Capstone Exploratory Data Analysis

Game data from a total of 2559 regular season NFL contests covering the ten-year period of 2010-2019 were obtained from the Pro Football Reference website (http://pro-football-reference.com). From this dataset, game conditions and various statistics averaged over previous contests were used to predict the total score for 1734 games. 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

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

reason, only games from the sixth week to the last week of the regular season were considered for prediction. A description of the features considered in this analysis is provided in Table 1.

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, the discrete nature of scoring in football is very apparent, with certain values (linear combinations of 7 and 3) showing dominance over others.

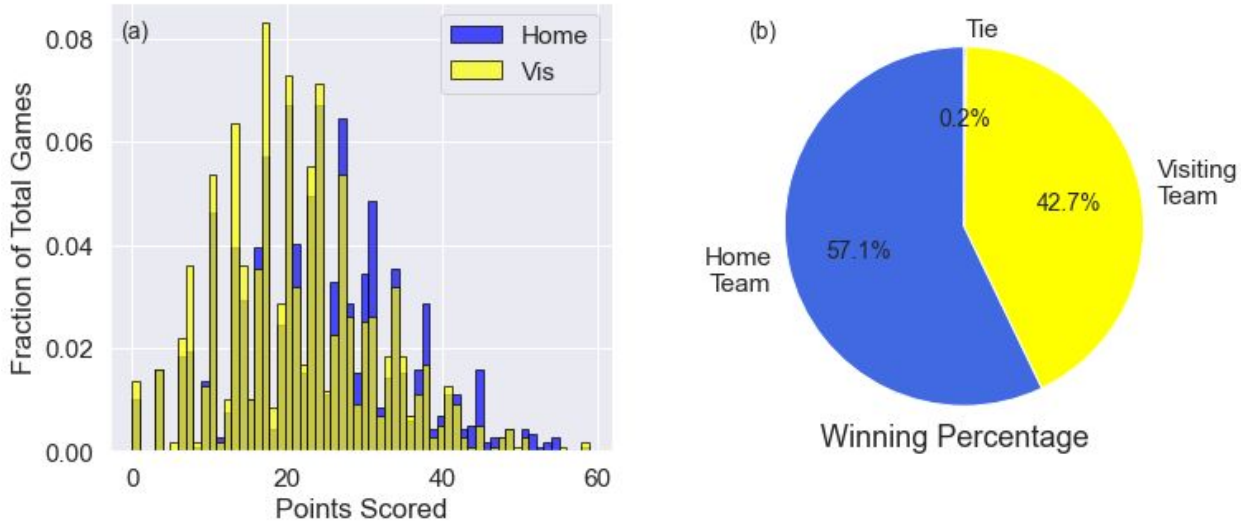

Figure 1

The values range from 0 to 59, with a mode of 17, clearly a very common score in American football. With the dark yellow representing the overlap between the two distributions, it is clear that visiting teams (bright yellow) tend to dominate in the smaller point values, whereas home teams (blue) tend to populate the higher values. The

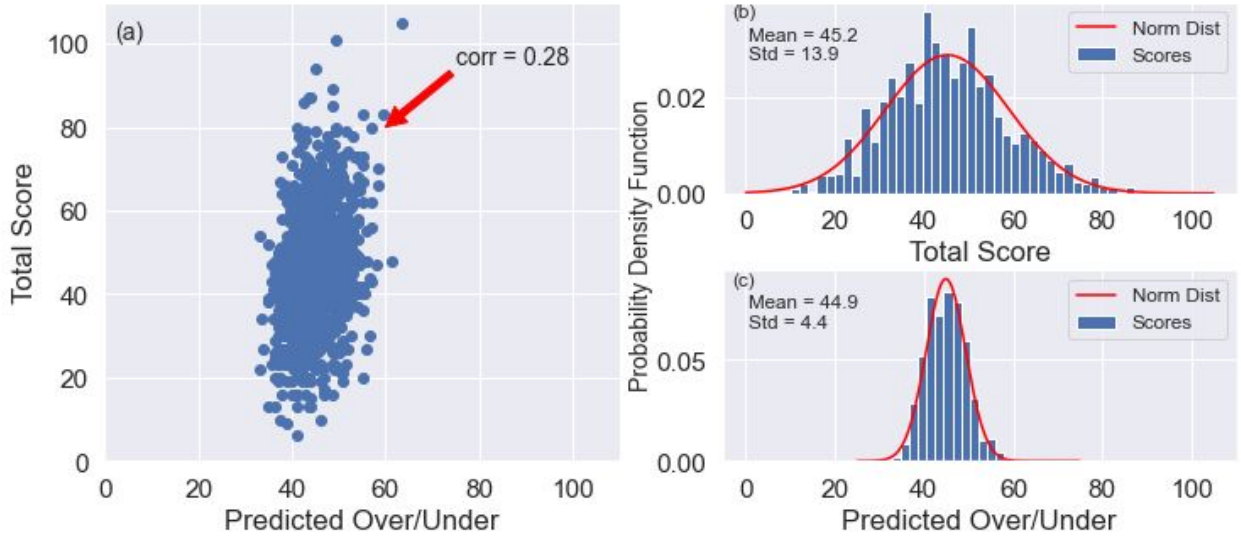

Figure 2

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.

