PREDICTING TENNIS MATCH RESULTS

Lauren Dellon

Springboard Data Science Career Track

Capstone Project

2021

THE PROBLEM STATEMENT

- Determine what statistics are most important in determining the winner of a tennis match
- Build a machine learning model capable of predicting tennis match outcomes (win/loss)
- This is a classification problem
- Possible clients:
  - Tennis players and coaches
  - Tennis bettors
  - Tennis sponsors

THE DATA

- Available on Kaggle
- Downloaded as 21 CSV files from: https://www.kaggle.com/pablodroca/atp-tennis-matches-20002019?select=atp_matches_2000.csv
- Original DataFrame shape: 59,340 rows and 32 columns
- Each row represented a single match, including statistics for both the winner and the loser

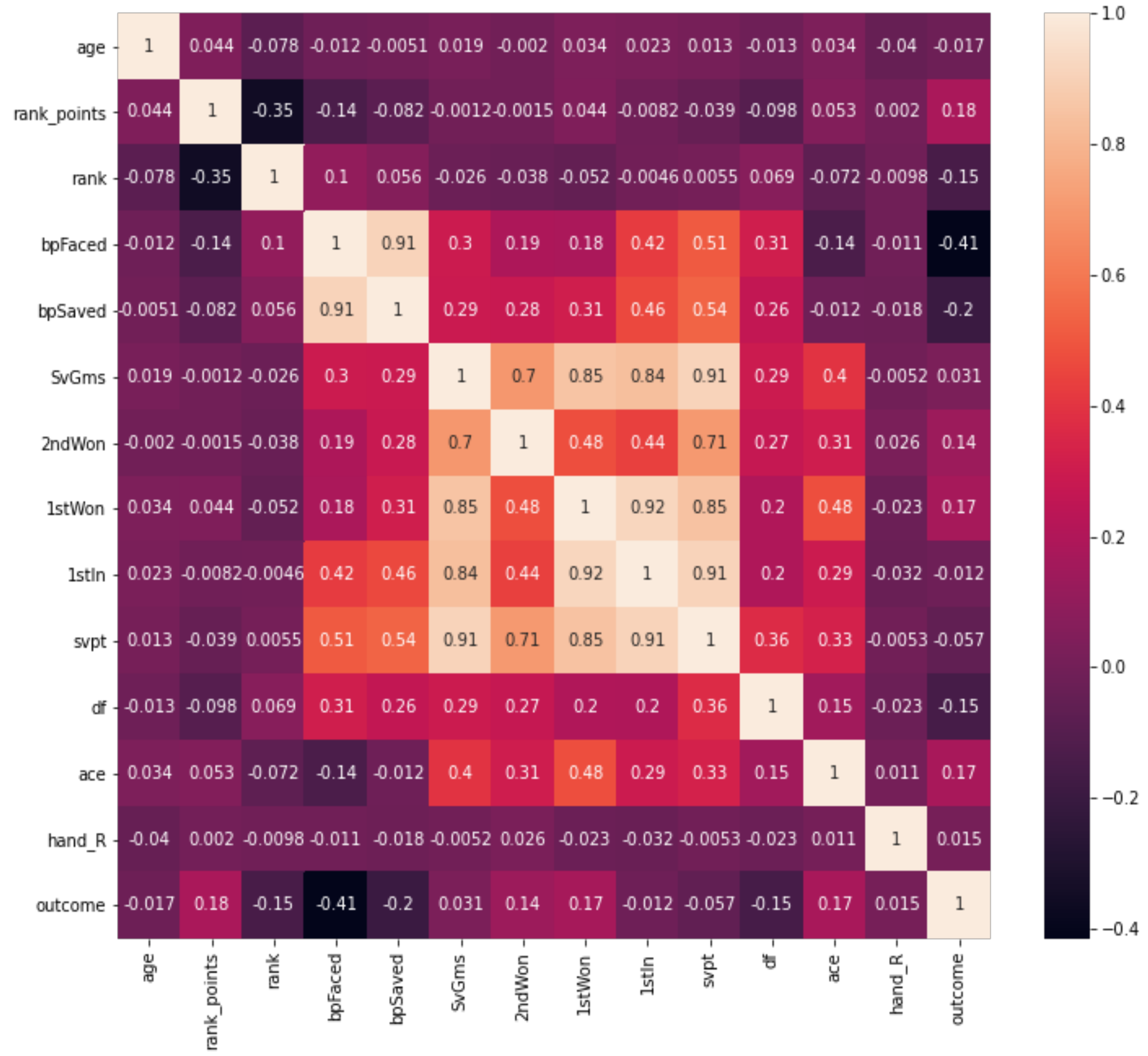
DATA TRANSFORMATION

