Tyler Miller

CS 555

Premier League Matchday Winner Classifier

**Research Scenario:**

This research uses the Premier League Kaggle dataset containing matchday information from 2000-2018. The project sets out to predict matchday winners with logistic and linear regression. Logistic regression is a binomial logistic regression. The linear regression predicts the continuous variable of goal difference and uses this to predict a winner.

**Research Question:** To what extent can team statistics and recent form predict the outcome of Premier League matches?

**Dataset Description:**

The dataset contains historical matchday data from the English Premier League, sourced from Kaggle. Each row represents a single football match, and the dataset includes a range of variables capturing team performance, recent form, and match outcomes. This data was used to build models predicting match results (win/loss) and goal differences.

Data cleaning included, adjusting FTR from strings representing home win and not home win to a binary classification. 1 represents home win and 0 represents not home win. Also cleaning and converting the Date from a string to a Date object. The HTH Win Percentage variable was added as a feature to improve performance. To do this, I looped through the data and found the last 3 years of matchday data for each specific matchup and provided a win percentage based on the home team. Finally, many columns were removed because they were either too complicated or deemed unnecessary.

Source: [English Premier League](https://www.kaggle.com/datasets/saife245/english-premier-league/data?select=final_dataset.csv)

| **Variable** | **Description** |
| --- | --- |
| FTR | Final match result: originally "H", “NH”, converted to binary (1 = Home win, 0 = Not home win) |
| HTGS, ATGS | Total goals scored by Home and Away team (season-to-date) |
| HTGC, ATGC | Total goals conceded by Home and Away team (season-to-date) |
| HTP, ATP | Season-to-date points for Home and Away teams |
| HTFormPts, ATFormPts | Recent form points from last 5 matches (W=3, D=1, L=0) |
| HTGD, ATGD | Goal difference (scored - conceded) for Home and Away |
| DiffPts | Difference in points between the teams (HTP - ATP) |
| DiffFormPts | Difference in form points (HTFormPts - ATFormPts) |
| HTWinStreakX, ATWinStreakX | Whether a team has a win streak of length X (3 or 5) |
| HTLossStreakX, ATLossStreakX | Whether a team has a loss streak of length X (3 or 5) |
| HMxW, AMxL, etc. | Encoded results of previous matches (e.g., "HM1W" = Home team won 1 match ago) |
| HTHWinPct | Historical win percentage of home team against away team in the last 3 years |
| Date | Match date (cleaned and converted to Date format) |

**Dataset Variables:**

**Statistical Methods:**

To answer our research question, we apply a variety of statistical and machine learning models. First, Elastic Net Regularization is used to determine the most relevant features while handling collinearity. The method used in R is the *glmnet()* library. The benefit of Elastic Net is that it combines Lasso and Ridge regularization and provides balanced feature selection, especially with correlated features.

Logistic regression is used through the generalized linear model library in R. Its purpose is to predict the probability that the home team wins a given match. P-values and coefficients are used to determine feature importance and further enhance feature selection.

Linear Regression is used to predict the goal difference between home and away team. By predicting goal difference, we can classify between home win and not home win like logistic regression. The *lm()* library is used in R to perform linear regression.

There are many ways to evaluate a model. In this experiment we use accuracy and confusion matrix for evaluating the models. We also calculate a correlation matrix to evaluate for collinearity. To ensure fair evaluation the data is split randomly into 70/30 training and testing. Finally, we use p-values to determine statistical significance.

**Results:**

Once the data is cleaned and extraneous variables are removed. The next step is to decide which features are relevant to the prediction. I use Elastic Net Regularization to select features, while also accounting for collinearity. Elastic Net Regularization is an effective way to find out feature importance. It is not the only way to do it, nor is it the only way I decide on features in this project, but it allowed me to get a feel for the data.

Figure 1 shows my feature importance according to elastic net; I select the important features accordingly from this representation. Some features have an extremely low coefficient, so they are removed. For example, I remove Home Team Goals Conceded (HTGC) as well as ATGC. This may be surprising because this is seemingly an important predictor. However, this may be elastic net accounting for collinearity with the goal difference statistic which is directly related to goals conceded. It is also important to note that goal difference has quite a high coefficient. A similar idea has occurred with goals scored but they just break the coefficient threshold to make it to the next step.

