STAT 301-3: Final Report

Albert Li

June 3, 2018

Introduction

During the the 2014 to 2015 regular season of the NBA, 128,069 shots were attempted. A Kaggle user, DanB, was able to release records of these shot attempts as well as many variables related to each shot attempt. These variables included the game match-up, the location of the game (home or away), game outcome (win or loss), the final score margin, the shot number, game period, time on the game clock, time on the shot clock, number of dribbles, touch time, shot distance, points type, shot result, closest defender, defender distance, and the shooter's name. For this final project, my ultimate goal was to develop models using these variables to accurately predict the shot outcome. Developing an accurate model would allow me to have a greater understanding of which variables impacts whether NBA players make or miss their shot attempt.

Data Exploration Summary

The main exploratory data analysis can be found in the EDA folder. I found that shot distribution on time left on the shot clock shows a fairly normal distribution, with jumps at 0 seconds and 24 seconds. I also found that the shot distribution on the distance to the basket is bimodal, which suggests that most shot attempts are either very close to the basket such as a lay-up, or are very far away from the basket and most likely three point shot attempts. The lack of mid range shot attempts (long range two point attempts) suggests that teams believe they are inefficient and tend to avoid them. Shooting percentages stay fairly consistent throughout each game period but they decrease a bit in the fourth quarter and decrease rather significantly during overtime. This could suggest that fatigue plays a factor in a shot attempt's outcome. Defender distance also has a smaller effect on a shot attempt's outcome that what I expected. I found that shooting percentages don't begin significantly increasing until a defender is at least 17 feet away.

Data Processing

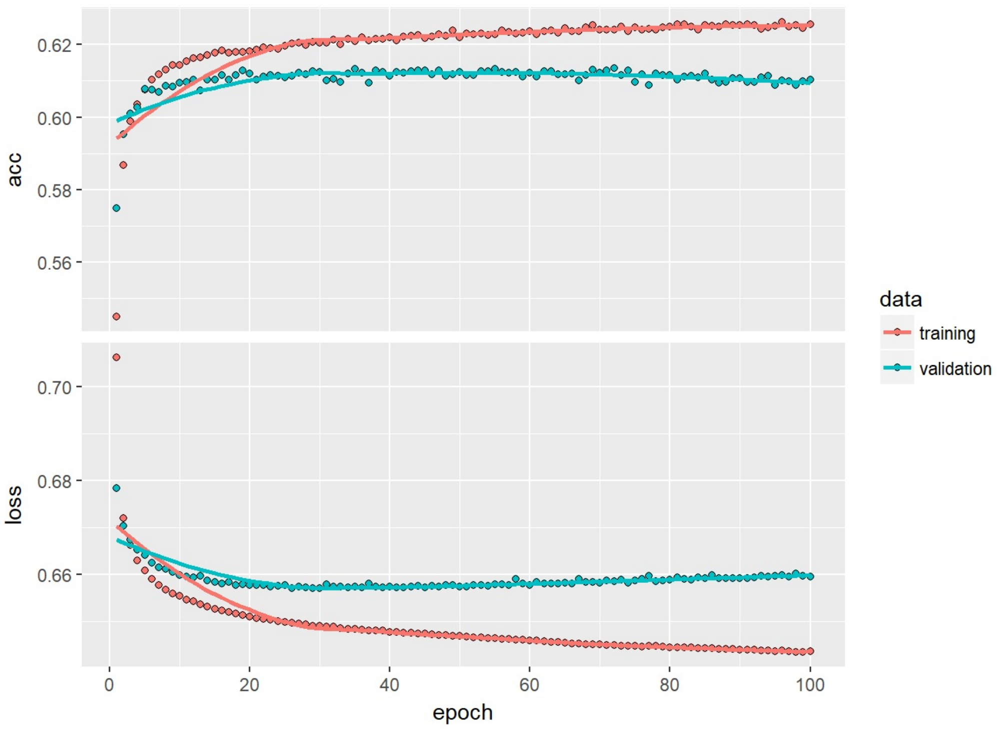
Before moving on to modelling, I needed to tidy up the dataset. There were many NA's in the shot clock column, which I believe are shot attempts that occur when the shot clock is turned off during the end of a period. To address this issue, I changed those NA's to 0. I excluded several redundant variables and variables that made no sense to use to predict shot outcome, such as final score margin and game outcome, since these variables are only recorded once the whole game is over. I changed game time left in the period to total number of seconds left in the period, and I also added a new variable stating the opposing team the shooter is playing against. The variables I ended up using in my models to predict shot outcome were location of the game, shot number, game period, time left on shot clock, touch time, shot distance, points type, closest defender distance, opposing team, and seconds left on the game clock. I was unable to use the shooter's name and closest defender's name as variables in my models since the modelling process when including those variables was unable to finish in less than 12 hours.

Modelling

The models I chose to implement in predicting shot outcome are the logistic model, linear discriminant analysis, quadratic discriminant analysis, bagging, random forest, boosting, and neural network. I attempted to use SVM methods but the model creation process was unable to finish in less than 10 hours. Rscripts on how I implemented these models can be found in the Modelling folder. Below is a table of the test accuracies from using each model.

Model	Test_Accuracy
Boosting	0.6195
Neural Network	0.6136
Logistic	0.6090
LDA	0.6087
Bagging	0.6031
Random Forest	0.6024
QDA	0.5585

Though the gradient boosting method produced the highest test accuracy, it is relatively low and not that much higher than the other models. The models seemed to range between .61 and .62, with the exception being quadratic discriminant analysis. Even when implementing a neural network, there were no significant gains made to accuracy after training a validation set for more than around five epochs. This can be seen in the graph shown below.



Potential for Improvement

I believe through more optimization some improvements to the models can be made, such as trying different model function arguments, figuring out how to efficiently use shooter name and defender name, and using SVM methods. Another possible feature I could take into account is a lagged variable that details whether or not the shooter's previous shot was a make or miss. Though these may produce some improvement, I question if these improvements can be significant enough. Even looking through some kernels on Kaggle similar to my project, the prediction accuracy also seems to max out in the mid .60s. I feel that modelling on this dataset could be more accurate if more advanced features are added such as height of shot release, jump height, or rotational speed of the ball. Currently, it may be unfeasible to capture these advanced features but with advancements in technology and willingness to collect data, more advanced features could be used to create a more accurate model to predict the outcome of a shot attempt.

Conclusion

The highest test accuracy I obtained was from my gradient boosting method, .6195, which I believe is not high enough to label as a useful model. The variables in the dataset are useful as the test accuracy is at least higher than .50. I believe that the inclusion of additional advanced variables can allow for more accurate and useful models. I look forward to when these variables are able to be recorded and are readily available to public datasets.

Citation

DanB. (2016). *NBA shot logs*. Retrieved from https://www.kaggle.com/dansbecker/nba-shot-logs (https://www.kaggle.com/dansbecker/nba-shot-logs)