Betting the NFL Over/Under — A Data Science Perspective

I. Introduction

Professional football is one of the most popular entertainment options in the United States, with 16.7 million spectators attending National Football League (NFL) games during the 2019 regular season and an additional 180 million viewers watching on television. Owing to this popularity and spurred on by the emergence of fantasy sports competition, sportsbooks and other sports betting sites have experienced lucrative growth in their bookmaker business, taking advantage of the fact that reportedly more than 38 million Americans planned to bet on NFL games during the 2019 season, according to the American Gaming Association.

Among the various betting options for NFL games offered by sportsbook outfits, the over/under (O/U) line is one of the most popular. The O/U line represents the predicted total score of the contest, taking into account all scoring including scoring in any overtime period. The bettor chooses to bet on the over or the under, meaning that the conjecture is that the final total score will be greater or less than the O/U line, respectively. If the selection of the bettor is correct, the bettor wins the amount of the bet placed; if not, the bettor loses the bet. In addition, the bookmaker charges a vigorish (usually 10%) for placing the bet. With this in mind, this study seeks to determine whether a system that uses regression analysis of historical football team statistics to predict the total score of a contest can be profitable, by defeating sportsbook O/U odds on a sufficiently regular basis.

In this study, team statistics are utilized as well as game data to obtain the target feature — total points scored. The team statistics include attributes for each participating team in the contest, such as total yards, average points yielded, and touchdown scoring efficiency, characterizing the abilities of the offensive, defensive, and special teams components of the squad over the previous five games of the respective season. In addition, event data such as the field type and temperature at kickoff is also included. Contests over a span of ten seasons are utilized in the analysis, with labels used for training and testing taken beginning with week six of each season and extending to the end of the respective regular season. These data are obtained from https://www.pro-football-reference.com/, which provides historical data from decades of NFL contests. Through the application of conventional statistical metrics, an optimal model is selected. Betting lines from the contests used in this analysis are also obtained from Pro-Reference-Football.com.

Despite the inclusion of an extensive amount of information regarding recent team history, this study contains a couple of limitations that may affect the reliability of its results. An important factor in the outcome of a football contest is the realization of injury to the members of the respective teams. Although injuries during a given contest can be regarded as a random factor that contributes to the variance of the result of that contest, injuries that occur during the week before such a given contest or in the most recent contests before can affect the betting line of the given contest and, thus, the reliability of the analysis. This study will not give consideration to such injuries, except as reflected in the statistics of the previous contests considered in the study. Another important factor in the result of a football contest is weather. The temperature of the event is indeed considered; however, the magnitude of such weather, of which the forecast could affect betting lines, is out of the scope of this study.

A regression analysis of historical NFL team data is performed, followed by thresholding on the prediction interval, to determine whether the total score of a football contest can be predicted with sufficiently low error for the model to be considered a profitable betting tool. Stakeholders for this study include anyone interested in evaluating whether a business involving the betting of O/U lines at sportsbooks could be profitable. Interested parties in this study may also include coaches and managers that are interested in what aspects of team performance are most relevant to team scoring in football today. The solution of this study is specifically focused on the effect of team performance on total score and does not directly cover individual player performance, nor does it address point-spread betting or other margin-of-victory considerations.

II. Dataset

Game data from a total of 2559 regular season NFL contests covering the ten-year period of 2010-2019 were scraped from the Pro Football Reference website (http://pro-football-reference.com) using BeautifulSoup. From this dataset, game conditions and various statistics averaged over previous contests were used to predict the total score. The columns depicting averaged statistics include statistics averaged over the previous five games of the respective season to provide a basis for prediction. For this reason, only games from the sixth week to the last week of the regular season are considered for prediction. In addition, play-by-play data from the website are manipulated to produce red zone success features. A description of the features considered in this analysis is provided in Table 1.

Table 1. Description of features used in analysis.

Descriptive Features - Used solely to describe the contest in question

- Home_Team - The name of the home team in the contest
- Vis_Team - The name of the visiting team in the contest
- H_Game - The number of points scored by the home team in the contest
- V_Game - The number of points scored by the visiting team in the contest

Analytical Features - Features used in analysis

- Tot_Pts - Total Points of the contest. **This is our target variable.**

- Over/Under - The predicted total points of the contest from professional bookmakers.

Averaged Features - Statistics averaged over previous five contests. In the columns, prefix "H_" stands for Home Team and "V_" stands for Visiting Team

- Pts - Points scored
- Pts_Opp - Points scored by opponent
- Off_Pass - Offensive passing yards
- Pass_Metric - Average of offensive passing rank (32 - best, 1 - worst) and defensive passing rank of opponent (1 - best, 32 - worst)
- Off_Rush - Offensive rushing yards
- Rush_Metric - Average of offensive rushing rank and defensive rushing rank of opponent
- Def_Pass - Passing yards given up on defensive
- Def_Rush - Rushing yards given up on defensive
- TD - Touchdowns scored
- TD_on_Def - Touchdowns scored by defense
- FG_Pct - Percentage of field goals made vs. field goals attempted
- RZ_Pct - Percentage of red zone possessions resulting in a touchdown (red zone possession is possession reaching opponent's 20 yard line)
- Def_RZ_Pct - Defensive percentage of red zone possessions giving up a touchdown
- Poss - Time of possession
- Plays - Total number of plays
- TO_Gain - Number of turnovers gained by defense
- TO_Lost - Number of turnovers lost by offense
- Yds_Pen - Number of yards penalized
- Sacks_Def - Sacks earned by defense
- Tackles_Loss - Number of tackles for loss earned by defense
- Kickret - Return yards from kickoff
- Puntret - Return yards from punt

Game Conditions - Conditions of the contest in question

- Temperature - Temperature of the contest at kickoff
- surface - The surface on which the contest was played

III. Exploratory Data Analysis

A. Overview

To get a feel for the dataset and how scoring is distributed in professional football, histograms of the scores for home teams and visiting teams are plotted in the same graph (Figure 1a). First, certain values (linear combinations of 7 and 3) clearly show predominance in the graph, displaying the nature of American football. The points scored range from 0 to 62, with a mode of 20, clearly a very common score in American football. Visiting scores (blue) tend to populate the smaller point values, whereas home scores (red) are more prevalent in the higher values. The superiority of home teams in professional football is further demonstrated in the pie graph, with the home team winning just over 57% of the total number of games. This home field advantage may suggest that contests containing home teams with potent offenses might tend to contain higher scoring than those containing visiting teams with similar offenses.

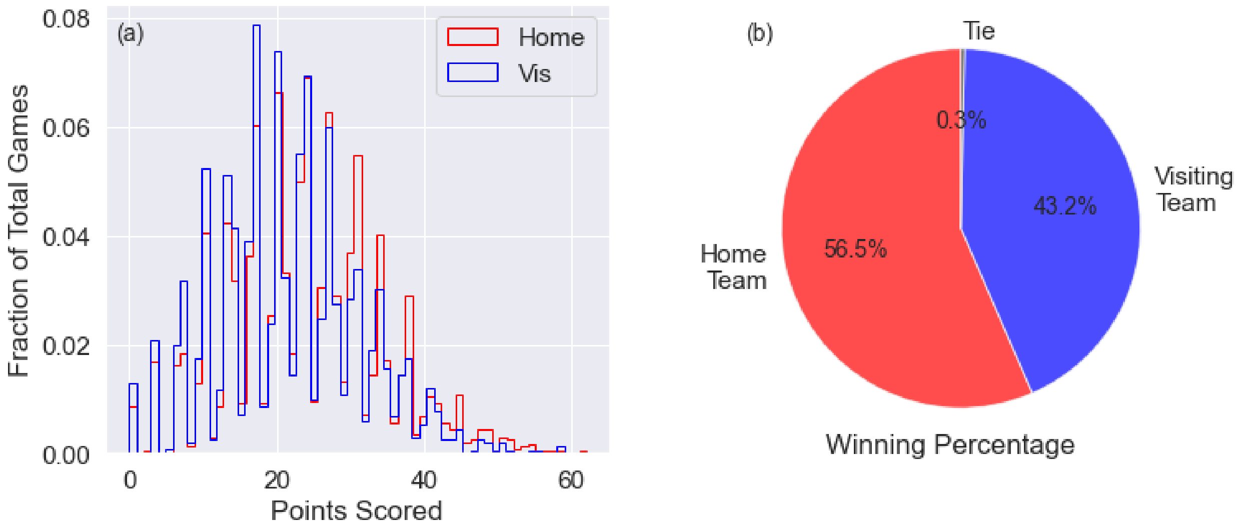

Figure 1. a) Scoring distributions for home and visiting teams. b) Pie chart showing the winning percentages for home and visiting teams.

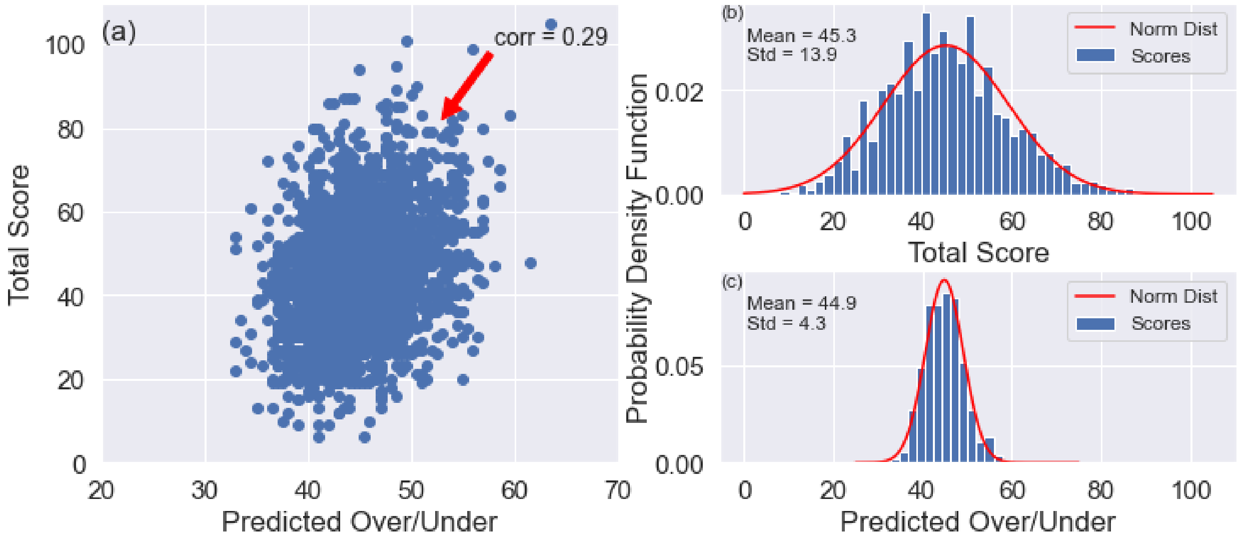

Figure 2. a) Scatterplot of Total Score vs. the sportsbook Over/Under. Probability density function of the b) total score and c) over/under, with the respective mean and standard deviation presented in the upper left corner.