The next step is to run my logistic regression with the features from elastic net and determine statistical significance of features. Table 1 shows the summary of the logistic regression model. The accuracy rating of this model is around .63. This is not a great accuracy score but is better than average. Following this, I removed some features that were deemed insignificant by p-value and also checked for collinearity between these features. The correlation matrix can be found as figure 2. I saw some high correlations but nothing above .80 so I did not care too much. However, I kept the high correlations in mind when removing features. Next, I run the second regression that can be seen in Table 2. The accuracy for this model is around .64. This is not much better but now almost all the features are significant, and the feature list is simpler. Finally, using p-value as an indicator the feature importance is ranked in figure 3.

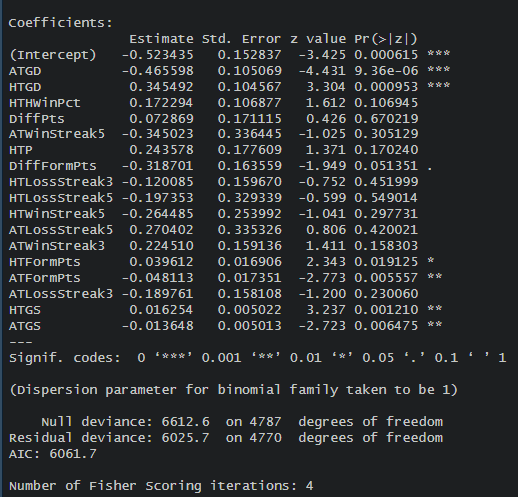
The final step is to run a linear regression on the matchday goal difference variable and determine the match winner from the regression’s prediction. The summary of my linear regression model can be found in table 3. While the accuracy and confusion matrix are in table 4. Overall, this model did not perform very well either, slightly worse than logistic regression.

**Conclusion and Limitations:**

The goal was to predict the outcome of English Premier League matches using team statistics and recent performance data. We built models to classify whether the home team would win and to predict the expected goal difference in each match. These statistical models performed okay. They were better than random guess but not by much with approximately 60% accuracy on both.

Limitations of this project include incomplete or noisy data and lack of complex models. The data is incomplete in a few ways, it lacks player information such as the lineup and injuries. Due to the nature of soccer being a strenuous sport, there are constant rest days for players and injuries. Weather and home team advantage would also be useful for this dataset. There are locations and stadiums that are known to be more hostile to play in (i.e. Stoke city weather, Liverpool stadium). A home team factor scale would be an interesting feature to include. Another way this data is incomplete is because the FTR in the dataset was binomial. In soccer draws are common and including a draw class, home win class and away win class would make a difference in results. Finally, the models could be more complex or different. Models such as random forest, decision tree, naïve bayes and SVM could produce better results due to the complexity of the problem.