- Created a new DataFrame where each row represented either a winner or loser
- Added outcome column where 1 represents a win and 0 represents a loss
- Final DataFrame shape: 86,230 rows and 14 columns (13 features and 1 outcome column)

	age	rank_points	rank	bpFaced	bpSaved	SvGms	2ndWon	1stWon	1stIn	svpt	df	ace	hand_R	outcome
0	26.0	810.0	63.0	4.0	4.0	17.0	26.0	58.0	73.0	117.0	4.0	8.0	1	1
1	29.0	1083.0	38.0	5.0	3.0	15.0	15.0	49.0	68.0	98.0	2.0	8.0	1	1
2	27.0	1835.0	19.0	7.0	6.0	10.0	12.0	37.0	43.0	76.0	6.0	9.0	1	1
3	26.0	275.0	185.0	0.0	0.0	11.0	10.0	39.0	43.0	58.0	0.0	12.0	1	1
4	31.0	1050.0	40.0	3.0	2.0	15.0	21.0	40.0	52.0	87.0	4.0	15.0	1	1

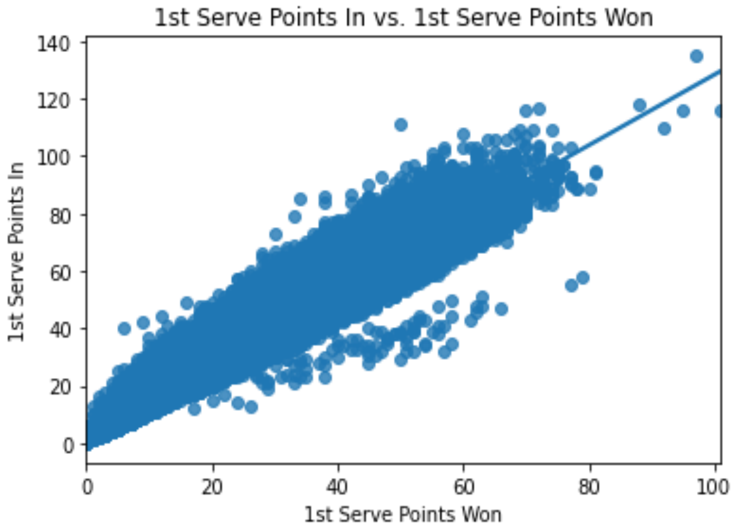
EXPLORATORY DATA ANALYSIS

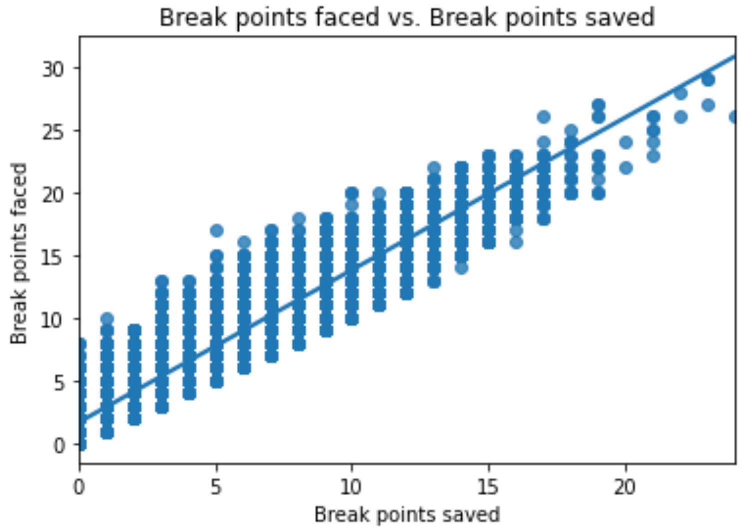
Heatmap showing correlations between features:



CORRELATIONS BETWEEN FEATURES

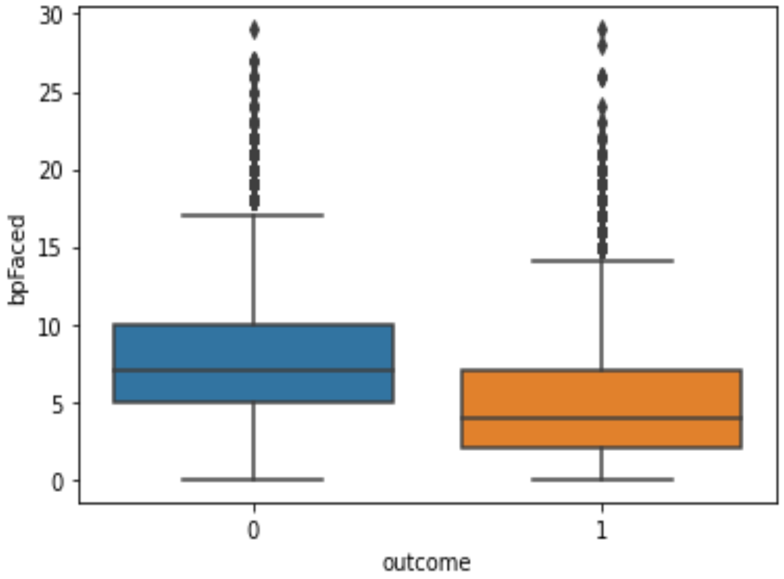
- From the heatmap, several features appear to be highly correlated
- Examples:

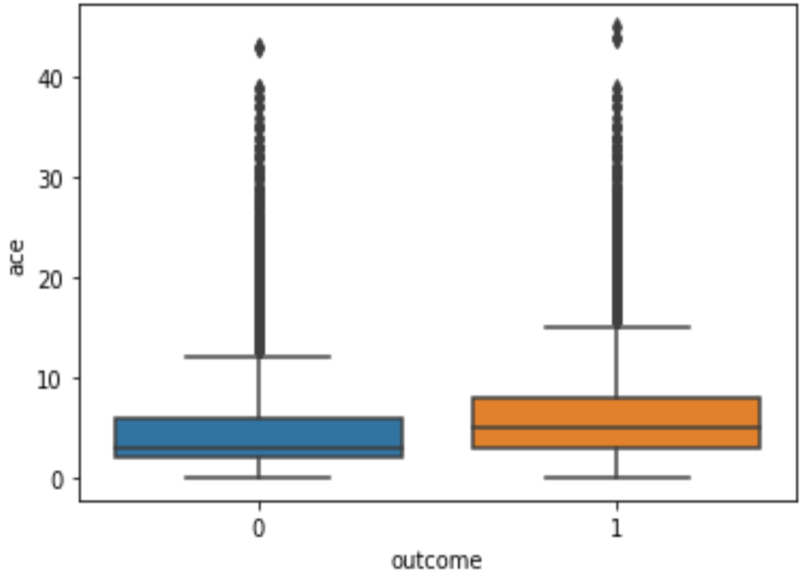




FEATURE EFFECTS ON OUTCOME

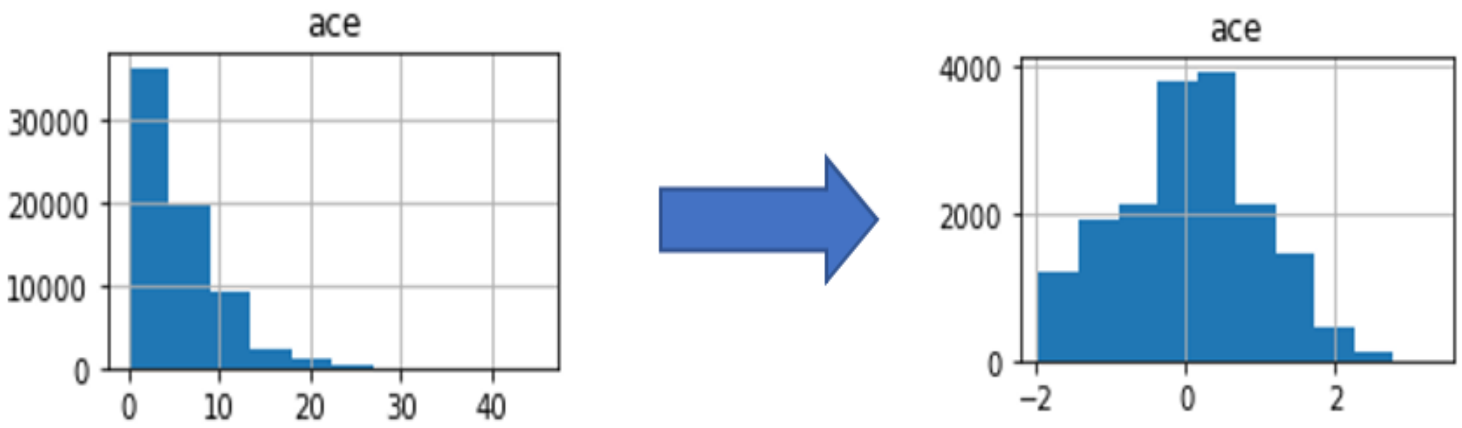
- Examples: bpFaced and ace





PREPROCESSING

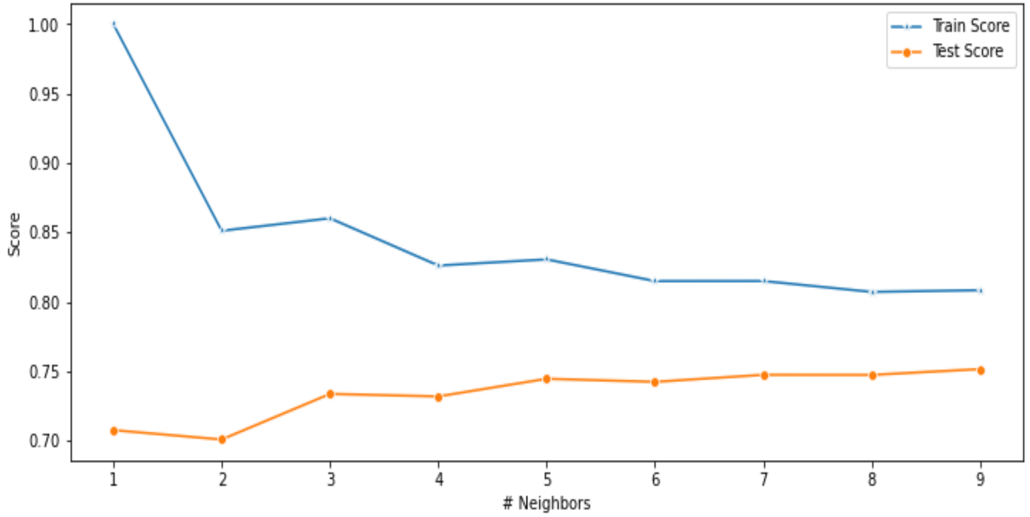
- Standardization and log transformation of features
- Example:



- Split data into training set (80%) and testing set (20%)
- Final shape of X_train: 68,984 rows by 13 columns