In terms of the overall total points of contests, there is a clear correlation with the predicted Over/Under (O/U) total but with a much larger range, as shown in Figure 2a. The total score distribution approximates a normal distribution with some deviation, while the O/U follows the normal distribution more closely (Figure 2b, c). This is confirmed by the quantile/quantile plots of the two quantities (Figure 3) for which there are deviations from the normal distribution, represented by the red line, only for the most extreme values.

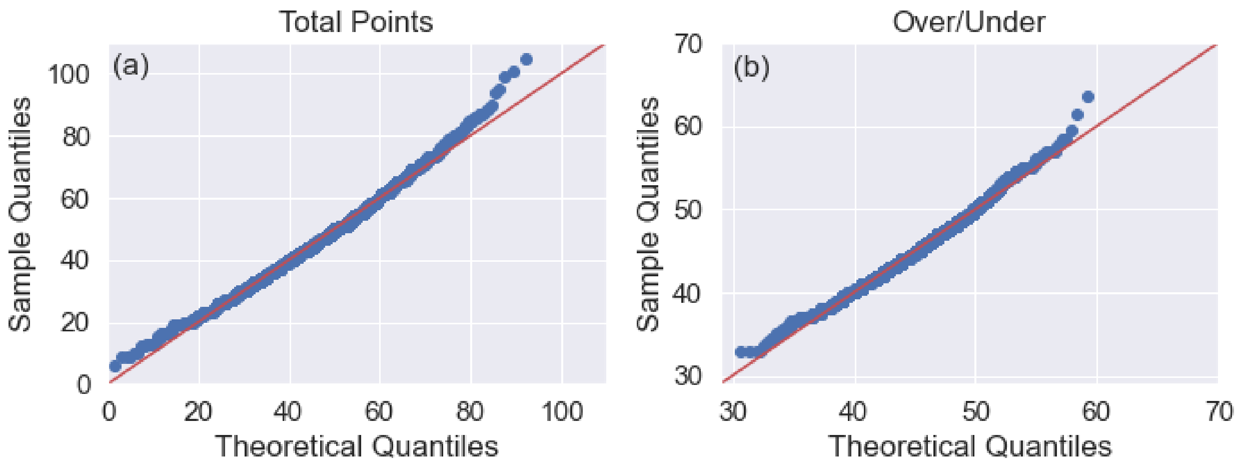

Figure 3. Quantile-Quantile plots for a) Total Points and b) Over/Under, with the red lines representing the normal distribution of the respective quantities.

The Over/Under error, defined as the difference between the O/U and the actual total score, approximately follows a normal distribution about zero (Figure 4a) with remarkable symmetry in the Over (median = 9.0, mean = 11.1) and Under (median = -9.0, mean = -10.3) results. The Quantile-Quantile plot (Figure 4b) shows greater deviation from normality to the upside than for the negative extreme values. This is corroborated by the boxplot in Figure 5, which shows that the outliers primarily occur to the upside. However, these extreme data points account for only 3.0% of the total distribution.

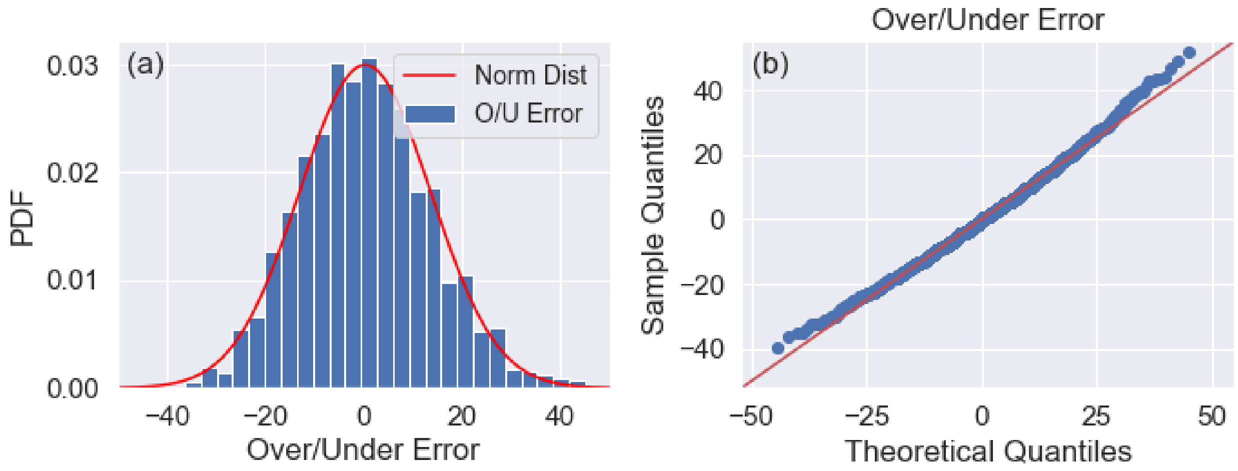

Figure 4. a) Probability density function of the over/under error, representing the difference between the resulting total score and predicted over/under. b) Quantile-Quantile plot for the over/under error.

The over/under error varies primarily between -40 and +40, regardless of the underlying O/U prediction (Figure 6a), demonstrating a near-zero correlation with O/U, with a coefficient of -0.04.

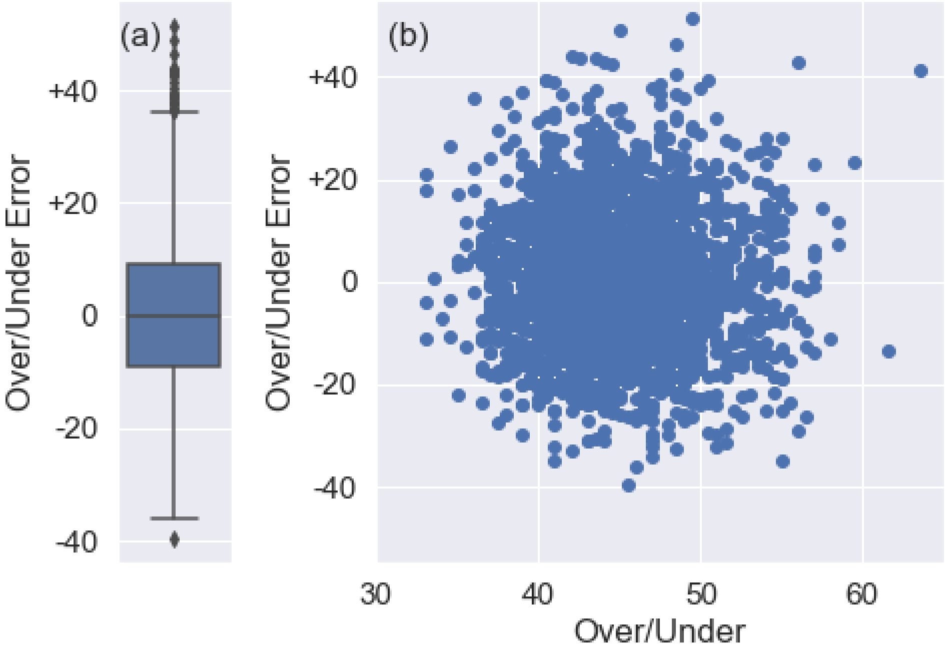

Figure 5. a) Boxplot of the Over/Under error. b) Scatterplot of the Over/Under error vs. the Over/Under.

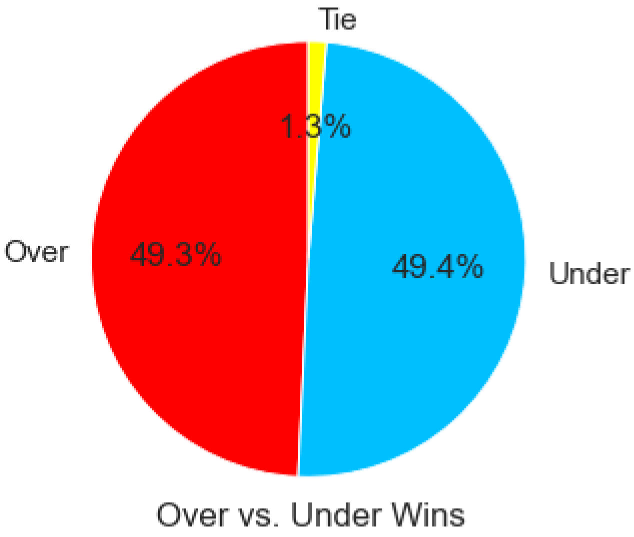

Figure 6. Over vs. Under pie chart from the dataset.