**Table 1:**



**Table 2:**

A screenshot of a computer

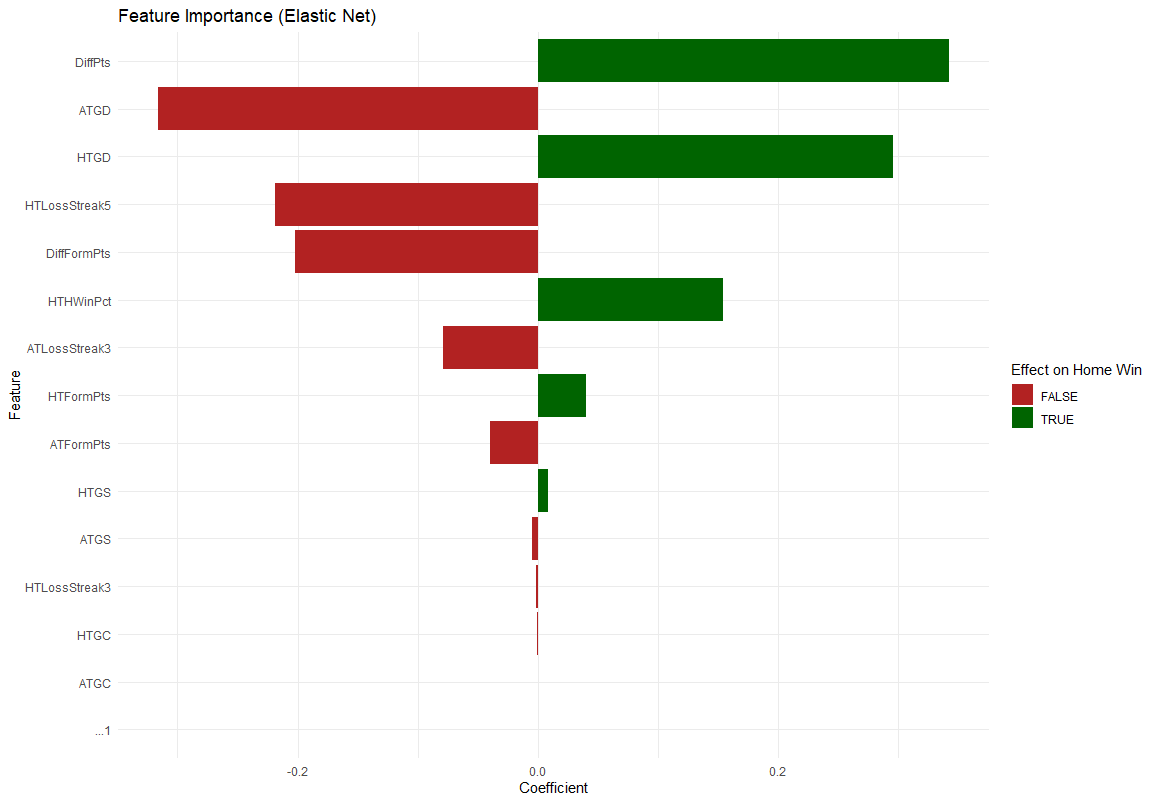
AI-generated content may be incorrect.

