**Machine Learning: Project Report**

Craig Shaffer

November 30, 2022

**Abstract**

In the United States today, the most viewed television broadcasts are games from National Football League. With people looking for new ways to interact with their favorite pastime, sports betting has exponentially grown. With money on the line, it is important to be able to determine who will win the game. This project aims to identify the winner of a football game through machine learning processes. Using historical data from the 2011-21 seasons, I built variables that are influential in determining the winner and constructed classification models to predict the likelihood of the home team winning the game.

Project Introduction………………………………………………………………...3

Data Description……………………………………………………………………3

Variable Description…………………………………………………………………...….4

Exploratory Analysis……………………………………………………………….7

Modeling…………………………………………………………………………..11

Variable Importance……………………………………………………………………...11

Generating and Evaluating Models…………………………………………………...….12

Results……………………………………………………………………………………13

Conclusion………………………………………………………………………...16

Areas of Further Study…………………………………………………………….16

References………………………………………………………………………... 17

**Project Introduction**

As fans look for new ways to interact with their favorite pastimes, sports betting has moved into the spotlight. The relationship between sports and gambling has always been there, however now it is becoming mainstream due to the relative ease of placing a bet through an app on your phone. Not only has sports betting grown, but viewership has also grown. In fact, since the Supreme Court legalized sports gambling in 2017, the NFL’s viewership has increased by over 20% (Grubman). Looking to be involved in the action or to make money, fans and other individuals partake in sports betting. By implementing machine learning practices, these people can be better informed when placing bets. Which ultimately increases the likelihood of turning a profit when gambling on such events.

This project aims to find a method to determine whether the home team will win the football game for sports betting purposes. Using data from the 2011-2021 NFL seasons, I perform the process of variable engineering, exploratory analysis, and modeling to find patterns to predict the winner of the football game. The goal is to be able to pick out games where the home team is predicted to win the game. This way, we know which team is favored to win and can place our bets appropriately.

**Data Description**

The historical data from the 2011-2021 seasons used for this project was sourced from the nflverse GitHub account and the website of Football Outsiders. The nflverse GitHub was created by data scientists and analysists to create an easily accessible, open-source package to analyze NFL data. The data used from this package contains individual game information. The Football Outsiders’ website contains many advanced statistics that can be used to determine the strength of a team, such as DVOA (Defense-adjusted Value Over Average). With the game data from nflverse and the DVOA data from Football Outsiders, I created a dataset that will be used to predict the winner of any given game. To complete the final data set, I performed some variable engineering in order to prepare it for analysis. The final data set consists of 2,956 rows and 50 columns. There are no missing values in the data set and each row represents an NFL game that has taken place during the 2011-21 seasons.

Note: In the NFL, though uncommon, games can end in a tie. In the dataset, 10 of the 2,956 games (roughly 0.3%) ended in a tie. For the purposes of predicting a winner and loser for betting purposes, the categorical variable *tie* will be used to remove these observations from the data set. However, these observations will be carefully examined as the project progresses as the data they present could be useful in predicting whether the home team wins or not.

Variable Description

The variable descriptions for the final dataset are displayed below:

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| game\_id | Categorical | Unique identifier for an NFL game |
| season | Categorical | The season when the game was played |
| game\_type | Categorical | The type of game that was played Ex - (regular, wildcard, super bowl, etc) |
| week | Categorical | The week of the season when the game was played |
| gameday | Date | The day the game was played |
| weekday | Categorical | The day of the week the game was played |
| gametime | Datetime | The time of when the game started |
| gametime\_hour | Categorical | The hour the game was played |
| gametime\_minute | Categorical | The minute of the hour that the game was played |
| away\_team | Categorical | The abbreviation of the away team |
| away\_score | Numerical | The score of the away team |
| home\_team | Categorical | The abbreviation of the home team |
| home\_score | Numerical | The score of the home team |
| location | Categorical | Whether the game was played at the home team’s stadium or a neutral site |
| result | Numerical | The away\_score subtracted from the home\_score |
| total | Numerical | The total points scored in the game |
| overtime | Categorical | Whether the game went into over time or not |
| pfr | Categorical | The ProFootballReference game ID |
| espn | Categorical | The ESPN game ID |
| away\_rest | Categorical | The week of the away team’s bye week |
| home\_rest | Categorical | The week of the home team’s bye week |
| away\_moneyline | Numerical | The moneyline for the away team |
| home\_moneyline | Numerical | The moneyline for the home team |
| spread\_line | Numerical | The spread generated by sportsbooks for the game |
| away\_spread\_odds | Numerical | The spread odds for the away team |
| home\_spread\_odds | Numerical | The spread odds for the home team |
| total\_line | Numerical | The total point line for over/under betting |
| under\_odds | Numerical | The odds for the under |
| over\_odds | Numerical | The odds for the over |
| div\_game | Categorical | Whether the game is played against a divisional opponent |
| roof | Categorical | The type of roof (if any) the stadium has |
| surface | Categorical | They type of grass/turf the stadium has |
| temp | Numerical | The temperature in fahrenheit the game was played at |
| wind | Numerical | The wind in mph speed during the game |
| away\_qb\_id | Categorical | Unique identifier of the away quarterback |
| home\_qb\_id | Categorical | Unique identifier of the home quarterback |
| referee | Categorical | The referee of the game |
| stadium\_id | Categorical | Unique identifier for the stadium |
| stadium | Categorical | The name of the stadium |
| outdoor | Categorical | Whether the game was played in a stadium without a roof |
| grass | Categorical | Whether the game was played on a grass field |
| playoff | Categorical | Whether or not the game was played during the playoffs |
| home\_DVOA\_Rank | Categorical | The home team’s DVOA ranking |
| home\_DVOA | Numerical | The home team’s DVOA score |
| away\_DVOA\_Rank | Categorical | The away team’s DVOA ranking |
| away\_DVOA | Numerical | The home team’s DVOA score |
| home\_win | Categorical | Whether or not the home team won the game |
| home\_afterbye | Categorical | Whether or not the home team plays the week after their bye week |
| away\_afterbye | Categorical | Whether or not the away team plays the week after their bye week |
| tie | Categorical | Whether the game ended in a tie or not |

**Exploratory Analysis**

Chart, histogram

Description automatically generated*Figure 1:*

The histogram shows the season for all home wins. The distribution is relatively even. The 2013 season had the larges number of home wins with nearly 160. While the 2020 season had the lowest number of home wins with around 135 wins.

*Chart, histogram

Description automatically generatedFigure 2:*

The histogram shows the season for all home wins. The distribution peaks in week 2 with over 100 and drops off near week 4 before going over 100 again in weeks 14, 16, and 17. The drop off mid-season is likely related to the bye weeks that occur. The wins then decrease dramatically as the playoffs begin and teams are eliminated.

*Chart, histogram

Description automatically generated*

*Figure 3:*

The histogram shows the number of home wins for each home rank of DVOA. The distribution is right skewed. Teams that are ranked higher in DVOA win more home games. Teams ranking 1st and 5th in DVOA won nearly 85 games in this time frame.

*Chart, histogram

Description automatically generatedFigure 4:*

The histogram shows the number of home wins for each away rank of DVOA. The distribution is left skewed with more home wins occurring when away teams have low ranks in DVOA. Home teams won the most (80 games) during the 2011-21 seasons when playing the 31st ranked team in DVOA.

*Chart, histogram

Description automatically generated*

*Figure 5:*

The histogram shows the count of home wins based on the spread line for the game. Most home wins occur when the spread is positive. The peak at over 550 wins during the time frame happens when the spread is between 3 and 7.

*Chart, histogram

Description automatically generatedFigure 6:*

The histogram shows the count of home wins based on the over/under total line for the game. The distribution is normal, but lightly right skewed. The peak of roughly 310 home wins occurred when the over/under total line was between 44 and 46.

*Chart

Description automatically generatedFigure 7:*

The histogram shows the count of home wins based on the start time of the game. The interesting takeaway from this chart is how close the 20 and 16 bucket are. There are typically more games starting at hour 16 (4-6 per week) than hour 20 (2-3 per week). Being separated by ~100 leads me to believe that home teams win more for late games.

*Chart, histogram

Description automatically generatedFigure 8:*

The histogram shows the count of home wins based on the bye week of the home team. Week 7 is the most common bye week in the NFL (over 26% of home teams in the data set have their bye week on week 7) so it is expected to be high. There are also small spikes in week 4, 10, and 14. Week 4 (2%) and 14 (2%) are uncommon bye weeks. That leads me to believe that it might be better to have early or late season bye weeks.