B. Do Week or Game Conditions impact Total Score?

Week and Temperature

Over the ten-year period of the study, the Over exhibits an almost identical winning percentage as the Under (Figure 6b). There is some variance from year to year; however, the more interesting result is how the Over varies on a weekly basis. Figure 7a shows the Over winning percentage vs. the week of the season over the dataset. In weeks 1-10, the Over% shows consistently higher levels than in the last seven weeks of the season. This is even more clearly shown in the boxplot in Figure 7b, in which the interquartile range of weeks 1-10 is significantly higher than the range representing weeks 11-17. A t-test suggests that the null hypothesis H_0 — the Over% of Weeks 1-10 is not statistically higher than the Over% of Weeks 11-17 — can be rejected as t(2360.1) = 2.991, p = 0.0014, although an effect size of 0.12 indicates that a greater sample size would be helpful in providing further confidence in rejecting the null hypothesis. Nevertheless, the preponderance of Over results in the first 10 weeks of the season should be given consideration in betting. The reason for this divergence can be seen in Figure 8, as the average Over/Under per week stays relatively constant throughout the season, not reflecting the clear decline in total points that occurs after Week 10 (Figure 8a). This decrease in Total Points is mirrored by a decline in average passing yards per game late in the season, peaking in Week 9 and followed by a reasonably steady decline throughout the rest of the season. There seems to be a significant relationship between total passing yards in a game and temperature, with the total passing yards peaking near the 60° F temperature (Figure 8b). Figure 9 shows that although there is a wide range in total passing yards for each decade of temperature, a characteristic of the wide-ranging dataset, a clear fall-off in the interquartile range of passing yards is

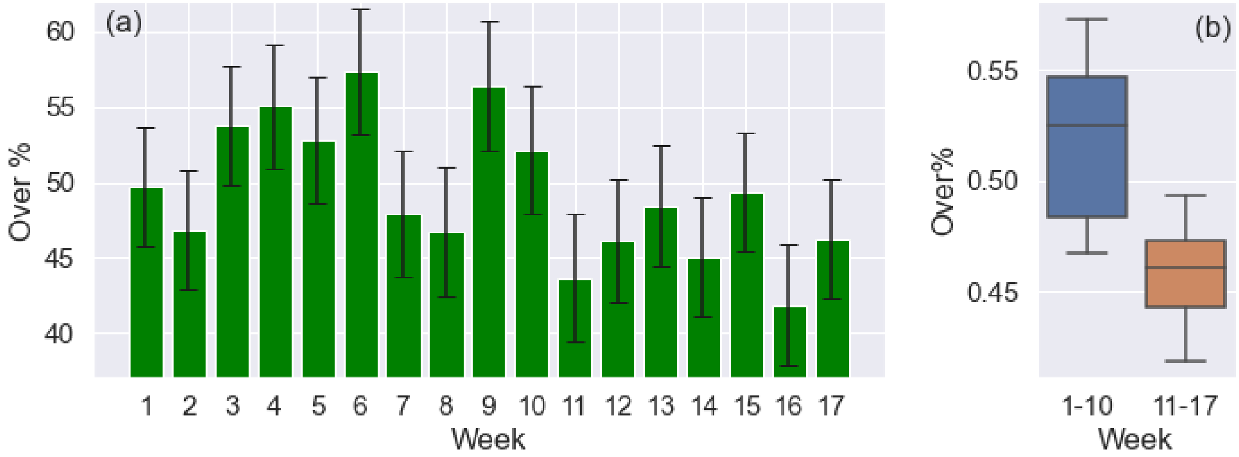

Figure 7. a) Bar chart displaying the Over winning percentage throughout the football season. b) Boxplots of Over winning percentage representing the first ten weeks and the last seven weeks of the season.

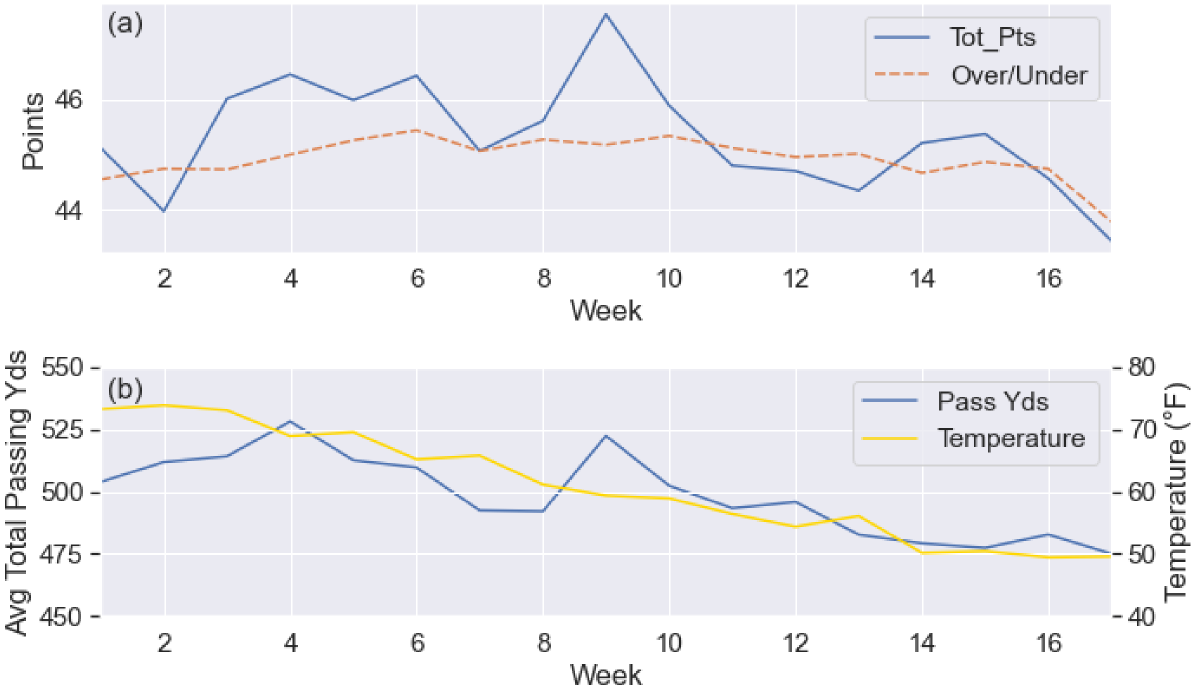

Figure 8. a) Total points and Over/Under vs. Week and b) average total passing yards per game and temperature vs. week during the NFL season.

exhibited in temperature bins lower and higher than the 60-70 °F temperature bin. As the average temperature falls below 60° after Week 9, this appears to be the main factor influencing the drop in total points during this period. This also reflects the increasing importance of passing in the modern game, as scoring and passing yards have increased over the past few seasons.

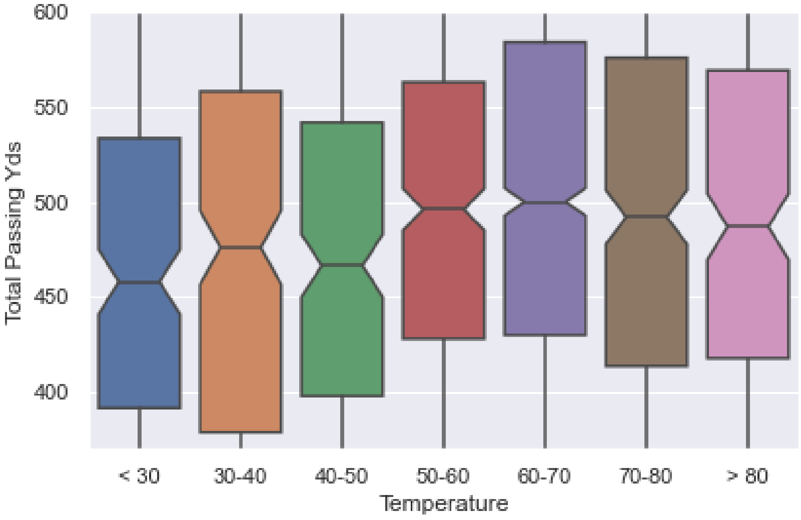

Figure 9. Boxplots of total passing yards vs. decade of temperature.

Other underlying factors responsible for the late season decline in total points may be a greater focus on defense for top teams as they prepare for the playoffs and greater experimentation with new player prospects for poor teams as their playoff chances are extinguished.

Wind

Another type of game conditions that has a significant effect on total score is wind. Of the 32 NFL teams, 8 currently play in domed stadiums under climate-controlled conditions (Atlanta Falcons, Arizona Cardinals, Dallas Cowboys, Detroit Lions, Houston Texans, Indianapolis Colts, Minnesota Vikings, and the New Orleans Saints). Furthermore, during the period of this study, the St. Louis Rams also played home games under domed conditions until their move to Los Angeles in 2016, and the Buffalo Bills played several "home" games in the early part of the study period in Rogers Centre in Toronto and Ford Field in Detroit. Thus, a little more than 26% of the dataset in this study comprises dome games. There is a clear scoring advantage to games played in a dome, where the weather elements are not a factor. In fact, during the period of the study, there was a two-point advantage in total score in dome games over those played outside. Perhaps, the most important factor in the elevated scores of dome contests is the lack of wind present in the stadium. A strong wind makes it difficult to throw the ball accurately or judge the direction of the ball with proper precision, impairing the passing attack. In fact, wind is an important determinative factor in the outcome of total score, as illustrated in Figure 10. Dome games, played under climate-controlled conditions, show the highest average total score at 46.8 points per game. Games that are outside but have low wind conditions less than 5 miles per hour (mph) also demonstrate relatively high scoring at 46.2 points per game. However, these games may provide some of the better betting opportunities, as the outside conditions are typically discounted in the O/U, resulting in an average that is 1.5 points less than the total score and an Over

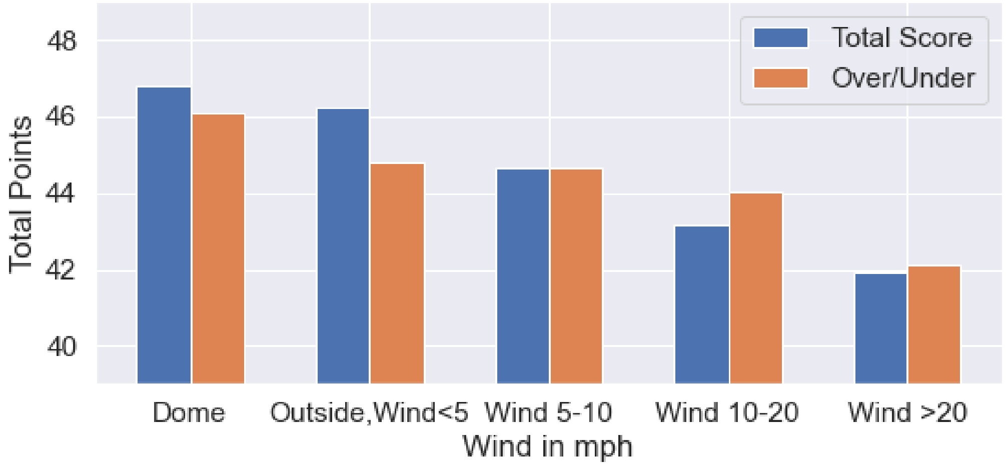

Figure 10. Total score and over/under under various wind conditions.

winning percentage of 52.5% versus 46.0% for the Under. Lower scores are reflected in higher wind conditions, with games with wind greater than 20 mph averaging just under 42 points per game. Interestingly, the Over/Under falls as well but does not seem to fully capture the wind effect, suggesting more betting opportunities. Wind is a feature with one of the largest absolute correlation coefficients in this study, displaying a large negative correlation to total points relative to the other study features.