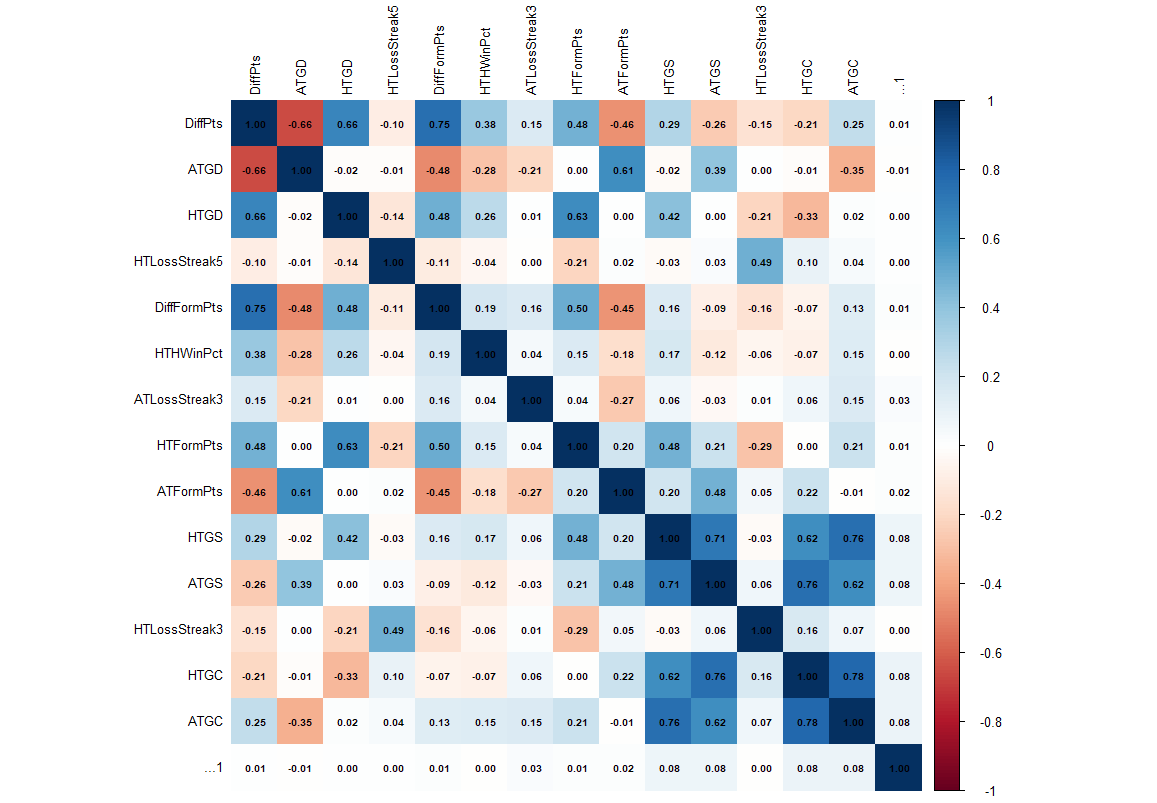
**Table 3:**A screenshot of a computer

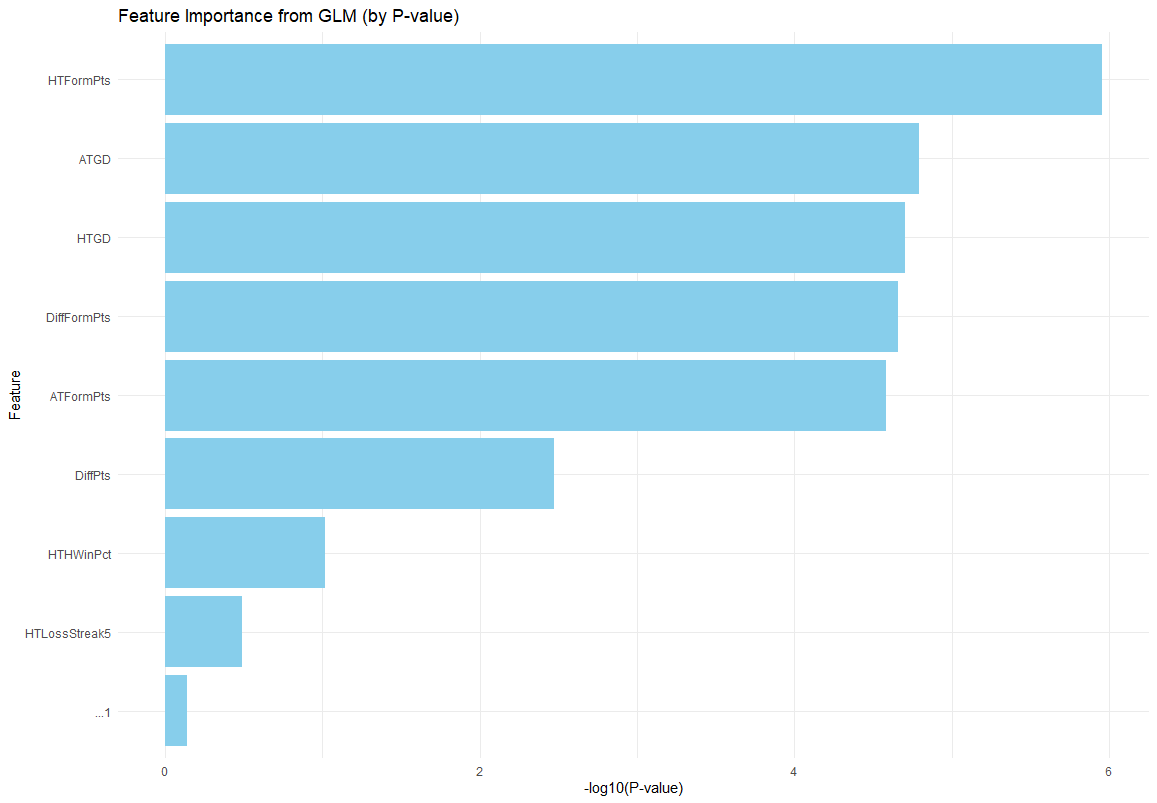
AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.**Table 4:**

**Figure 1:**

**Figure 2**

**Figure 3**