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).

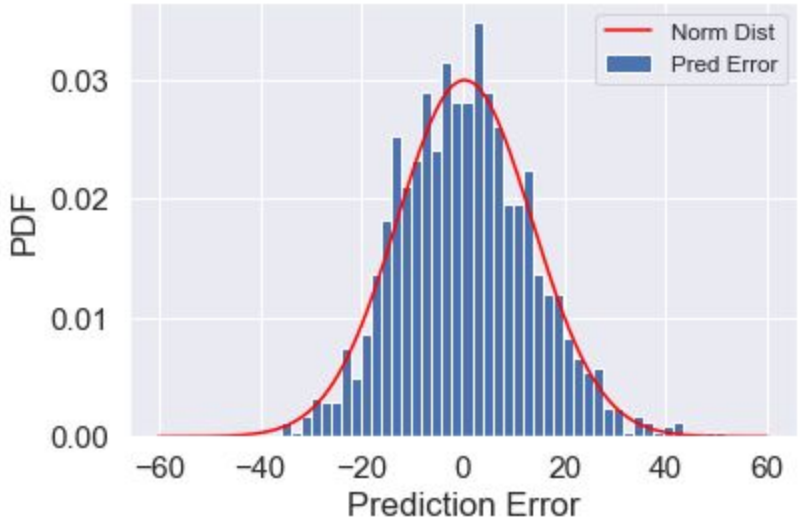

Figure 3.

The prediction error, defined as the difference between the O/U and the actual total score, approximately follows a normal distribution about zero (Figure 3), with remarkable symmetry in the Over (median = 9.0, mean = 11.0) and Under (median = -9.0, mean = -10.4) results. The following boxplot (Figure 4) shows that the outliers primarily occur to the upside but accounts for only 3.4% of the total distribution. The production error varies primarily between -40 and +40, regardless of the underlying

Figure 4.

O/U prediction (Figure 5a), demonstrating a very small correlation coefficient with O/U of -0.04. Over the ten-year period of the study, the Over has an almost identical winning percentage as the Under (Figure 5b). However, the evenness between the two sides

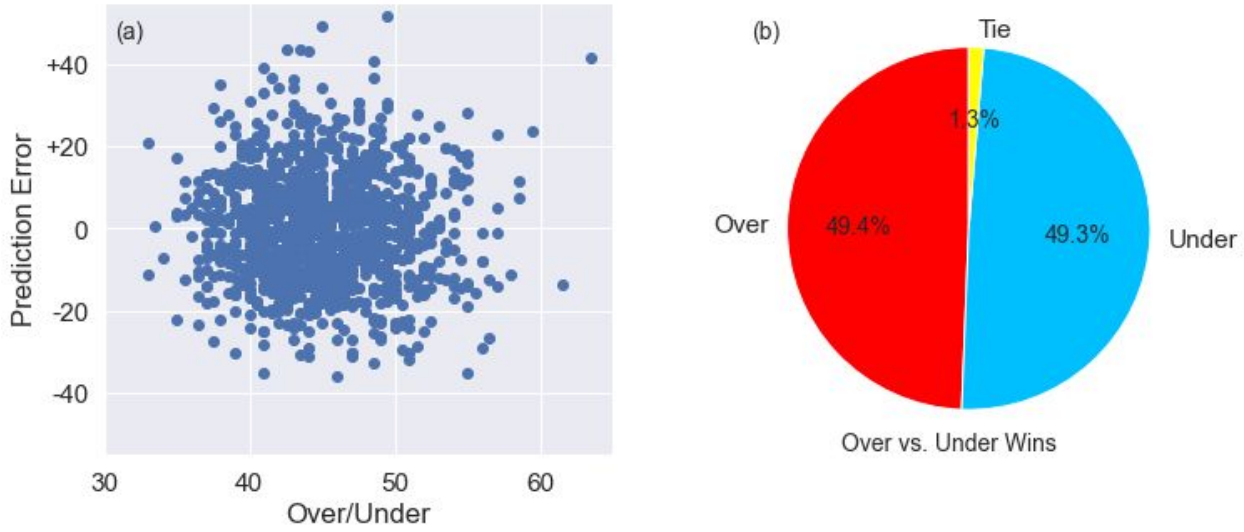

Figure 5.