Summary

- In contests in the first 10 weeks of the season, the Over tends to win more, while the Under is more successful towards the latter part of the season
- Temperature has a significant effect on passing yards, which has a significant impact on total points output. This is not adequately considered in Over/Under prediction
- The presence of wind is also a major factor in total points output
- Dome effect may be given too much consideration in O/U prediction, as there is a smaller difference between dome and low-wind outside games than assumed in O/U prediction

C. Outliers and Leverage

To take a closer look at the data and its impact on determining the relationship between past team performance and productivity forecasting, we consider outliers and high leverage observations and whether they exert outsize influence on the modeling aspect. Applying ordinary least squares regression to the dataset, the base model produces studentized residuals, which are plotted against the corresponding leverage values in Figure 11. The influence plot shows that the observations with greatest influence are those with large residuals, both on the positive and negative sides. In contrast, the extremely high leverage observations do not exert significant influence on the model predictions. Nevertheless, there is not a great division in influence but rather a more or less gradual increase in influence with an increase in absolute residual value. Furthermore, the high influence points are proportionally distributed between positive and negative residual values in a manner similar to the entire dataset and are well distributed among leverage values. An example of a highly influential point is the 2018 meeting between the Los Angeles Rams and the Kansas City Chiefs. Both teams came into the Monday night meeting with potent offenses and were considered to be two of the best teams in the NFL. The game was expected to be high scoring, with the Over/Under placed at 63.5 points, the highest Over/Under in our dataset. The base model responded accordingly with a prediction of 57.3, a score in the 99.5 percentile of base model predictions. The contest was a classic shootout, a 54-51 victory by the

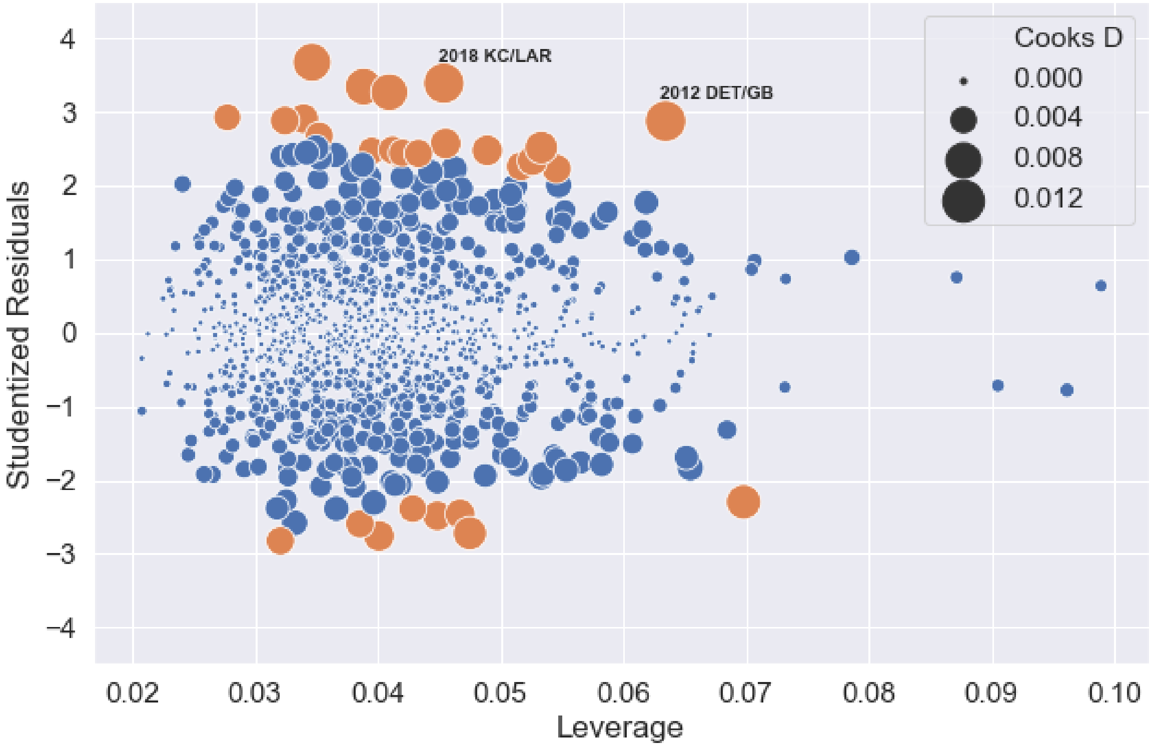

Figure 11. Influence plot with influence represented by the size of the point. Points with Cook's diameters greater than three standard deviations from the mean, representing very high influence, are presented in orange.

Rams, with the 105 total points placing it as the third highest scoring game in NFL history. The resulting large residual is a significant factor in the observation having the second largest Cook's distance. Another influential observation, this one with relatively high leverage, is the 2012 contest between the Green Bay Packers and the Detroit Lions. In this final regular season game of the 2011 season, the Packers, led by eventual NFL MVP Aaron Rodgers, exhibited a potent offense and a strong defense, illustrated by their strong statistics in these areas over the previous five games. Meanwhile, the Lions, on their way to the playoffs, also demonstrated a strong passing attack toward the end of the season. As a result, the primary predictors for total score were far from the mean values of those predictors, leading to elevated statistical leverage for this observation. The Over/Under at 42 points did not reflect the potential offensive output of this meeting, perhaps because backup Matt Flynn was scheduled to start for the Packers in place of Aaron Rodgers. Nevertheless, the base model predicted a total score of 54.5, and the two teams did not disappoint, with the contest resulting in a 45-41 shootout. This high leverage, high residual observation is the most influential of the dataset. However, as Figure 11 shows, most of the high leverage observations in this dataset are not very influential; the influential points comprise the high residual outliers. Whether eliminating some of these influential points is beneficial to the modeling of this problem is explored in the Modeling section.

Summary

- Highly influential points in dataset are primarily outliers.
- High leverage observations are not mainly highly influential.
- Influence gradually increases with residual value

IV. Modeling

A. Metrics Considered in the Study

In this modeling exercise, two metrics are used to guide model evaluation — mean absolute error (MAE) and O/U accuracy.

MAE calculates the average of the absolute value of the error between the predicted value and the resulting score. In this study, this value will generally be relatively high, largely due to the varying nature of the sport, with total scores in the dataset ranging from 6 to 105 points. Nevertheless, it is a goal of this study to minimize this value to provide a more accurate model and enable more accurate decisions. This is the primary metric used to distinguish among the different models.

Another important metric used in this study is O/U accuracy. Presumably, the main purpose of this model is to aid in the decision-making process of choosing between the over and under in betting NFL contests. Thus, it is important to know how accurate a given model is in performing this function. O/U accuracy considers how likely it is that a prediction of a score higher (lower) than the posted O/U would coincide with an actual result that is also higher (lower) than the O/U. This metric is primarily used in threshold tuning.

B. Balancing Training/Test Sets to Ensure Proper Representation

To conduct the modeling and feature selection of this dataset, an important step in this process is to ensure that the training and test sets sufficiently represent the dataset as a whole. The target feature, *Tot_Pts*, is pretty well normally distributed, as shown in Figure 2a but has a much wider range (6 to 105) than the O/U (30 to 63.5) with which it is compared. As such, a randomly selected subset can reveal an inordinate number of Overs or Unders, which may affect the predicting ability of the model. To avoid this possibility, the training and test datasets and/or cross-validation folds are generated in such a way as to ensure balanced subsets of data while maintaining randomness to simulate future as yet unseen data to the greatest extent possible.

To do this, the labels are sorted, with labels of equal value placed in random order in the sorting. Then, each set of N labels is randomly placed in N groups when choosing a training/test set or when dividing a training set into cross-validation folds. Here, N is the greatest common denominator between the multiple sets. For instance, a training ratio of 70% and test ratio of 30% would yield N = 10, where 7 of every 10 sorted observations would be chosen for the training set, while the other 3 would be chosen for the test set. In the event that cross-validation is chosen, the training set would be divided into folds in the same way. For instance, five-fold cross-validation would give N = 5, such that for every five observations from the sorted list, each observation is selected for a different test fold. This ensures that the characteristics of the entire dataset are represented in the data subsets, regardless of the random seed chosen, while maintaining the large degree of randomness necessary to train for and simulate unseen data.

