Predicting NFL Outcomes Using Machine Learning

Nicholas Romano, Robert Kelley

Bellarmine University Data Science Program

Abstract

In recent years, the NFL has seen a rise in the use of sports analytics alongside an increasing volume of available data, in applications ranging from player health and safety to evaluating player performance. This project aims to develop a predictive analytics model to predict the results of upcoming NFL games by using team statistics of previous games. The dataset used for this project includes game data that was pulled from ESPN and includes a wide variety of team statistics for both the home and away team, ranging from third down and fourth down efficiencies to passing yards gained, rushing yards gained, and offensive time of possession. The data manipulation, visualization and modeling for this project was performed using Python.

Introduction/Objectives

This project focuses on incorporating offensive and defensive team statistics over the previous few weeks and somehow integrate them into a model that will predict who would win an upcoming NFL game.

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.

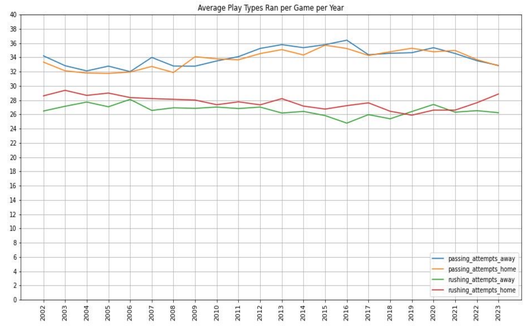
This project focuses on the future research directions proposed by Matt Clifford and Tuncay Bayrak in their *Decision Analytics Journal*:

- Limiting the data to a shorter window of seasons to train the model on
- Trying to primarily focus and train a model on previous game data.

Objectives

- The goal of this project is developing a predictive analytics model that can be used to predict the outcomes of NFL games using team's previous game statistics.

- Although the dataset used for this project includes data dating back to 2002, the model will be trained on data since 2012, with the idea of including game data when the NFL become more of a pass heavy offensive league.



Materials/Methods

Data Collection:

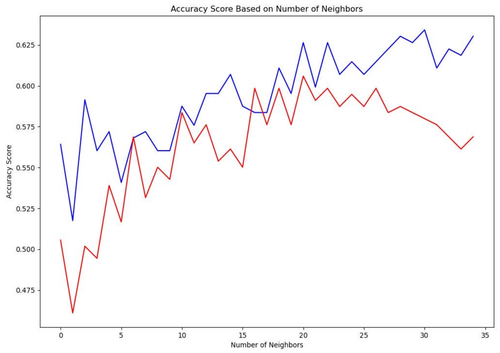
The data for this project was from a dataset from Kaggle posted by user CVIAXMIWNPTR. At the time of retrieval, the dataset contained game data dating back to the 2002-2003 to the 2022-2023 NFL seasons and includes both regular and postseason games. The user obtained the data by scraping the data from ESPN.

Data Cleaning

- Dealing With Games Resulting in a Tie
  - Out of the total number of games played in the entire dataset 14 games resulted in a tie (0.2482%). These 14 games were dropped from the dataset, enabling the target variable (game_result) to be binary (either the home team or away team wins).
- Dealing With Postseason Games
  - With the dataset including postseason games, this in turn causes teams to have differing number of games played. The solution to this was to obtain the dates of the first and last postseason games of each season and drop them from the dataset.
- Obtaining Team and Average Team Statistics
  - With the data containing both the home and away team statistics for each game, 32 separate data frames had to be created (one for each team).
  - For the team, the statistics were modified to be at the previous 4 game averages for the team for both offensive and defensive statistics.
- Condensing Number of Seasons
  - During the EDA of the dataset, it was found that beginning in 2012, there was an uptick in the number of passing plays ran per game alongside a slight downturn in the number of rushing plays ran per game. This provides the season to trim the dataset down to.

Model Selection

The model selected was the K-Nearest Neighbor (KNN) Classifier Model. The KNN Classifier Model looks at the k closest known datapoints to a datapoint that the model is trying to predict. In this case, the model will be looking at games where the home and away teams had similar statistics entering the games as the game the model is trying to predict.



Results

Model Selection (Cont.)

Train-Test-Validate Split of Data

- To train, test, and validate the model, the data need to be split. The validation data was the most recent year in the dataset (the 2022-203 NFL Season).
- The remaining games were used to create the training and test data. To match the roughly 270 games in the 2022-2023 NFL season, 10% of the remaining data was used to create the testing data. (2309-257-269).

Optimal Number of Neighbors

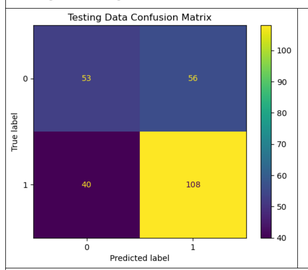
- The optimal number of neighbors to evaluate the future games was determined by using a for loop that increased the number of neighbors for the model while fitting the model on the training data, and then testing and validating the predictions on the test and validation data.

- The optimal number of neighbors that returned the highest accuracy score for both the testing and validation data was 21 neighbors.
- **Testing Data Accuracy:** 62.64591439688716%
- **Validation Data Accuracy:** 60.59479553903345%

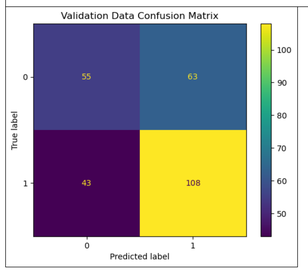
Model Results

Running the model with 21 neighbors, the following classification reports and confusion matrices were reported.

```
Testing Data Accuracy: 0.6264591439688716
              precision    recall  f1-score   support

           0       0.57      0.49      0.52       109
           1       0.66      0.73      0.69       148

    accuracy                           0.63       257
   macro avg       0.61      0.61      0.61       257
weighted avg       0.62      0.63      0.62       257
```




```
Validation Data Accuracy: 0.6059479553903345
              precision    recall  f1-score   support

           0       0.56      0.47      0.51       118
           1       0.63      0.72      0.67       151

    accuracy                           0.61       269
   macro avg       0.60      0.59      0.59       269
weighted avg       0.60      0.61      0.60       269
```



Conclusions

Conclusions

Using the KNN Classifier Model to predict the outcome of future games produces accuracy scores in the mid-to-upper fifty percent and lower sixty percent range. One possible reason for the lower accuracy scores is that the model has difficulties when it comes to accurately predicting away team victories. When trying to predict both the testing and validation data, the model has a large number of false negatives relative to the number of away team victories. This could be due to there being more games in which the home team wins.

Future Work

Possible future work for this project is trying different ranges for the number of games that are included in the rolling averages of the team statistics, similar to that of the optimal number of neighbors. Perhaps other models need to be taken into consideration like Support Vector Machines.

References

Tools

- Python v. 3.11.5
- Pandas v. 2.1.4
- NumPy v. 1.26.3
- Matplotlib v. 3.8.0
- Sci-kit learn v. 1.2/2

References

Kaggle Dataset

- cviaxmiwnptr. NFL Team Stats 2002 - Feb. 2023 (ESPN). *Kaggle*. Retrieved 18. Jan 2024 from https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-team-stats-20022019-espn.

Decision Analystics Jorunal NFL Article

- Clifford, 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

ESPN Schedule Webpage

- "NFL Schedule." ESPN. https://www.espn.com/nfl/schedule/_/week/5/year/2022/seasontype/3. Accessed 22 Feb. 2024.

Contact

Nicholas Romano

School Email: nromano@bellarmine.edu