MODELING

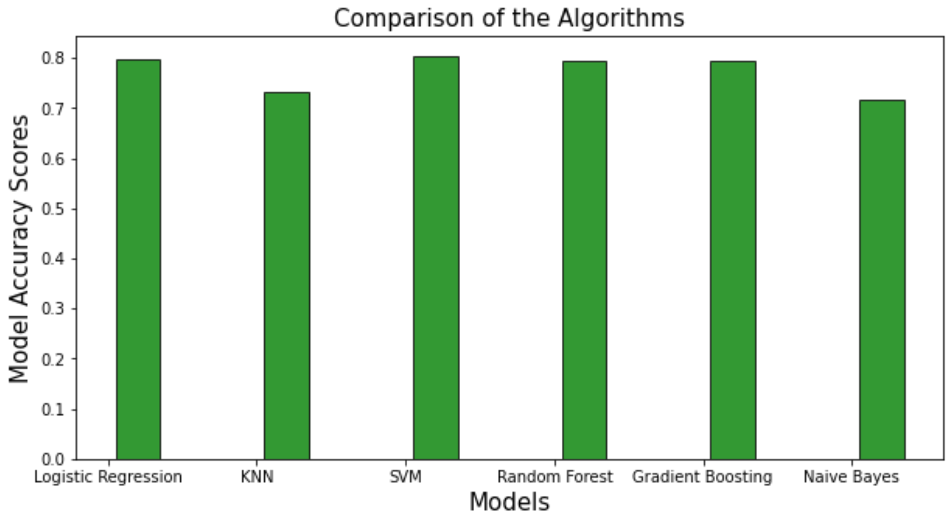
- Logistic Regression
  - Best initial value of regularization parameter, C: 0.01
- K-Nearest Neighbors (KNN)
  - Ideal number of neighbors: 3



- Support Vector Machine (SVM)
- Random Forest
- Gradient Boosting
- Naïve Bayes

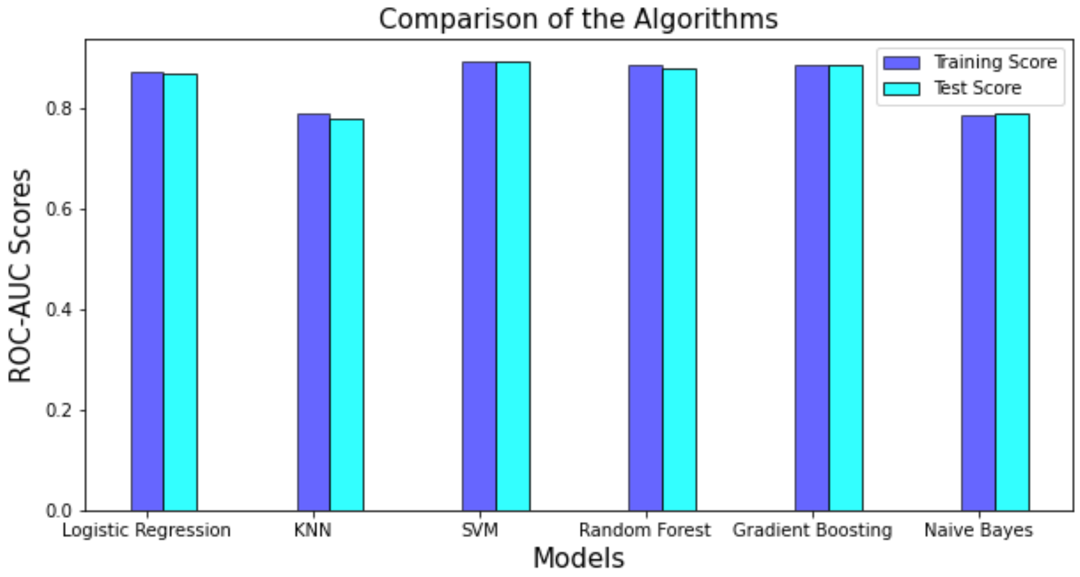
RESULTS: ACCURACY SCORES

Algorithm	Model Accuracy Score
Logistic Regression	0.798736
KNN	0.733677
SVM	0.803897
Random Forest	0.793633
Gradient Boosting	0.794329
Naive Bayes	0.715354



