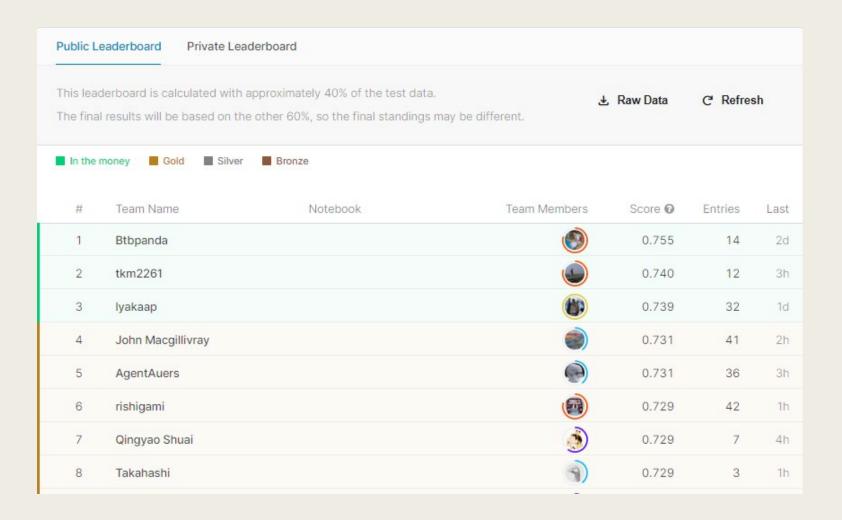
# PREDICTIVE ANALYSIS COMPETITION (PAC)

Hosted on Kaggle

#### Competition

- Compete to generate the best predictions.
- Goal is to generate the best predictions at the end of the month-long competition.
- Every submission is scored and results posted to leaderboard in real time.
- Can submit up to four predictions each day.
- Complete transparency. Positions of all participants are visible throughout the competition.

## Sample Leaderboard





- Hosted on Kaggle, an online platform that runs data science competitions
- 1M registered users and 60K active users compete on Kaggle 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
  - gave a chance to showcase skills to recruiters, and
  - earn points

# ABOUT PAC

#### About PAC

#### Description

 How much is your car worth? Predict the sale price of a car based on its features and condition\*

#### Goal

 Construct a model to predict sale price of a used car based on its features and condition and use it to generate predictions for a set of unlabeled data.

#### Metric

Submissions will be evaluated based on RMSE (root mean squared error)
 (Wikipedia). Lower the RMSE, better the model.

<sup>\*</sup> 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
- Presentation
- Report and supporting R code for
  - best model
  - data wrangling and experimentation in arriving at the best model

### **Grading Criteria**

- Commitment to the Project (25 points)
  - Worked consistently on the Project.
  - First submission before October 29<sup>th</sup> and a total of at least ten submissions.
- Quality of Modeling (50 points)
  - Demonstrated adequate knowledge of data exploration, suitably prepared data for analysis, used a variety of predictive analysis techniques discussed in class, and communicated results effectively.
  - Assessed by a brief report summarizing the data analysis process supplemented by neatly commented R code for the best submission.
- Prediction Accuracy (75 points)
  - Accuracy of predictions at the end of the Project.
  - Assessed by Rank on Leaderboard

#### Methods You Are Allowed to Use

#### PCA is ok

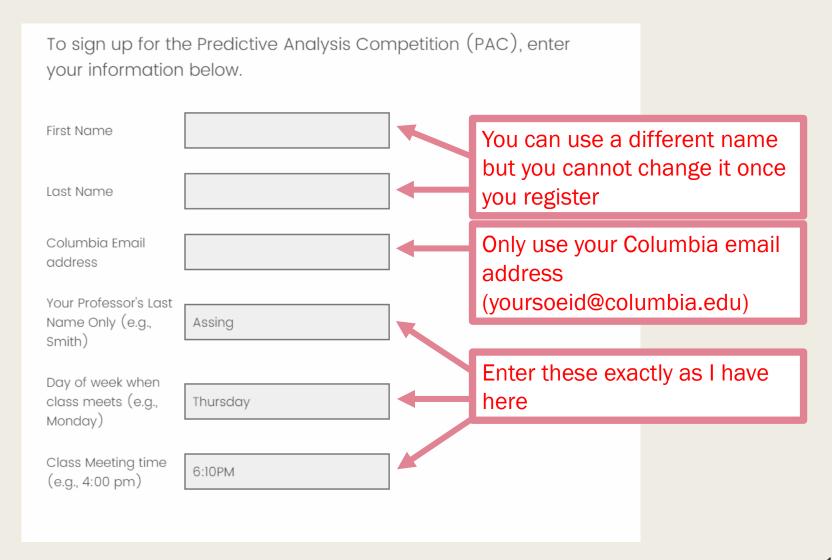
- You can only use models and techniques we covered in Frameworks I! No neural nets. No unsupervised learning methods. The following list is acceptable for the competition.
  - All tidying and modeling techniques discussed in Modules 1-3
  - Multiple Regression Module 4
  - Logistic Regression Module 5. I will also allow the use of the multinomial function from the NNET library – no other functions from NNET.
  - All feature selection techniques discussed in Module 6.
  - Regression and Classification Trees Module 8
  - Tuning and Ensemble Models (Random Forests, XGBoost) Module 9
  - Support Vector Machines Module 10
  - Parallel Processing techniques Module 11

# GETTING STARTED

# Registration

■ To register for PAC, <u>click here and follow directions</u>

### Registration



#### First Submission – Due October 29<sup>th</sup>

- 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 - Due October 29th

Read data and construct a simple model

```
data = read.csv('analysisData.csv')
model = Im(price ~ daysonmarket,data)
```

- Read in scoring data and apply model to generate predictions scoringData = read.csv('scoringData.csv') pred = predict(model,newdata=scoringData)
- Construct submission from predictions

```
submissionFile = data.frame(id = scoringData$id, price = pred)
write.csv(submissionFile, 'sample_submission.csv',row.names = F)
```

#### PAC Timeline

- October 16<sup>th</sup>: Registration Opens
- Oct 29<sup>th</sup>: Deadline for entering first submission
- Nov 17<sup>th</sup>: Competition Closes
  - You are responsible for making sure your best model is submitted correctly by the deadline. I cannot reopen the competition for you to resubmit your model after the deadline. If you submit a poor model accidentally you are stuck with the RMSE. I cannot consider a better model in your grading after the deadline.

This counts toward your final grade!

## Good Luck