Summary

- Labels undergo a process of sorting and random selection to ensure proper balance and sufficient randomness.
- Routine is performed for train-test splitting and determination of cross-validation folds

C. Feature Engineering

In determining the features to use in modeling the NFL dataset, a necessary consideration is the correlation among the features. Figure 12 shows a correlation heatmap that includes the 15 features with the greatest absolute correlation coefficient values in relation to the target variable *Tot_Pts*. Although *Over/Under* is not included in the modeling, its corresponding correlation coefficients are shown here for the purpose of analysis. An examination of this heatmap reveals some interesting points. First, the features, as a whole, are not as correlated to *Tot_Pts* as *Over/Under*. This is not terribly surprising since, as shown above, *Tot_Pts* has a much greater spread than *Over/Under*, and, clearly, randomness plays a much greater role in the total score. *Tot_Pts* seems to be more or less equally correlated with *V_Off_Pass*, *H_Off_Pass*, *H_Pts*, *V_Pts*, *H_TD*, and *V_TD*. However, the home and away *TD* features are shown to be highly correlated with the corresponding *Pts* features (R = 0.95). Therefore, to avoid multicollinearity, *H_TD* and *V_TD* are dropped from the feature analysis. Despite the relatively low individual correlations demonstrated for *Tot_Pts*, there does seem to be some opportunity for optimization of the prediction of *Tot_Pts* against *Over/Under*. *V_Def_RZ_Pct* shows a similar correlation coefficient with respect to *Tot_Pts* as those of the categories mentioned above, whereas the correlation with *Over/Under* is significantly reduced in comparison with those of the previously mentioned categories.

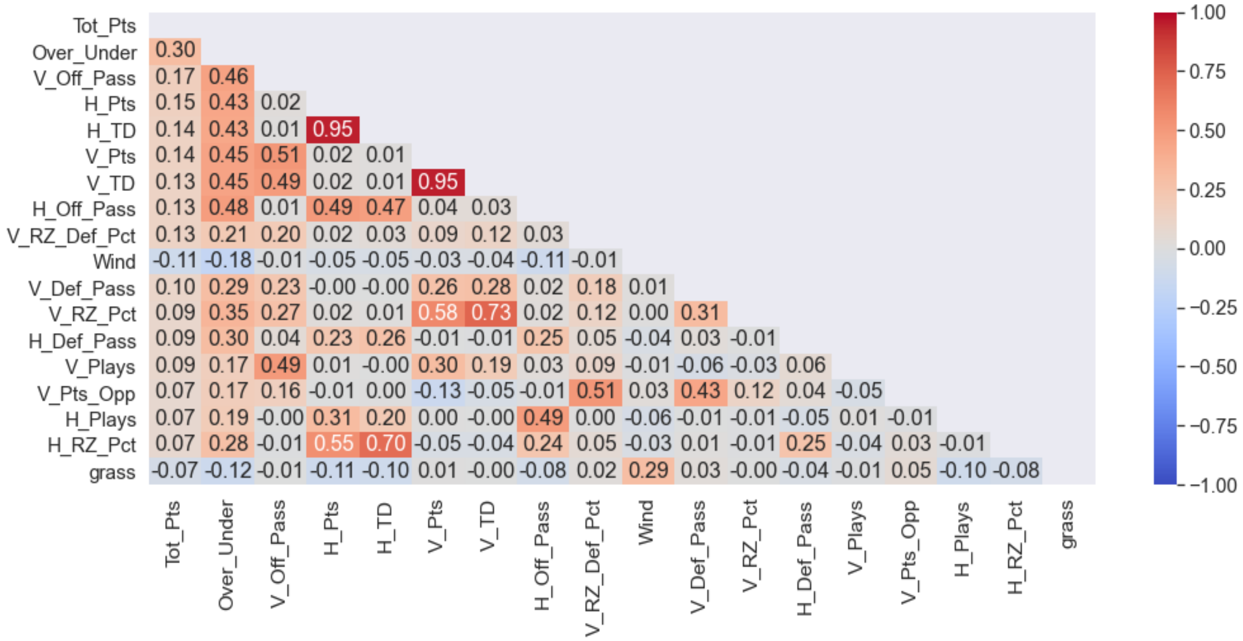

Figure 12. Correlation heatmap for the features with the largest absolute correlation magnitudes with respect to *Tot_Pts*.

This suggests that the defensive red-zone percentage may be more significant in the eventual total score than what is considered in the *Over/Under* determination. Interestingly, *H_Def_RZ_Pct*, the feature representing home team defensive red-zone percentage, is not even within the top 17 features in terms of correlation with *Tot_Pts*. Overall, the correlation results suggest that the accurate prediction of the total score in an American football contest may be determined by the combination of many factors; however, those most determinative are related to scoring, passing proficiency, ability to defend the pass, defensive red-zone percentage for visiting teams, and wind.

Along with the traditional passing, rushing, and special teams statistics, as well as the data describing game conditions, two additional features were created to characterize the matchup of a particular contest. *_Pass_Metric* and *_Rush_Metric* are features engineered to consider the effects of matching the offense of one of the participants with the defense of the other and vice versa. To generate these features, the *Off_Pass*, *Off_Rush*, *Def_Pass,* and *Def_Rush* statistics, representing the average of the respective stats over the previous five games, are ranked for all the teams for a given week, from 1 to 32, where 32 represents the highest average value for the respective statistic. Thus, for example, a high *H_Rush_Metric* for a game would mean that the home team has averaged a relatively high number of rushing yards over the past five games and is facing an opponent that has given up a high number of rushing yards over the same timespan. Because of the significant correlation between *_Off_(Pass,Rush)* and *_(Pass,Rush)_Metric* features, some multicollinearity is inevitable, necessitating the elimination of the *Metric* features when determining feature importance. However, these features, whose combining aspects provide information that no other features do, can be quite useful in the final modeling.

Summary

- Features most correlated with Total Points are associated with scoring, passing offense and defense, visiting defensive red-zone proficiency, and wind
- Visiting defensive red-zone percentage has a much stronger relationship with total points in relation to other features than with Over/Under
- *Wind* is the only significant feature to have a negative correlation with Total Points
- Rushing features do not have a relatively high correlation with Total Points.

D. Scaling and Feature Importance

Regression Analysis of Feature Importance Using VIF

To gain a more complete understanding of what drives the determination of total score, the most important features and their relationship to that target variable are identified. To perform this task, we utilize the variance inflation factor (VIF) to eliminate features that are significant sources of multicollinearity in the feature set. As mentioned

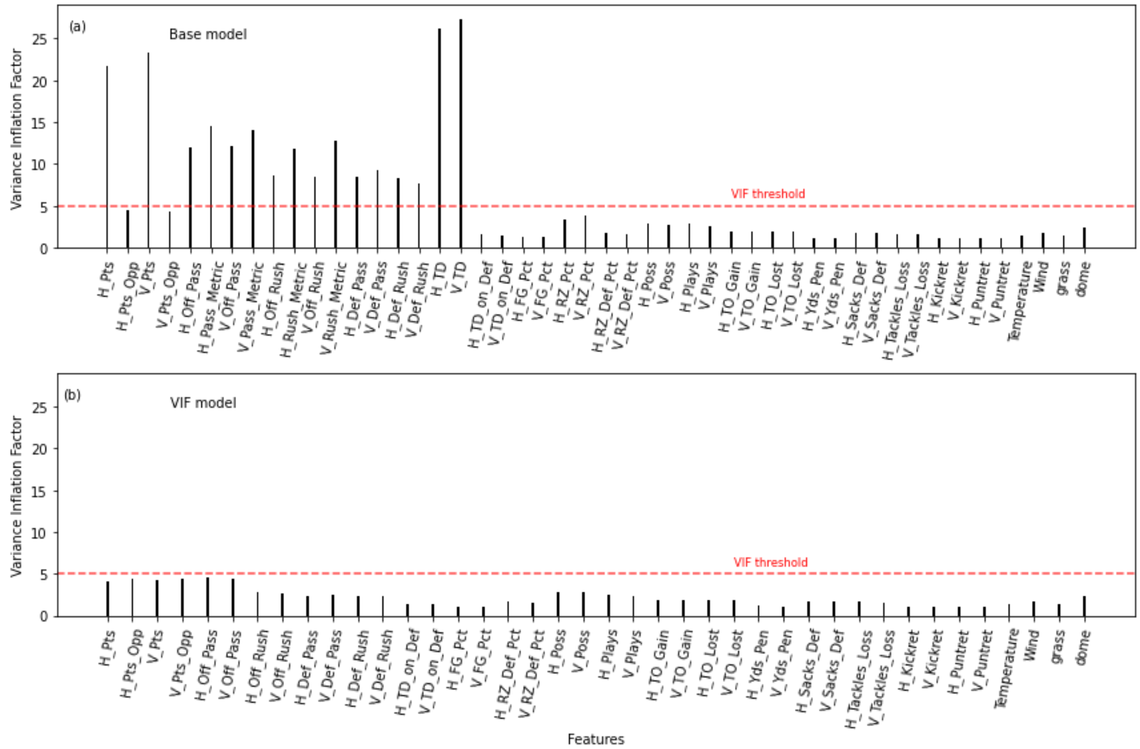

Figure 13. Variance inflation factor for the features in the a) base model and b) model with the *_TD*, *_Pass_Metric*, *_Rush_Metric*, and *_RZ_Pct* features removed, labeled as the VIF model.

previously, the *_TD* features exhibit high correlation with *Tot_Pts*. Figure 13a shows the VIF for the features of the base model, with the *_TD* features clearly demonstrating the highest values. Apart from the *_Pts* features, which are quite essential in the total points determination, the *_Pass_Metric* and *_Rush_Metric* features also show high VIF values. A good rule of thumb for maintaining multicollinearity under control is for the VIF values of the model features to be below five. This is achieved by removing the *_TD*, *_Pass_Metric*, *_Rush_Metric*, and *_RZ_Pct* features, denoted as the VIF model as shown in Figure 13b. This set of features gives confidence that the corresponding derived coefficients are well-representative, without significant multicollinearity.