does not carry over year after year. In fact, there is considerable variance in the annual Over winning percentage (Figure 6), which varies from a low of 40% in 2014 to a high of 65% in 2010. Whether this disparity is simply the result of natural variance or whether

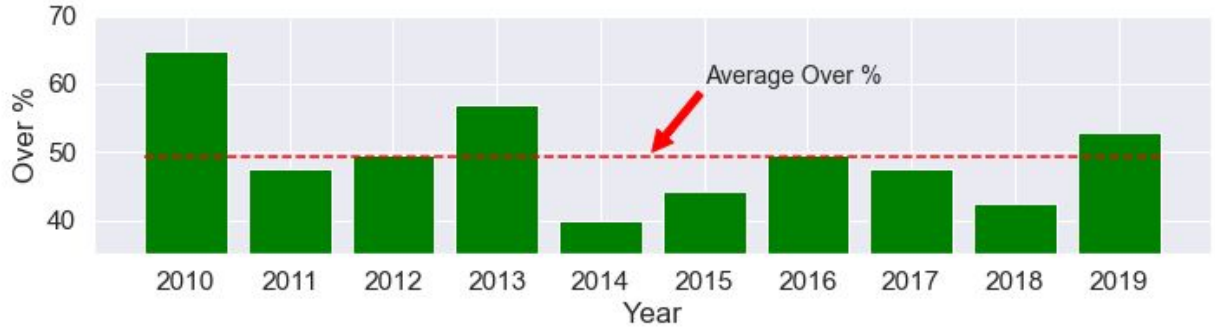

Figure 6

there are some underlying causes for this difference will be explored in the Hypothesis Testing portion of the analysis.

Although the prediction error outliers are not numerous, analyzing them might reveal some patterns that could help direct regression analysis. Figure 7 shows how the values of this subset relate to the standard deviations of the entire population for the various features considered in our analysis. The outlier subset is populated with scenarios in which home teams have strong passing offenses, particularly those combined with weak passing defenses, as reflected in the “Pass_Metric” category.

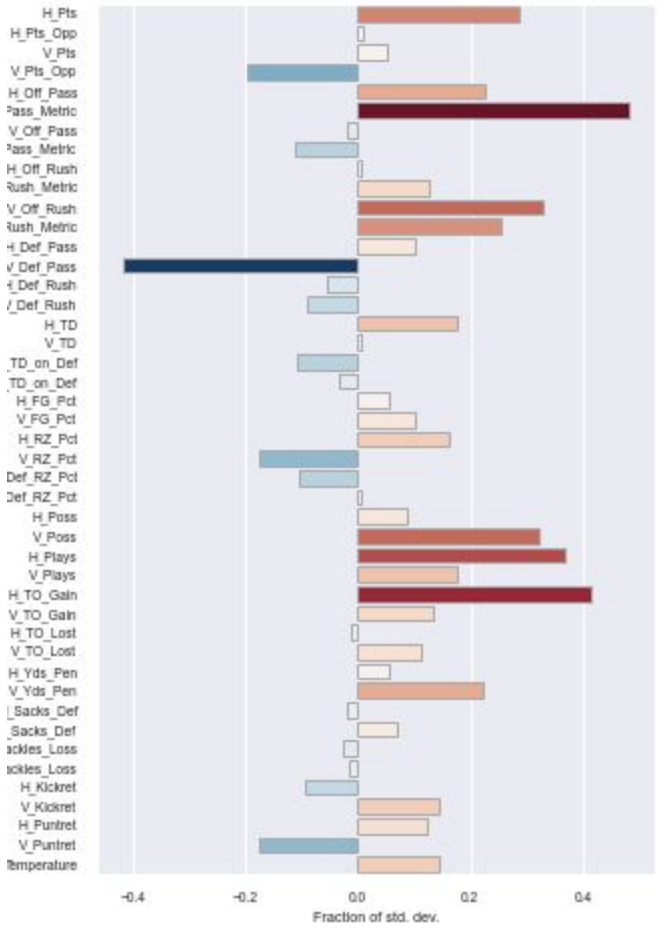

Figure 7

Games with home teams that have a high tendency for obtaining turnovers are also included in this population. These categories and their impact on total score will be further analyzed in the Hypothesis Testing and Regression Modeling portions of this analysis.