**Modeling**

Following completing the initial exploratory analysis, I now have a much better understanding of the data and how I will move forward with finding a model to predict the winner of a football game. In the next stage of the project, I will build and evaluate multiple classification models that predict the likelihood of the home team winning the game. Prior to building the models, I will perform variable selection to find the 10 most important variables from Random Forest, AdaBoost, and Gradient Boosting classification models. This step is vital to finding which variables are important for determining the winner of a football game. Following finding the important variables, I will build multiple classification models that I have learned throughout this course. These models include Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, AdaBoost, and Gradient Boosting. I will also consider hyper-parameter tuning where it is necessary. The hyper-parameters that will be considered are decision tree depth, number of estimators, learning rates, and kernels. I will also consider different cut off values for making predictions. During the model evaluation stage, I will consider the accuracy, recall, and F1 scores for choosing the best model. The ultimate goal of this project is to find a classification model that can accurately predict whether the home team will win the football game. By following the steps I have laid out, I will be able to determine which model is best and the parameters that can be tuned for the best performance.

Variable Importance

For the win prediction model, I evaluated variable importance through building three classification models and extracting their most important variables. Below are the steps I performed to find and select the most important variables from the data set.

|  |
| --- |
| 1. Remove variables that are not helpful for predictions (identifiers, final results, etc.) 2. For 100 iterations:    1. Split the data into train and test sets    2. Build a Random Forest model and extract the variable importance scores    3. Build an AdaBoost model and extract the variable importance scores    4. Build a Gradient Boosting model and extract the variable importance scores 3. Find the average importance score for all of the variables 4. Return a final data set “games\_Final.csv” with the top 10 most important variables from the previous steps |

Generating and Evaluating Models

The steps for generating and evaluating each model are shown below.

|  |
| --- |
| 1. Read the “games\_Final.csv” file into a pandas data frame 2. Split the data into input (10 most important variables) and output (home\_win) 3. Define data frames with all hyper-parameter combinations for each model type 4. For 100 iterations:    1. Split the data into train and test sets    2. For all hyper parameter combinations       1. Build a Logistic Regression model with the training data set and evaluate it on the test set for all cut off values. Then append the cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score).       2. Build a Support Vector Machine model with the training data set and evaluate it on the test set for all cut off values. Then append the kernel, cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score).       3. Build a Decision Tree model with the training data set and evaluate it on the test set for all cut off values. Then append the depth, cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score).       4. Build a Random Forest model with the training data set and evaluate it on the test set for all cut off values. Then append the depth, number of trees, cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score).       5. Build an AdaBoost model with the training data set and evaluate it on the test set for all cut off values. Then append the depth, number of trees, learning rate, cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score).       6. Build a Gradient Boosting model with the training data set and evaluate it on the test set for all cut off values. Then append the depth, number of trees, learning rate, cutoff value, accuracy, recall, F1 score, and an overall score (average of accuracy, recall, and F1 score). 5. Export a csv file for each model type’s result 6. Compute average performance scores for each possible hyper-parameter combination for each model 7. Identify the best model of each model type |

Results

By following the steps above, I was able to use the training data to predict on the testing set of NFL games. From these predictions, I was able to extract the best model from each model type. The best models for Logistic Regression, Support Vector Machines, Decision Trees, Random Forest, AdaBoost, and Gradient Boosting models are displayed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Type** | **Accuracy** | **Recall** | **F1 Score** | **Overall** |
| Logistic Regression | 0.676 | 0.903 | 0.757 | 0.779 |
| Support Vector Machine | 0.666 | 0.910 | 0.753 | 0.776 |
| Random Forest | 0.628 | 0.959 | 0.742 | 0.776 |
| Gradient Boosting | 0.642 | 0.940 | 0.746 | 0.776 |
| AdaBoost | 0.623 | 0.961 | 0.741 | 0.775 |
| Decision Tree | 0.618 | 0.964 | 0.738 | 0.773 |

To further explore the results, we can see how cut-off values played a significant part in selecting the best model. The figures below demonstrate how vital selecting the right cut-off values impacts a model’s results.

