Final Project Report

ajestridge40

April 2022

1 Introduction

In the world of sports, there are many different factors that lead to a teams success. Experts try to predict the outcome of sporting events by looking at the teams recent performances and the recent performances of their opponent. These experts tend to be correct more often than they are incorrect, but they are not perfect. Personal bias towards selecting the wining team will always subconsciously affect the expert decision. What if there was a machine learning algorithm that could accurately predict the winner of sporting events? This would be useful for teams to learn what factors lead to wins. They could make it a priority to refine that aspect of their team to win more games. There is also a large community around making bets on sports and wagering money on predicting the winner of the game. If this machine learning algorithm existed, The companies that host these bets would be more informed about the wager they are offering. If the algorithm was 75 percent accurate, it would already outperform many experts and would give insight on to the factors that lead to athletic success.

2 Literature Review

For the literature review portion of the proposal, I separated the important aspects of the articles into four themes. Each themed subsection will discuss how the individuals in the article approached the theme and how it is relevant to my project.

2.1 Data

In general, most of the peer reviewed articles followed a similar method for data collection. They would pull data from a teams previous six games to predict the outcome of the next game. This is because a teams recent performances speaks to how they will perform in their next game. Data should not be pulled from too far back because a lot can change for a team throughout the season. Players improve, they can digress, they can get injured, and also team morale can be a factor in the game. If data is pulled from previous seasons to make a prediction, it won't reflect the current state of the team.

2.2 Features

There were a few different ways that the articles discussed selecting features. One method was to use prior knowledge of the sport to select features you believe are important. This method makes sense if you are already an expert on the sport and know what factors lead to a victory. In the article titled A Critical Comparison of Machine Learning Classifiers to Predict Match Outcomes in the NFL, the authors used 42 different features that dealt with the teams recent performance (Beal et al.). Rather than selecting the features they believed were the best, they used any feature that they deemed to be relevant. Feature selection algorithms were mentioned in the papers but were not used in the projects.

2.3 Algorithms

Throughout the readings there were many algorithms that were used. The algorithms that were used most frequently were random forest, Bayesian nets, and decision trees. Bayesian nets was chosen because they represent the knowledge of the expert and the relation they perceive between features. Decision trees and random forest help to manage outliers in the data. They also mentioned that one possible improvement to the accuracy is to use ensemble learning techniques with the results of multiple algorithms. This will combine the results of multiple algorithms to find a consensus answer.

2.4 Results

Each models results hovered around the 63 - 67 percent accuracy range. There was one model that had an 80 percent accuracy but they only tested the model on one week of NFL play (Uzoma and Nwachukwu). I do not believe that when putting this model through a whole football season it would continue to show those results.

3 Technical Material

While working on this project, I had to learn about how to combine the machine learning algorithm results into a blender that then make its prediction. To do this I had to have a strong understanding of each machine learning algorithm I planned on using and how to combine them together in the blender.

3.1 SVC

I chose to use the SVC machine learning algorithm because it works well with high dimensional data sets. When working with the NFL, there are only fourteen games per week so I believed that the machine learning algorithms would have difficulty when training on only the 2016 season. Since the SVC algorithm is able to operate when there are more features than observations, I believe it would be a good selection for the project.

3.2 Random Forest

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training. The prediction is based on the output that was given by the highest number of decision trees. I chose to use the random forest machine learning algorithm because it is able to reduce over fitting and variance which helps to improve the accuracy.

3.3 Decision Tree

Decision Trees are machine learning algorithms that can be used for classification or regression. They create simple decision rules that they infer from the data. I chose to use the decision tree algorithm because it is simple and easy to understand. Decision Trees also offer good visualization to help understand the decision making process.

3.4 Logistic Regression

Logistic Regression is an algorithm that looks for relationships in the features to produce an output. This is very effective when there exists a linear relationship between the variables but that is not always the case. I chose to use this algorithm because when conducting my research I found that some people were able to get a high accuracy with the algorithm. I believe that there may be a linear relationship in NFL data which would mean that Logistic Regression would have a high accuracy.

3.5 K-Nearest Neighbors

K-Nearest Neighbors is an algorithm that assigns a class to each data point. It then looks at the data points k nearest neighbors and outputs the majority vote. I chose to use the K-nearest neighbors algorithm because it has a fast run time due to the fact that it has no training step and does all of its work during classification.