Considering these features, coefficients and their respective confidence intervals are determined by ordinary least squares linear regression. In this analysis, the data are scaled using RobustScaler, a scaling algorithm from Scikit-Learn that is robust to outliers. Coefficients with absolute values greater than zero at the 95% confidence level are shown in Table 2, with the scaled values and corresponding error bars plotted in Figure 14.

Upon first glance at the table, one can see a variety of features with similar magnitudes. In fact, there is not one particular feature that stands out from the rest. This is a reflection of both the low correlation coefficients of the features with respect to Total Points and the similarities in correlation coefficients among the different features, as illustrated in the correlation heatmap in Figure 12.

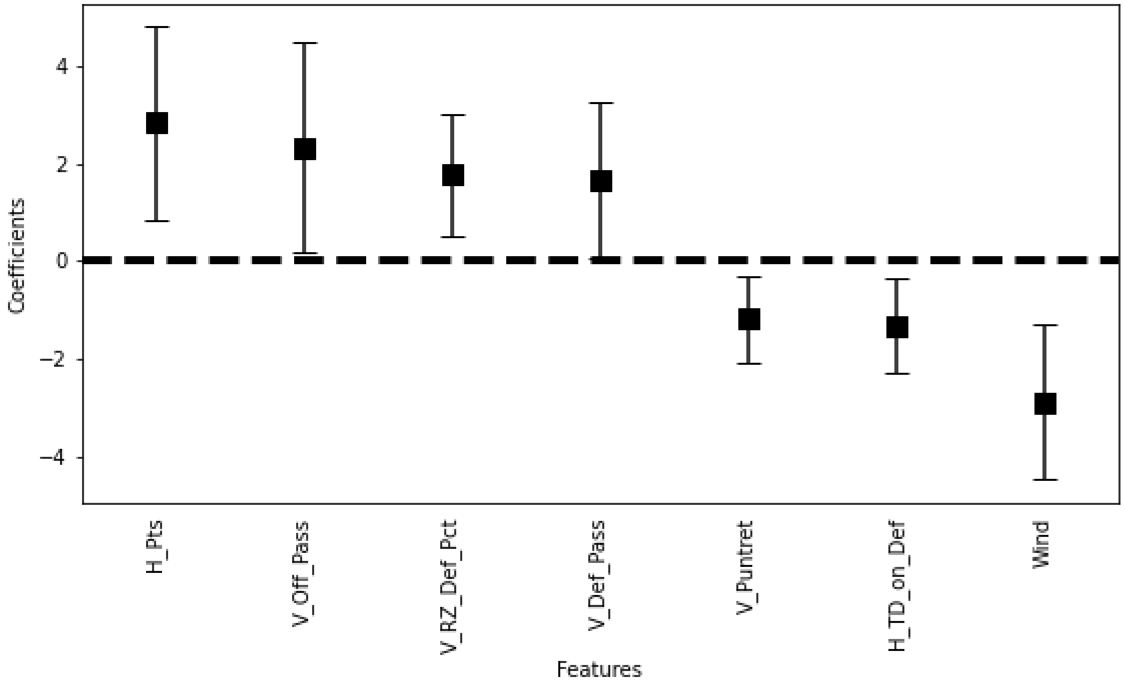

Figure 14. Scaled coefficients of the most important features in the VIF model.

Table 2. Most important features as obtained by the Ordinary Least Squares linear regression model, sorted by absolute value of scaled coefficients.

Features	Unscaled	Scaled	p-coef
Wind	-0.2881	-2.881	0.000
H_Pts	0.3850	2.830	0.005
V_Off_Pass	0.0377	2.329	0.034
V_RZ_Def_Pct	10.8855	1.767	0.006
V_Def_Pass	0.0317	1.643	0.043
H_TD_on_Def	-6.5788	-1.316	0.008
V_Puntret	-0.2501	-1.175	0.009

The number of features with p-values less than 0.05 is 7, with scaled coefficients, an indication of feature importance, ranging from 1 to approximately 3. This suggests that a successful total points predictor model will have to include a variety of features. Furthermore, the maximum potential accuracy of such a model may be limited, owing to the scattered nature of the NFL data. Nevertheless, note that in an O/U betting scheme, a prediction accuracy of only just over 52% is necessary to yield profitability.

Another notable characteristic of the feature set shown in Table 2 is how the feature type is distributed. Besides Wind, three of the top four features are visitor-related. This is despite the fact that home teams have a significantly greater winning percentage, as mentioned above, and a greater than two-point advantage (23.6 vs 21.5) over visiting teams in average score.

Analyzing Feature Relationships

To make sense of these feature regression coefficients and the role of the features in predicting *Tot Pts*, an analysis of some of the feature interrelationships is conducted.

1) *_Off_Pass* and *_Pts*

The *_Off_Pass and _Pts* features are some of the most important features in predicting *Tot_Pts*. *_Off_Pass* has a ~0.5 correlation with *_Pts* (Figure 12); thus, the two sets of features deliver much of the same information. *H_Pts* has a larger correlation with *Tot_Pts* than *H_Off_Pass*, while *V_Off_Pass* has a larger correlation than *V_Pts*. However, it seems that most of the relevant information for the home team is provided by the previous scoring performances of the team, contained in *H_Pts*, whereas there are more relevant features on the visiting side.

2) *V_Off_Pass* and *H_Off_Pass*

Figure 15 shows scatterplots of both *V_Off_Pass* and *H_Off_Pass*. Though both features demonstrate a circular blob-like distribution, the *V_Off_Pass* distribution demonstrates a positive tilt, with positive slopes for the relationship between *Tot_Pts* and *V_Off_Pass* clearly evident. In contrast, the regression slopes for *H_Off_Pass* are not significantly positive. Digging deeper, the data indicate a 16% greater correlation between *V_Off_Pass* and *V_Game* than that between *H_Off_Pass* and *H_Game*, leading to a 26% increase in correlation with *Tot_Pts* for *V_Off_Pass* over *H_Off_Pass*.

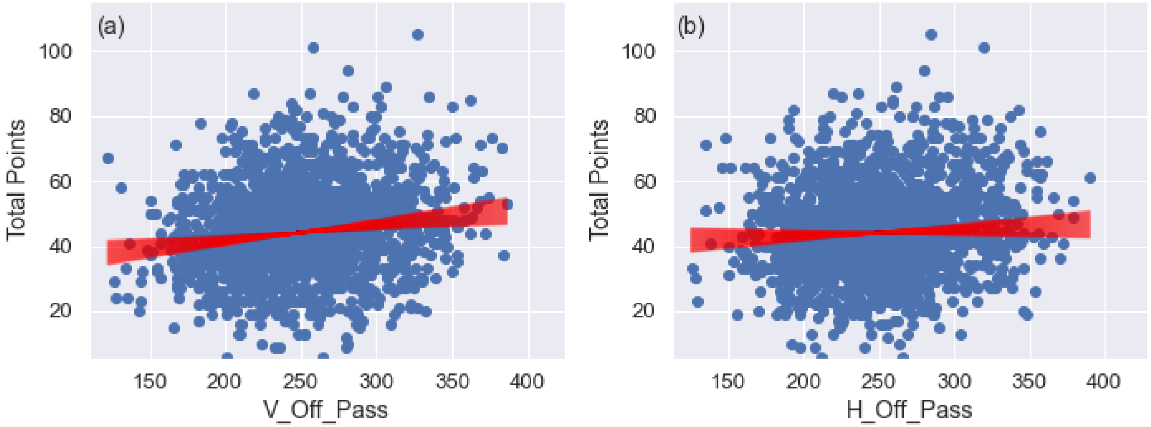

Figure 15. Scatterplots of a) *V_Off_Pass* and b) *H_Off_Pass* with the ranges of regression coefficients displayed in red.

3) *_Off_Pass* and Total Points

Figure 16 shows the distributions of *V_Off_Pass* and *H_Off_Pass* with respect to Total Points superimposed onto each other, focused on the two upper quadrants. For passing yards greater than the median (~250 yards), there is a more or less even distribution between home and visiting scores among the higher scoring contests. For instance, among contests in which the total points exceeded 60 points, 59 of the contests involved home teams that averaged more than 300 passing yards in their previous five games (*H_Off_Pass*) and 53 involved visiting teams with more than 300 passing yards averaged over the same period (*V_Off_Pass*). However, for *H_Off_Pass* less than 200 yards, there is a significantly larger number of games with high total scores than for *V_Off_Pass* less than 200 yards. For example, among the contests in which the Total Points were greater than 60, 24 involved *H_Off_Pass* less than 200 yards, whereas only 13 games had *V_Off_Pass* less than 200 yards. This pattern continues throughout the first quadrant set of values. The fact that high point totals are more distributed among different levels of offensive passing productivity of the home

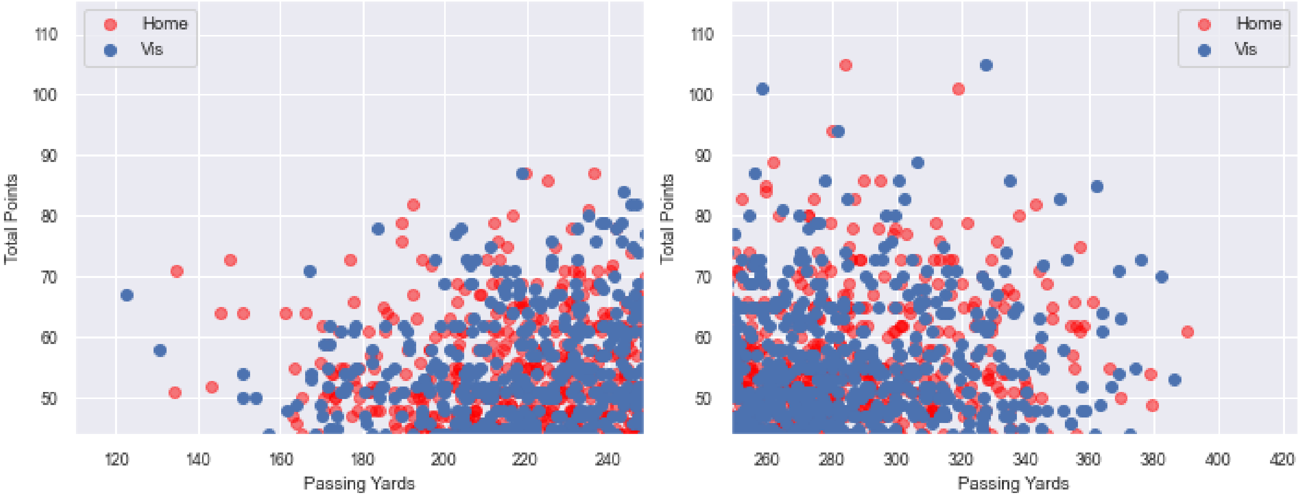

Figure 16. Upper left and right quadrants of superimposition of Figure 15 scatterplots, with *H_Off_Pass* in red and *V_Off_Pass* in blue.

team, as opposed to that of the visiting team, in which high scores seem to correspond chiefly to high passing yardage, suggests that home teams without prolific passing offenses may have more viable ways to score and produce high scoring games, whereas a potent passing offense might be more necessary for the visiting squad.

