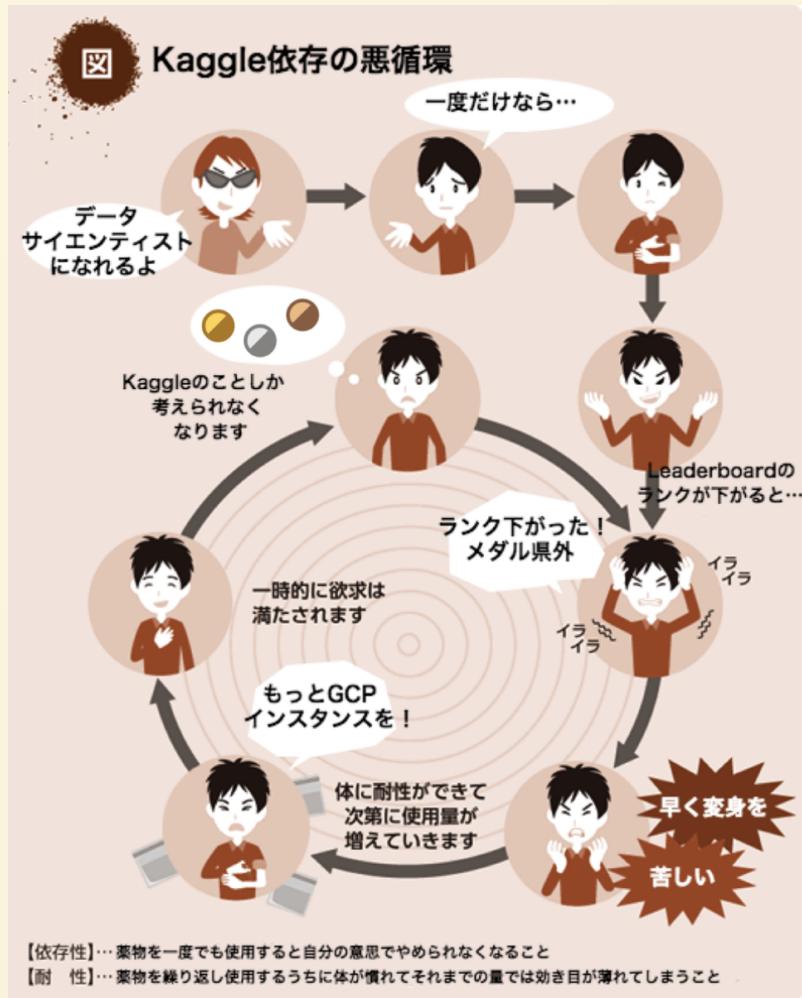
- ・ 様々なデータに触れられる（企業が実データを提供してくれる。レアい）
- ・ 他の参加者から学べる
- ・ 入賞すれば賞金 + 良い仕事をGET！
- ・ 楽しい

# 主催者側のメリット

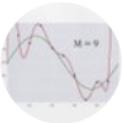
- 世界中のデータサイエンティストが問題解決の手法を試行錯誤してくれる
- ブランディング・PR
- データサイエンティストの採用

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つ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



  8  48  

ということで

# 早速kaggleをやってみた

# 1. コンペを選ぶ

kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ... 1

## Competitions

Documentation InClass

General InClass Sort by Grouped All Categories Search competitions

### 3 Entered Competitions

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

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

New Visitor

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** 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.

**Timeline**



## 2. コンペの内容を読む

1. 概要: 大まかにわかると色々アイディアが出て楽しい
2. 評価指標: これが一番大事ってわかんだね(｀・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

# 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 graph of revenue over time and a "\$45,000 Prize Money". Below the banner, the competition details are listed: "RStudio · 847 teams · 2 months to go (2 months to go until merger deadline)". The navigation bar includes "Overview", "Data", "Kernels" (which is underlined), "Discussion", "Leaderboard", "Rules", and "Team". A "New Kernel" button is located in the top right of the navigation bar. The main content area displays a list of kernels sorted by "Hotness". The list includes:

- 111 A Very Extensive GStore Exploratory Analysis (4h ago)
- 58 Google Analytics EDA with screenshots of the app! (12h ago)
- 26 Fixing Conflicts in the geoNetwork Attributes (17h ago) data cleaning
- 9 EDA -- Clustering --- 80:20 rule (14h ago)
- 19 rstudion LGB Single model LB1.6607 (1d ago) @ 1.6607

Each kernel entry includes a user profile icon, a title, a timestamp, and a set of buttons for viewing outputs, languages (R, Py), and comments.