*Figure 9: Plot of accuracy for top models and their cut off values*

Chart, line chart

Description automatically generated

*Figure 10: Plot of recall for top models and their cut off values*

Chart, line chart

Description automatically generated

*Figure 11: Plot of F1 scores for top models and their cut off values*

Chart, line chart

Description automatically generated

*Figure 12: Plot of overall scores for top models and their cut off values*

Chart, line chart

Description automatically generated

From the table and figures above, it appears that the best model for predicting the winner of an NFL game is the Logistic Regression model with the predictor variables: away\_moneyline, home\_moneyline, spread\_line, home\_spread\_odds, total\_line, temp, home\_DVOA\_Rank, home\_DVOA, away\_DVOA\_Rank, and away\_DVOA. Since it was a logistic regression model, there were no hyper-parameters to be tuned, only the cut off value. The ideal cut off value for this model was 0.35.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Type** | **Cut Off** | **Accuracy** | **Recall** | **F1 Score** | **Overall** |
| Logistic Regression | 0.35 | 0.676 | 0.903 | 0.757 | 0.779 |

This model had the highest accuracy and F1 scores of all the models that were considered. This model had the highest accuracy (67.6%) which means that it was the best model for correctly classifying the winner of football games. This model also had the highest F1 score, meaning that it was the best at its ability to both capture positive cases and be accurate with the cases it does capture. While the recall (90.3%) was the lowest of the top performing models, it was still very high. This tells us that it was very good at identifying when the home team is going to win the game. Ultimately, this model offers the best blend of accuracy and recall allowing us to make accurate predictions on the winner of an NFL game.

**Conclusion**

As sporting events, like NFL games, have taken the forefront in American entertainment, more viewers are wanting to get involved in the action. With the legalization of sports gambling in many states, fans are now able to interact with their favorite pastimes. The aim of this project was to find a method for determining whether the home team will win the football game for sports gambling purposes. This way, we know which team is favored to win and can place our bets appropriately. Through variable engineering, I added and created several variables to assist in predicting the winner of a football game (for example, home\_afterbye tells us the home team is coming off their bye week). Through visualization, I was able to see which variables had strong relationships with the home team winning (for example, refer to Figure 3’s distribution of home wins for each DVOA ranking). Through the modeling phase, multiple classification models with many hyper-parameter combinations were built and evaluated based on their accuracy, recall, and F1 scores. The worst model type was the Decision Tree model. Its best model (cut off of 0.25 and max depth of 3) had the lowest accuracy (61.8%) and lowest F1 score of 0.738. However, it had the highest recall (0.964) of all of the models. This tells me that it struggled to identify when the home team lost and probably has a high rate of false positives. The best performing model was the Logistic Regression model with a cut off value of 0.35. This model had an accuracy of 67.6% and an F1 score of 0.757. Its F1 score tells us that it has a good balance of precision and recall. Using this model, we are more likely to predict the winner of a football game accurately than before.

**Areas of Further Study**

While a model with 67.6% accuracy gives us an advantage for predicting the winner of a football game, there is still a lot of room for improvement. I think there are more variables to be found to increase the prediction power of these models that weren’t easily accessible for historical data. For example, a new statistic taking the NFL by storm is EPA (Estimated Points Added) per play. This variable would give us greater insight into team efficiency and could give us further insight into which team is likely to win the game. With analytics just starting to become more popular in football over recent years, more of these key variables are being created. Unfortunately, it is difficult to find these new analytical figures for data from past seasons. Ultimately, finding a way to find a way to access and incorporate statistics that describe a team’s strengths and weaknesses would be key to improving the prediction power of the model.

Another idea that I would like to explore further would be incorporating multiple high performing models into an ensemble. This way we could take key features from each model and hopefully find a model that outperforms the models that were evaluated over the course of this project. I would consider weighting the results of each model so that the more accurate ones have more sway in the ensemble prediction.

**References**

NFL Game Data available at: [nflverse (github.com)](https://github.com/nflverse)

DVOA Data available at: [Team DVOA Ratings: Overall | Football Outsiders](https://www.footballoutsiders.com/stats/nfl/team-efficiency/2022/regular)

Grubman Gambling article: [Ratings Rise With Expansion Of Legal Sports Betting (forbes.com)](https://www.forbes.com/sites/jacobgrubman/2022/02/11/nfl-ratings-rise-with-expansion-of-legal-sports-betting/?sh=152293245bc5)