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

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

kaggle

- 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*

kaggle

- 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.

* 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 29th 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, click here and follow directions

Registration

To sign up for the Predictive Analysis Competition (PAC), enter your information below.

First Name	
Last Name	
Columbia Email address	
Your Professor's Last Name Only (e.g., Smith)	Assing
Day of week when class meets (e.g., Monday)	Thursday
Class Meeting time (e.g., 4:00 pm)	6:10PM

You can use a different name but you cannot change it once you register

Only use your Columbia email address (yoursoeid@columbia.edu)

Enter these exactly as I have here

First Submission – Due October 29th

- 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 = lm(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

This counts toward your final grade!

- October 16th: Registration Opens
- Oct 29th: Deadline for entering first submission
- Nov 17th: 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.*

Good Luck