# 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")
```

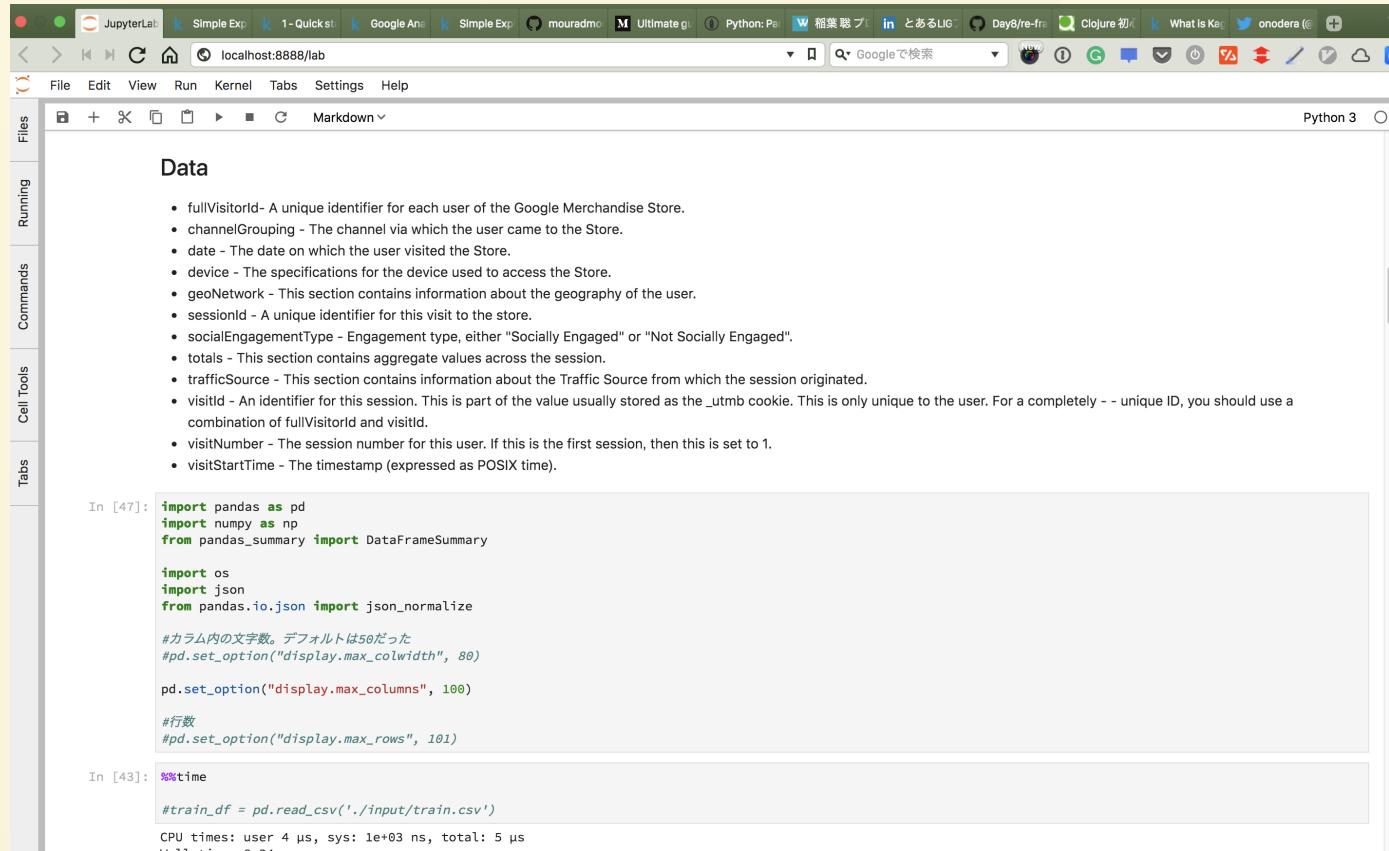
```
Loaded train.csv. Shape: (903653, 55)  
Loaded test.csv. Shape: (804684, 53)  
CPU times: user 5min 6s, sys: 12.9 s, total: 5min 19s  
Wall time: 5min 19s
```

```
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 with the following details:

- Header:** The browser title bar shows "localhost:8888/lab". The menu bar includes "File", "Edit", "View", "Run", "Kernel", "Tabs", "Settings", and "Help".
- Sidebar:** On the left, there are tabs for "Files", "Running", "Commands", "Cell Tools", and "Tabs".
- Content Area:**
  - Data:** A section containing a bulleted list of 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.
    - 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]:** A code cell containing Python imports and configuration:

```
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]:** A code cell containing a magic command and a file read:

```
%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 ^ 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

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 profile icon. Below the navigation bar, the main content area has a blue header with the title "Simple Exploration+Baseline - GA Customer Revenue" by "SRK". The header also displays a profile picture of the author, a large letter "R" icon, and a "275 voters" badge. Below the header, there are tabs for Notebook, Code, Data (1), Output, Comments (65), Log, Versions (14), Forks (529), and a prominent blue "Fork Notebook" button. Underneath these tabs, there's a "Tags" section with "starter code", "data visualization", and "eda" tags. The main content area is divided into sections: "Notebook", "Objective of the notebook:", "Objective of the competition:", and "About the dataset:". The "Notebook" section contains a brief description of the goal: "In this notebook, let us explore the given dataset and make some inferences along the way. Also finally we will build a baseline light gbm model to get started." The "Objective of the competition:" section describes the challenge: "In this competition, we are challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer." A small "Code" button is located at the bottom right of the main content area.

# forkしたコードを実行するだけ( `・ω・` )

« Simple Exploration+Baseline - GA Customer Revenue Draft saved Python Commit

**Objective of the notebook:**  
In this notebook, let us explore the given dataset and make some inferences along the way. Also finally we will build a baseline light gbm model to get started.

**Objective of the competition:**  
In this competition, we are challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer.

```
[ ]: 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 CPU 0% GPU OFF RAM 9.5GB/17.2GB Disk 299.2MB/5.2GB

**Sessions**  
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

2 committed versions  
V2 12m -0 -0 ✘  
V1 36m -0 -0 ✓

**Draft Environment**  
+ Add Data  
input (read-only) Google Analytics Customer Revenue Prx

**Settings**  
Sharing Private, 0 collaborators  
Language Python  
Docker Latest available  
GPU BETA GPU off  
Internet BETA Internet blocked  
Packages No custom packages

Docs API

# 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.

Now let us create development and validation splits based on time to build the model. We can take the last two months as validation sample.

Sessions

Interactive Session 17m:27s / 6h

CPU 0% RAM 9.5GB/17.2GB

GPU Off Disk 299.2MB/5.2GB

[ ]:

```
# Impute 0 for missing values
train_df["totals.transactionRevenue"] = train_df["totals.transactionRevenue"].fillna(0)
train_y = train_df['totals.transactionRevenue']
train_id = train_df['transactionId']
test_id = test_df['transactionId']

# label encode the categorical variables
cat_cols = ["channelGrouping", "deviceCategory", "geoNetwork.city", "geoNetwork.continent", "geoNetwork.latitude", "geoNetwork.longitude", "trafficSource.adwordsClickInfo.page", "trafficSource.adwordsClickInfo.query", "trafficSource.adwordsClickInfo.referrer", "trafficSource.adwordsClickInfo.source", "trafficSource.adwordsClickInfo.type", "trafficSource.adwordsClickInfo.visitNumber", "trafficSource.adwordsClickInfo.visitSource", "trafficSource.adwordsClickInfo.visitType", "trafficSource.adwordsClickInfo.visitNumber", "trafficSource.adwordsClickInfo.visitSource", "trafficSource.adwordsClickInfo.visitType"]
for col in cat_cols:
    print(col)
    lbl = preprocessing.LabelEncoder()
    lbl.fit(list(train_df[col].values))
    train_df[col] = lbl.transform(list(train_df[col].values.astype('str'))))
    test_df[col] = lbl.transform(list(test_df[col].values.astype('str'))))
```

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

Console

CPU 0% GPU OFF RAM 9.5GB/17.2GB Disk 299.2MB/5.2GB

Sessions

Interactive Session 17m:27s / 6h

Commited draft

kiNagae's draft

Commited version

Committing...

Environment

+ Add Data

Output (read-only)

Google Analytics Customer Revenue Project

Logs

Private, 0 collaborators

Language Python

Latest available

GPU off

Internet blocked

No custom packages

Docs API

# 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

R RStudio · 1,031 teams · 2 months to go (2 months to go until merger deadline)

\$45,000 Prize Money

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

Your most recent submission

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

 In the money  Gold  Silver  Bronze

#  $\Delta 1w$  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・` )

689	new	data ninja		1.7342	1	1m
-----	-----	------------	---	--------	---	----

Your Best Entry ↑  
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