Summary

- Dropping of *_TD*, *_Pass_Metric*, *_Rush_Metri*c, and *_RZ_Pct* features enable determination of feature importance in Ordinary Least Squares regression base model
- Variety of features are similar in importance, including *H_Pts*, *V_Off_Pass*, *V_RZ_Def_Pct*, and *Wind*
- There is a larger number of visitor-related features in most important features, suggesting importance of home team to Total Points is mainly captured in *H_Pts*, whereas more diverse features dictate visitor-team contribution to Total Points
- In generating high point totals, visiting offensive passing may be more determinative, whereas the home team may be more likely to participate in high-scoring games without a potent passing offense.

E. Outlier Detection and Handling for Model Optimization

As mentioned previously, this dataset does contain a few outliers, primarily to the high side. Figure 17 shows a residual plot from the base model. As can be seen, there is little correlation between the residuals and the predictions of the model. However, as Figure 11 shows, a number of the outliers that are present are influential observations, based on the Cook’s distance parameter. Thus, we consider whether removing the most influential points would benefit the modeling of this problem. Using the training set of

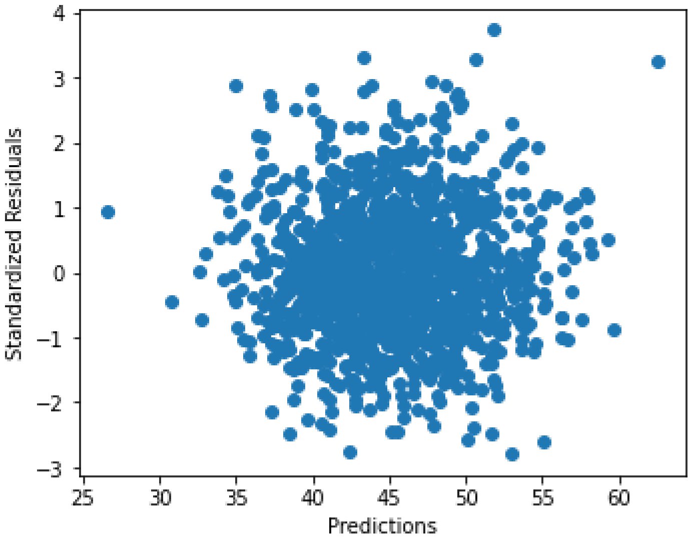

Figure 17. Residual plot showing the standardized residuals vs. the predictions for the base model.

1301 observations, the numbers of observations with Cook's distances that exceed 2, 3, and 5 standard deviations above the mean Cook's distance are 67, 28, and 6, respectively. We consider cases with those observations dropped from our set as well the case in which the two most influential observations, described in Section III.C. are dropped and the case in which no observations were dropped. With these five training sets of [0, 2, 6, 28, 67] observations dropped from the original training set, we use Recursive Feature Elimination from Scikit-Learn, an algorithm that recursively selects the features most relevant to predicting the target variable, given the number of features desired. With the number of features varied from 6 to 48, the model is run with each of the five training sets to determine if removing influential observations would yield more optimal results.

In this analysis, 25% of the dataset is left aside as the test set, while the remaining 75% is divided into training and validation sets at a ratio of 60%/15% under the technique described in Section IV.A. The validation set is used for evaluation, with the results shown in Figure 18. There is a general decrease in MAE for cases with greater than 30 features, with the exception of the 67- dropped case. The best results occur for the 0- dropped and 2- dropped cases. Because the 2- dropped results are not significantly superior to the 0- dropped results, we err on the side of retaining as much information as possible, choosing the 0- dropped case and retaining all 1301 observations from the training set.

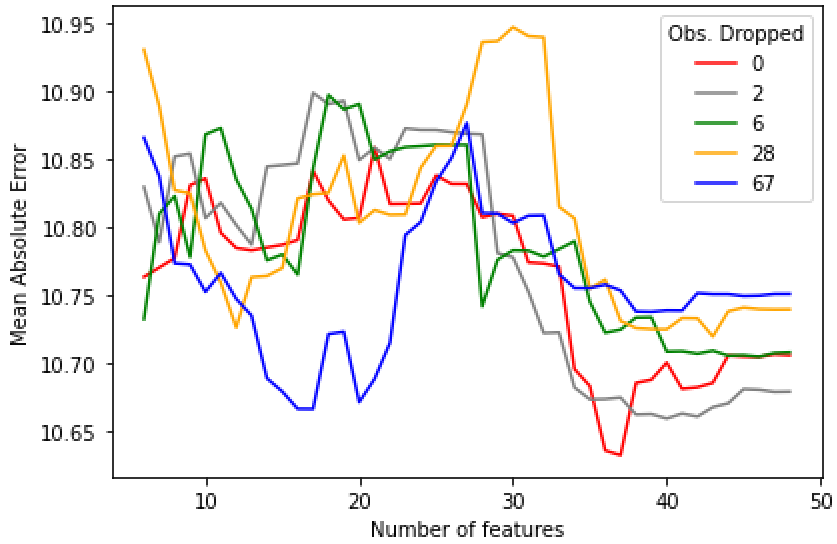

Figure 18. Mean absolute error obtained by the recursive feature elimination algorithm for the following training sets: 0 observations dropped (red), 2 observations dropped (gray), 6 observations dropped (green), 28 observations dropped (yellow), and 67 observations dropped (blue).

Summary

- The effect of removing highly influential points is examined
- Mean absolute error values are generally lower for models with greater than 30 features than for those with less than 30 features
- Best results are shown for models in which no observations or only two observations are dropped. The model chosen is that with no observations dropped.

F. Feature and Model Selection

To determine the best model for predicting total points for the contests in our dataset, four regression models are examined — optimized Ordinary Least Squares (OLS), Ridge regression, Random Forest (RF), and the XGBoost gradient boosting method (XGB). The optimal number of features for each model is determined by recursive feature elimination (RFE) with 5-fold cross-validation. Once the optimal number of features is obtained, RFE is then fitted on the entire training set at once, yielding the select subset of features, which is then applied to the test set for final evaluation. In addition, the hyperparameters of the respective models are tuned by applying this process to a selection of hyperparameter values for each model, yielding the ideal set of parameter values. The optimal models with the calculated metrics on the evaluation test set are described in Table 3 and Figure 19.

Table 3. Optimization results of four regression models applied to the NFL dataset. The models were tuned through recursive feature elimination with cross-validation and hyperparameter tuning.

Regression Model	Ordinary Least Squares	Ridge	Random Forest	XGBoost
# of Features	38	45	37	41

Tuned Hyperparameters		alpha = 0.44	n_estimators = 25,
			max_depth = 5,
			min_samples_leaf = 1
				learning_rate = 0.11,
				n_estimators = 50,
				max_depth = 3,
				subsample = 1.0,
				colsample_bytree = 0.3,
				gamma = 0.0,
				min_child_weight = 1,
				reg_alpha = 0

Top five features	V_Off_Pass, H_Pts, V_Pass_Metric, Wind, H_Pass_Metric	Wind, V_Off_Pass, H_Pts, V_RZ_Def_Pct, grass	V_Off_Pass, H_Off_Pass, H_Pass_Metric, V_RZ_Def_Pct, V_Pts_Opp	H_TO_Lost, H_Off_Pass, V_Off_Pass, V_TD, H_TD
Test MAE	10.701	10.516	10.505	10.571

Comparing the four models, they all exhibit large numbers of features, ranging from 37 to 45, demonstrating the nature of this total points prediction exercise in which no one feature or small set of features stands out as dominant. In fact, in the RF model, with the exception of *V_Off_Pass* (0.11) and *H_Off_Pass* (0.08), no other feature shows a feature importance greater than 0.045, and in the XGB model, no feature exceeds this value in feature importance. There are a few features that consistently rank as among the most important features across the models, such as *V_Off_Pass*, *H_Off_Pass*, *H_Pts*, *V_RZ_Def_Pct*, and *H_Pass_Metric* (Table 3). The RF model exhibits the lowest

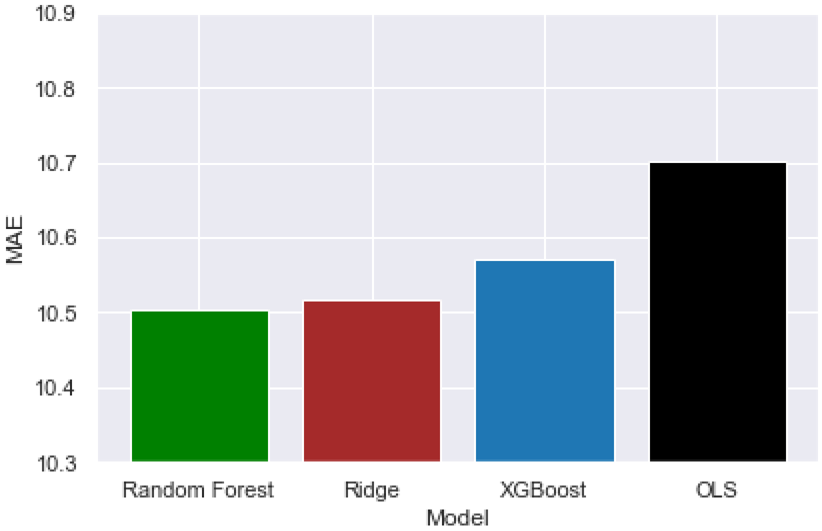

Figure 19. Mean absolute errors of the four examined regression models as obtained on the test set.