RESULTS: ROC-AUC TRAIN/TEST SCORES

Algorithm	ROC-AUC Train Score	ROC-AUC Test Score
Logistic Regression	0.871921	0.870983
KNN	0.790149	0.779892
SVM	0.894223	0.893458
Random Forest	0.888146	0.881503
Gradient Boosting	0.887809	0.888341
Naive Bayes	0.786312	0.789141



RESULTS: TRAIN AND PREDICT TIMES

Model	Train Time (s)	Predict Time (s)
Logistic Regression	0.1546	0.0028
K-Nearest Neighbor	0.2018	3.6273
Support Vector Machine	81.6413	8.9006
Random Forest	7.4065	0.3300
Gradient Boosting	6.9962	0.0343
Naïve Bayes	0.0230	0.0052

- Based on efficiency and model accuracy score, **Logistic Regression** is the optimal model.

HYPERPARAMETER TUNING

- Performed a Grid Search for Logistic Regression for the regularization parameter (C) and the solver

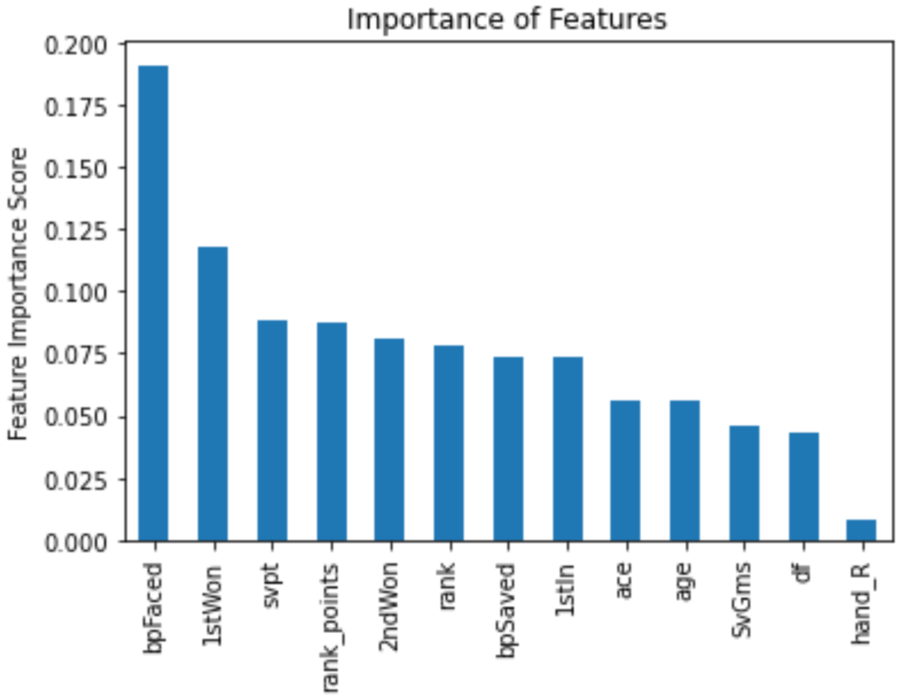
- Optimal parameters:
  - C = 1
  - Solver = newton-cg

- **Final model accuracy score: 79.42%**

FEATURE IMPORTANCE

- Using Random Forest Model to get feature importance:

- 2 most important features:
  - Break points faced
  - 1st serve points won



CONCLUSION

THE MOST IMPORTANT STATISTICS IN A TENNIS MATCH ARE THE NUMBER OF BREAK POINTS WON AND THE NUMBER OF 1ST SERVE POINTS WON

TENNIS PLAYERS NEED TO FOCUS ON THOSE ASPECTS OF THEIR GAME TO WIN MATCHES

FUTURE DIRECTION

- Here I have used data from the years 2000-2019, but data is available starting from 1968. The model could possibly be improved if I use data from earlier years.
- I could investigate model stacking, which utilizes multiple learning algorithms.
- Finally, I used all 13 features in my models. I could define an importance cutoff and use features only with an importance higher than the cutoff. This could improve the model accuracy.

ACKNOWLEDGEMENT

- Mentor: Hassan Waqar Ahmad
- Kaggle
- Springboard Team

THANK YOU!