NFL Score Predictor

Randall Li, Adam Askari, Robert Crosby

Baskin Engineering
UC SANTA CRUZ

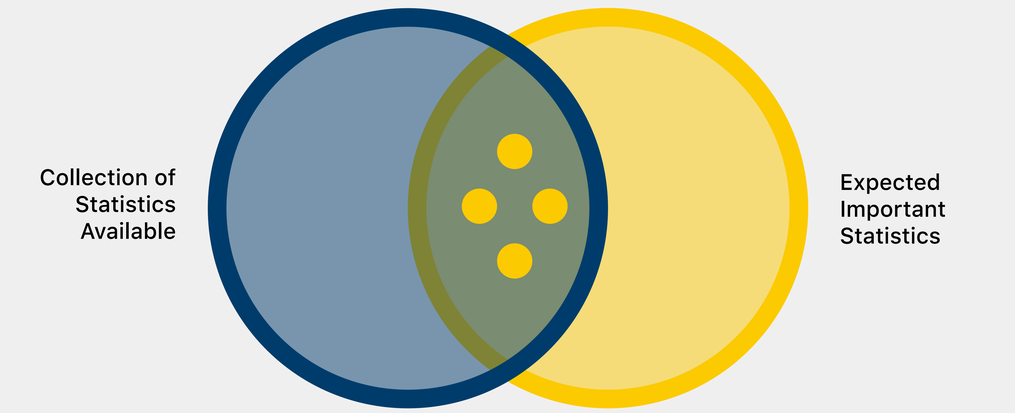
Abstract

Using data from scrapped from Pro-Football-Reference.com, our grouped used that data to predict the overall outcome of a NFL football game in terms of score. The purpose of this study was to find the key contributors that lead towards the outcome of a football game. The models which we used to test our data were all found from the Keras and scikit-learn libraries. These models include linear regression, random forest, and a neural network. From our findings, we noticed a significant trend between yards gained and allowed by either team that contributed to the overall score.

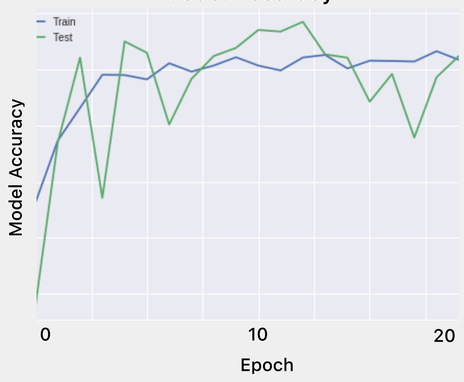
Approach

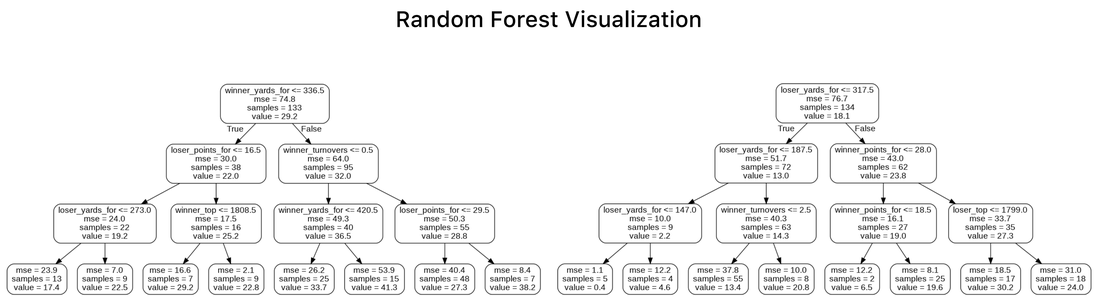
The first step of the project was gathering a significant dataset in order to have enough data for our machine learning algorithms to study from. We decided to initially only look at football statistics from the past 5 years of the NFL in order to have contemporary data that would more accurately predict future results since rules change every year. We also narrowed the types of statistics we would gather which would include offensive/defensive yards per game for a team, penalties committed, time of possession for a team etc. Once we found a website called Pro-Football-Reference.com and scrapped and parsed all the data we needed, we started to play around with some of the machine learning models. When running the machine learning models, we split the data into winner and loser statistics. The first model we configured was a linear regression model since our problem was regressive. However, the linear regression model was not outputting a high enough accuracy which transitioned us towards a neural network. From there we increased the datasets we used to include every season of the NFL from the 2000s onwards. The neural network gave us a good enough accuracy but it was hard to see what factors were determining the output, which lead us to the final model we used, random forest. Random forest helped us visualize our data but was not as powerful as a neural network.

Overview



Data for every NFL team between 2000 and 2018 was collected and stored locally. Then, every regular season game between 2000 and 2018 was scraped and stored. Prior to passing the data through a neural network, we iterated over every single game, setting a vector of the four (eight total) above statistics using the participating teams, and produced an output based on whichever team won. The model was trained on this data.





Analysis

Upon reviewing the graphs given to us by the random forest algorithm, it is clear that the biggest contributor to the score offensive and defensive yards by each team. A surprise to us was that time of possession had a very minimal influence on the outcome of the game since one would think the longer a team has the ball, the more that team would be able to score. Penalty yards were the second biggest factor which should tells teams that giving up free yards will not help them score points. The decision tree produced by random forest is also incredibly useful for analysis since if one follows the tree down, they too can predict the score of the game. It is also clear that from the outcome of neural network that we were able to achieve a model that does not overfit since the test and train lines match at the final result. Another contributor to that was running a small amount of epochs in order to reduce the chance of overfitting.

Results

The linear regression model produced an accuracy of around 50% percent for both the training and test set each time we ran it. Random forest overfit by an extreme margin by having a train set accuracy of 93% and a test set accuracy of 50%. Even through various testing on different parameters, we were unable to prevent the overfitting tendency of random forest on our dataset. Our neural network performed the best out of the three by having a consistent training set and test set accuracy of about 65%. There were some extreme fluctuations in the neural network test set during the first few epochs.

Conclusion

The dataset we collected from Pro-Football-Reference.com worked well with the machine learning algorithms we decided to employ. Linear regression was consistent but was not as effective as a neural network. Random forest helped visualize a correlation in our data while also being a fairly effective algorithm. Lastly, the neural network was able to produce the best results on the data compared to the other algorithms. We think the results we were able to produce were really good since having over 60% accuracy on a non binary variable shows that the model was effective.