NFL Game Outcomes and Predictive Modeling

Nicholas V Romano

nromano@bellarmine.edu

15 January 2024

Executive Summary

As someone who is an avid football fan and who just so happens to be a data science major, the best project topic had to be finding a way to incorporate my interests. Therefore, for my project, I plan on using the K-Nearest Neighbor, Support Vector Machine, and Decision Trees to predict results of NFL games using the teams previous game performances.

Project Idea

The goal of this project is to use machine learning methods and models to be able to predict the outcome of an NFL game. Initially, I was thinking about using previous player data to predict future player performances but at times player performances can vary a lot and injuries are difficult to predict. So, I decided to pivot slightly and focus on using previous team performances to predict whether they are going to win their next game. The approach that is being considered is to include the offensive and defensive team statistics for both teams over the previous few weeks and somehow integrate them into a model that will predict who would win. The reason for considering the performance of a team over a span of a few games is because it might help take into consideration teams are either going through somewhat of a slump and teams that have been playing on the top of their game. Also, I could validate the models with the NFL regular season that just ended.

Background

For this project, I will be using data from Pro Football Reference. Recently, I personally reached out to Sports Reference to get permission to use the data that they provide om their website and got permission. However, I wouldn't be able to use a web scrapper to automatically retrieve the data for me, rather I would have to manually get the data myself through exporting data to a csv file. The data I will be retrieving the box scores for each game and the schedule of games played over the past 4 to 5 seasons.

Matt Gifford and Tuncay Bayrak wrote an article in the *Decision Analytics Journal* titled, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression", in which they predicted NFL game results using a decision tree and a logistic regression model using team statistics from each game from 2002-2017. In doing so, they had misclassification rates of 21.6% and 16.9%, meaning their models were about 83% and 78% accurate in predicting game results (pg. 4-7). However, when it

comes to their models, Gifford and Bayrak explain, “it does not allow for the prediction of a game’s outcome without the final statistics” (pg 9). Furthermore, they go on to note two more things: (1) “This research can be further improved by segmenting the data by year or smaller groupings of years to understand how the impact of team statistics has changed over time” (pg. 9); (2) “the utilization of performances from previous games … one could predict a game before it happens” (pg.9). So that is the reasoning behind the project, use a smaller range of seasons – but big enough to have enough data – so the style of play doesn’t introduce anything that can mess with the modeling and use past team performances to predict games the upcoming games. It appears that Gifford and Bayrak used the team stat lines of a particular game as the inputs and the score of the target value to predict. So, using previous performance data to predict an upcoming game instead of using the future game’s own stat line to predict the game is a different approach than Gifford and Bayrak that might have less variability in team stats from week to week.

Modeling

As of now, the machine learning models that are in consideration in using for this project is K-Nearest Neighbor model, Support Vector Machines, and a Decision Tree. The K-Nearest Neighbor and Support Vector Machine models are being considered since teams with stronger stats might be nearer to each other and have a better chance of winning against their next upcoming opponent, and the same goes for teams with less impressive states but who have a lesser chance of winning. Therefore, the difference in these data points may be easy to identify, so these models are being considered. With the Decision Tree model, I would want to compare my different approach using a similar model as Gifford and Bayrak.

Tools

For this project, the tools that will be used includes mainly Python. Python was chosen as a tool for this project for its coding style and for the many libraries that it provides. The libraries in Python that will be incorporated into this project is Pandas, Scikit-Learn, NumPy, Seaborn, Matplotlib. Pandas and NumPy provide tools for storing and manipulating data within Python. Scikit-learn provides the resources in creating machine learning within Python. Seaborn and Matplotlib help in creating visualizations within Python. Another tool that might be used is Tableau. When it comes to creating visualizations, tableau provides many different graphs and is

rather easy to learn. Tableau come in handy with the exploratory data analysis of the project. Furthermore, when exporting the data on the Pro Football Reference, Excel might be of help in making sure the data is properly transferred over and might also help in the data preparation step of the project.

Conclusion

Inspired by my interest in football and wanting to improve previous works in predictive modeling, my project is to use predictive analysis to predict NFL football outcomes from past performances leading up to the game. More specifically, my project entails: gathering data from Pro Football Reference; cleaning, transforming, visualizing, and learning from the data in Python with the Pandas, NumPy, Seaborn, Matplotlib, and Scikit-Learn libraries. The models that are currently being considered to perform the predictive analysis are: K-Nearest Neighbor model, Support Vector Machines, and Decision Tree. In having a different approach than that of Gifford and Bayrak in what data we are feeding into the models, I hope to able to compare the results of my model with theirs.

References

Gifford, Matt, and Tuncay Bayrak. "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression." *Decision Analytics Journal,* vol. 8, 2023. *Science Direct,* https://doi.org/10.1016/j.dajour.2023.100296. Accessed 15 Jan. 2024

Sports Reference LLC. Pro-Football-Reference.com - Pro Football Statistics and History. https://www.pro-football-reference.com/.