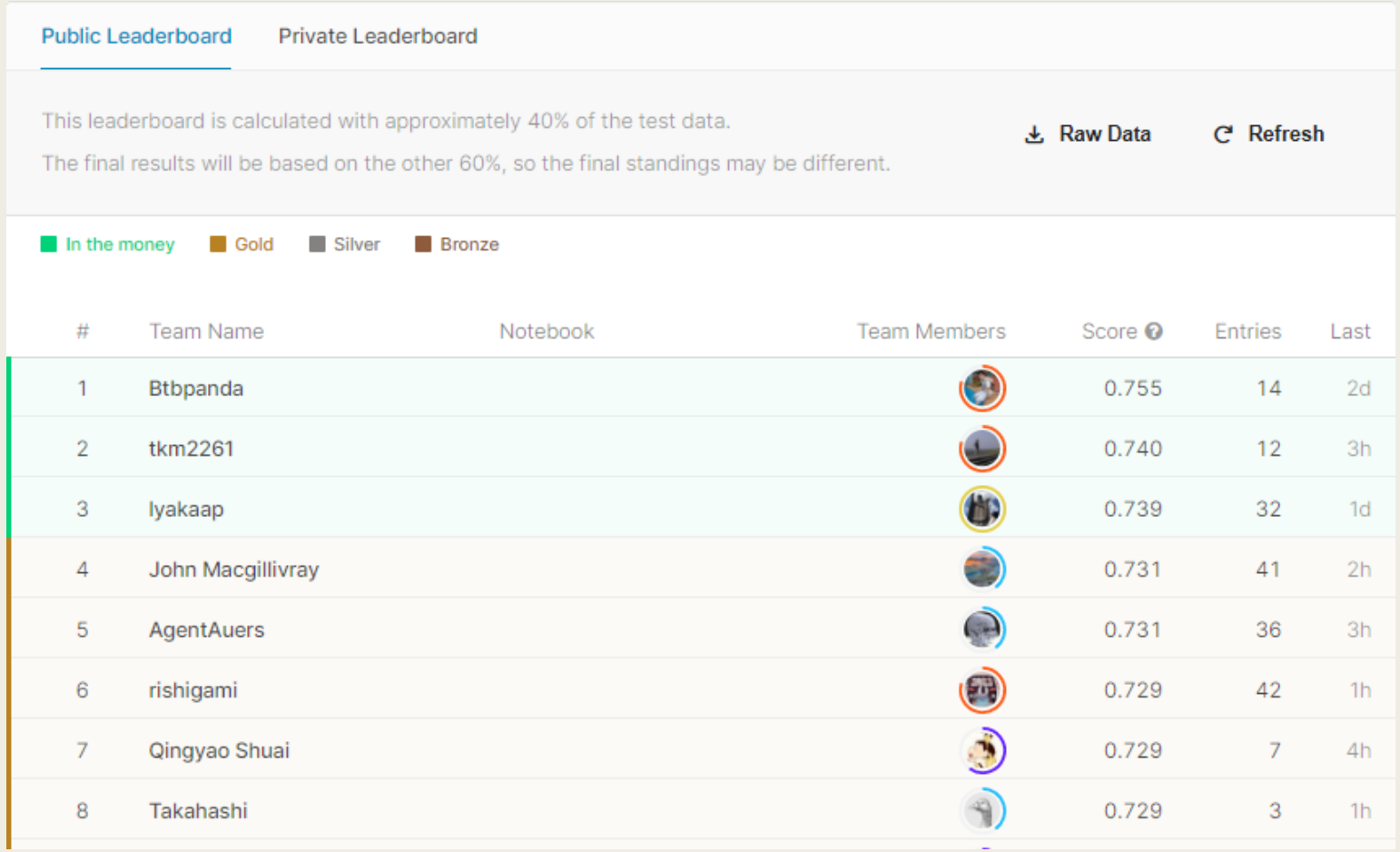
# PREDICTIVE ANALYSIS COMPETITION (PAC)

Hosted on Kaggle

# Competition

- Goal is to generate the best predictions at the end of the month-long competition.
- Allowed up to four submissions per day
- Results of the submission available soon after the submission.
- Position relative to others visible on competition Leaderboard.

# Sample Leaderboard



Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 40% of the test data.
The final results will be based on the other 60%, so the final standings may be different.