3.6 Blender

A blender is an ensemble learning technique where after multiple machine learning algorithms make their predictions, they are put into another machine learning algorithm to make the final prediction. This can improve accuracy because multiple machine learning algorithms will be working together to achieve the final result. I chose to incorporate a blender into my project because I did not see any other projects use this technique and I believe that multiple machine learning algorithms working together will increase the accuracy.

4 Methodology

When working with machine learning, there are multiple decisions that need to be made in regards to data, features, and algorithms. choosing the wrong data, features, or algorithms could have a significant impact on the accuracy of the model.

4.1 Data

I used weeks one through six of the 2016 NFL season to construct my data set. I knew that I wanted the final model to predict the outcome of the game based off the average statistics from each teams previous six games. This means that I had to incorporate similar averages into the training data. I created my own data set because I could not find a data set that was created in the way that would maximize the possible accuracy of the model. I first added each game from weeks one through six. After that I needed to incorporate training data that used average statistics from previous games. I added weeks three through six again, but this time I used the average of each teams previous two performances. This would allow the algorithm to learn how to manage the averages and how the performances from previous weeks impact the classification. I followed this same method to incorporate data that averaged the teams previous three games, four games, and five games respectively. After this was done, my training data set had 136 games.

4.2 Features

I used 34 features in my data set. 17 features represent the home teams performance and the other 17 features represent the away teams performance. The features I used are yards, passing yards, rushing yards, yards per play, fumbles lost, interceptions thrown, sacks allowed, third down efficiency, penalties, yards allowed, passing yards allowed, rushing yards allowed, average yards per play allowed, fumbles recovered, interceptions, sacks, and third down efficiency allowed. The home teams 17 features are put in first followed by the away teams 17 features. I selected the features based off prior knowledge of the sport and what experts believe lead to wins.

4.3 Algorithms

I used 5 different machine learning algorithms which passed their output to a blender. I used SVC, Random Forest, Decision Tree, K-Nearest Neighbor, and Logistic Regression algorithms. The reasoning behind their selection can be found in the Technical Material section of the report. The blender is a Logistic

Regression algorithm that uses five features (one for each machine learning algorithm). A blender that takes in the output from the five machine learning algorithms will likely perform better than any individual classifier.

5 Results

Since the model was trained on the 2016 NFL season, I did not want to test it on that same season. I created a new data set with the same features but data from the 2021 NFL season. I did this because I wanted to see if the model would perform well on a season that happened five years after the data that it was trained on. Each feature in the testing data set contained the average of the teams previous 6 games. The blender finished with an accuracy of 70 percent on predicting the outcome of 40 games. This accuracy was better than any of the individual machine learning algorithms. The best performing individual algorithms were the Decision Tree algorithm and the K-Nearest Neighbor algorithm which both achieved a 65 percent accuracy. All of the other algorithms achieved higher than 55 percent. This validates my claim that using a blender would improve the accuracy of machine learning algorithms when working with NFL data. Individual algorithms had errors that they did not share with other algorithms. The blender was able to take in the predictions and come to the correct result more effectively than any individual classifier did.

6 Future Work

There is still lots of work that can be done in sports prediction using machine learning. In the case of my project, adding more training data would likely be beneficial. Specifically, more data that contains the averages of a teams previous performance. Using better features would also help the accuracy of the model. Some possible features to investigate would be weather and total starting players that are injured. Often times good teams lose because they have key starting players who suffered a recent injury. As it stands, this model would have no way to predict that. Experimentation with other algorithms and hyper parameters to put into the blender could also help the accuracy. Most of my time on the project was spent experimenting with algorithms and hyper parameters to find the best models. I believe that with more time and experimentation that 75 percent accuracy is possible.

7 Code

All of the code I used for this project can be found at
https://github.com/ajestridge/Estridge-Final-Project

8 References

Ryan Beal, Timothy J. Norman and Sarvapali D. Ramchurn (2020) *A Critical Comparison of Machine Learning Classifiers to Predict Match Outcomes in the NFL*, Addison-Wesley Professional.

Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." Applied computing and informatics 15.1 (2019): 27-33.

Polikar, Robi. "Ensemble learning." Ensemble machine learning. Springer, Boston, MA, 2012. 1-34.

Uzoma, Anyama Oscar, and E. O. Nwachukwu. "A hybrid prediction system for american NFL results." International Journal of Computer Applications Technology and Research 4.01 (2015): 42-47.