MAE value, followed closely by the Ridge model (Figure 19). Thus, the model of choice is the RF model, with the optimal hyperparameters listed in Table 3.

Summary

- Recursive feature elimination and hyperparameter tuning is used to obtain optimal model
- No one feature or set of features is dominant in the models, but passing yards is consistently among the most important features
- Random Forest (RF) model with 37 features gives the best results

G. Threshold Tuning

For a betting strategy to be successful against a sportsbook, a success rate of at least 52.38% is necessary to overcome the 10% vigorish typically charged by the bookmaker. The O/U accuracy obtained by the chosen Random Forest model over the entire 433-game test set, at 54.9%, is sufficiently accurate to produce profitability. However, to enhance profitability, one might be interested in applying this prediction model to a subset of games that may carry with it a higher probability of correct prediction of Over vs. Under with respect to the posted O/U. To consider this, a *threshold* is applied, under the assumption that greater deviation from the O/U value corresponds to a greater likelihood that the final result with respect to the O/U would be in that direction. Applying a threshold is desired when profits of equal or greater amount can be attained with betting fewer games, as risk is reduced and bets placed are more efficient, yielding greater returns. Threshold values of 0.01 to 5 are applied in both the Over and Under directions, with 0.01 (-0.01) effectively representing all Over (Under) predicted games. The training set is used to tune the threshold for the chosen Random Forest model, with the results presented in Figure 20.

Profits calculations are performed for a typical week's worth of games, in which $100 is wagered for each contest that is bet. In a typical week of NFL football, 16 contests take place. With the application of a threshold, not all games are bet. Here, the percentage of games that are eligible for betting is calculated and applied to the calculation. For a bet of $100, it is assumed that a winning bet pays $90 to the bettor, while a losing bet loses $100. Earnings are then divided by the number of eligible contests, as determined by the threshold, to yield the percent return per bet (Figure 20b). This value is then multiplied by 100, the assumed bet size, multiplied by 16, the assumed total number of games per week, and multiplied by the percentage of games eligible for betting to yield the profit, shown in Figure 20a. Naturally, the profits are going to be considerably larger for results obtained from the training set; however, these results can be useful in obtaining an appropriate threshold to apply to other datasets.

Table 4. Threshold tuning applied to the training and test datasets.

Threshold	Training Set	Test Set	Threshold	Training Set	Test Set
-5	$92 $0.55	-$24 -$0.19	All Over	$179 $0.24	$42 $0.05
-4.5	$112 $0.53	-$5 -$0.03	0.5	$175 $0.27	$31 $0.04
-4	$120 $0.47	$7 $0.04	1	$173 $0.31	$18 $0.03
-3.5	$118 $0.38	$32 $0.12	1.5	$163 $0.33	$53 $0.09
-3	$144 $0.39	$39 $0.12	2	$165 $0.40	$29 $0.06
-2.5	$163 $0.38	$29 $0.07	2.5	$154 $0.43	$13 $0.30
-2	$194 $0.39	$19 $0.04	3	$132 $0.44	$9 $0.03
-1.5	$204 $0.35	$8 $0.02	3.5	$109 $0.46	$33 $0.13
-1	$222 $0.34	$9 $0.01	4	$96 $0.51	$47 $0.25
-0.5	$225 $0.29	$21 $0.03	4.5	$83 $0.57	$63 $0.44
All Under	$222 $0.26	$45 $0.06	5	$75 $0.63	$47 $0.43

Each cell in the training set and test set columns shows the weekly profit above and the return per bet below in green (red) for positive (negative) values.

From these results, a threshold of -1 in the Under direction and +2 in the Over direction appear to be the most optimal. For the Under case, the profits remain nearly constant for the -0.01, -0.5, and -1 thresholds. As the increase in return per bet levels off after -1, the profit starts to decrease. For the Over, the smaller profits as compared with the Under are primarily due to the smaller number of eligible games, as the chosen model produces results corresponding to the Over only 46% of the time. Nevertheless, the accuracies for the all Over and all Under cases are similar. When applied to the test set, this threshold yields a subset of 67% of games eligible for betting, producing a weekly return of 3.5% or a seasonal return of 42%, with an O/U accuracy of 53.8%.

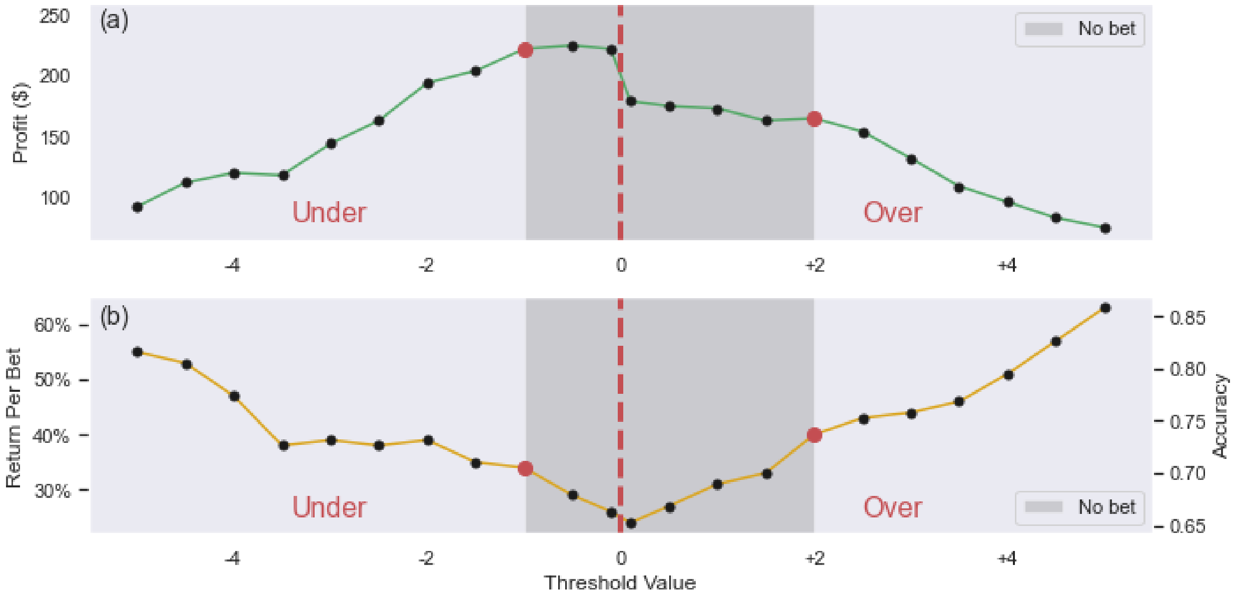

Figure 20. Applying a threshold for the margin between the prediction and the posted over/under value over the training set, the a) total profit and b) return per bet and accuracy for a week's worth of games, assuming a $100 bet for all wagered contests.

Summary

- Thresholds of total score result with respect to posted Over/Under are applied to optimize betting strategy
- A threshold of -1 (Under) and +2 (Over) demonstrates to be the most optimal, permitting bets on 67% of games played

V. Conclusions

Main Points

Data analysis and regression modeling has been applied to a dataset consisting of 2559 NFL games spanning a period of ten years. From this set, a total of 1735 observations and 48 features were used in model training and testing to determine if using a model-predicted final total score of an NFL contest to bet the Over/Under can be a profitable betting strategy. The major findings are as follows:

- Passing yards is consistently the most important single statistic in affecting total score output.
- Passing yardage is significantly affected by temperature. This is not adequately taken into account in Over/Under determinations, leading to an abundance of Over wins in the first ten weeks of the season and Under wins in the final seven weeks.

- Wind is also a very significant factor in the total score, so much so that it appears as a top five feature in importance in a variety of models
- The most important defensive feature relates to visiting defense red-zone percentage. This feature also exhibits a relatively low correlation with Over/Under, suggesting its lack of consideration
- There is no one dominating feature in the prediction of total score, as many features exhibit similar levels of importance, and all optimized models employ a large number of features.
- The Random Forest model with 37 features produces the best results, demonstrating a profit when applied to the test set. These results can be further optimized with respect to the training set with the application of a threshold of -1 for the Under bet and +2 for the Over bet.
- Applying this threshold to the test set yields a seasonal return of 42%.

Future Work

This study can be improved with a couple of considerations. First, the dataset runs from 2010-2019. Presently, another year and a half of data are available, which would improve the training of data. This would promote the use of a validation set, which would have benefitted some analyses in the study but was forgone on account of lack of data. Although more data would be beneficial, care must be taken when including data from before the study period. The NFL game has changed considerably over the past 10-15 years, and relationships, particularly involving passing statistics, may not be the same when extending further into the past. The study included the assumption that the relationships among the features are constant throughout the study period. This allows the prediction of contests throughout the time period without incurring data leakage. However, the evolution of these relationships over time may be a worthwhile study that could improve model prediction. This would merit a time series analysis. In addition, the threshold tuning can be further optimized by varying the placed bet based on the deviation from the sportsbook O/U. Further study would investigate the optimal amount to increase the bet as the deviation extends past the threshold. Nevertheless, this study demonstrates that a betting strategy of Over/Under can definitively benefit from the application of data science/machine learning.

Summary

- More recent data is available for implementation into the model
- Considering the dynamic nature of football, the evolution of features over time may be an important consideration
- Varying the threshold tuning can provide further model optimization