Raw Data | Refresh

In the money | Gold | Silver | Bronze

| # | Team Name | Notebook | Team Members | Score | Entries | Last |
|---|---|---|---|---|---|---|
| 1 | Btbpanda | | | 0.755 | 14 | 2d |
| 2 | tkm2261 | | | 0.740 | 12 | 3h |
| 3 | Iyakaap | | | 0.739 | 32 | 1d |
| 4 | John Macgillivray | | | 0.731 | 41 | 2h |
| 5 | AgentAuers | | | 0.731 | 36 | 3h |
| 6 | rishigami | | | 0.729 | 42 | 1h |
| 7 | Qingyao Shuai | | | 0.729 | 7 | 4h |
| 8 | Takahashi | | | 0.729 | 3 | 1h |


- Hosted on Kaggle, an online platform that runs data science competitions. Part of Google Cloud.
- Kaggle has over 15 million users in 194 countries (Wikipedia) who compete for
  - *Sport and Bragging rights*
  - *A Job with competition sponsor*
  - *A chance to showcase skills to recruiters*
  - *Prize Money*


- Through this competition, you will
  - *earn bragging rights*
  - *gain valuable hands-on experience with building models*
  - *have a chance to showcase skills to recruiters, and*
  - *earn points*

# ABOUT PAC

# About PAC

- Description
  - A credit card company has provided historical data on 50,000 of its active customers. The data includes demographic and behavioral information on customers.
- Goal
  - Build a predictive model to estimate monthly credit card spending of individual customers based on a rich set of customer attributes, including demographics, credit behavior, transaction activity, and lifestyle indicators.

* Disclaimer: This data is to be used solely for the purpose of this course. It is not recommended for any use outside of this competition.

# About PAC

- Metric
  - You must train and validate predictive models that estimate monthly_spend. Submissions will be evaluated based on RMSE (root mean squared error) on the scoring data. Scoring data is split into a public and private dataset. Performance on the public dataset will be posted on Public Leaderboard which will be visible during the competition. Score on the private dataset will be shared on the Private Leaderboard at the conclusion of the competition. Your performance will be judged based on your rank on the Private Leaderboard. Lower your RMSE, higher your rank.
  - *Kaggle allows you to select the submission to use for Private Leaderboard. Unless you are extremely confident in the virtues of a particular submission, it is best to allow Kaggle to automatically use your top Public Leaderboard submission.*

* Disclaimer: This data is to be used solely for the purpose of this course. It is not recommended for any use outside of this competition.

# Deliverables

- Predictions submitted on competition site hosted on Kaggle
- Python code for
  - *best model*
  - *data wrangling and experimentation in arriving at the best model*
- Report
- Presentation

# Grading Criteria

- Commitment to the Project (25 points)
  - *Worked consistently on the Project.*
  - *First submission before specified date and a total of at least ten submissions.*
- Quality of Modeling and Presentation (50 points)
  - *Demonstrated adequate knowledge of data exploration, suitably prepared data for analysis, used a variety of predictive analysis techniques, and communicated results effectively.*
  - *Assessed by a brief report summarizing the data analysis process, neatly commented Python code for best model, Python code demonstrating the data wrangling process and models not used, and presentation.*
- Prediction Accuracy (75 points)
  - *Accuracy of predictions assessed by Rank on Private Leaderboard*

# GETTING STARTED

# Registration

- To register, click on PAC Registration on Classes under the module, PAC Launch.
- See instructions to register.

# First Submission

- Download data from Kaggle
- Read Data
- Construct Model
- Read scoring Data and apply model to generate predictions
- Construct submission from predictions
- Upload to Kaggle

# First Submission Code

- Read data and construct a simple model

```
import pandas as pd
data = pd.read_csv('analysis_data.csv')
X = data.loc[:,['credit_score', 'travel_frequency']]
y = data.monthly_spend
from sklearn.linear_model import LinearRegression
reg = LinearRegression()
reg.fit(X,y)
reg.predict(X)
```

- Read in scoring data and apply model to generate predictions

```
scoring_data = pd.read_csv('scoring_data.csv')
scoring_data_X = scoring_data.loc[:,['credit_score', 'travel_frequency']]
pred = reg.predict(scoring_data_X)
```

- Construct submission from predictions

```
submission_file = pd.DataFrame({'customer_id': scoring_data.customer_id, 'monthly_spend': pred})
submission_file.to_csv('submission_file_1.csv',index = False)
```

# PAC Timeline

October 30th

Registration Opens

November 6th

Deadline for entering first submission

December 9th

Competition Closes

# Good Luck