English Premier League Predictions and Analysis

Andrew Lujan

Regis University

MSDS 696: Practicum II

John Koenig

December 12, 2021

Premier League Predictions and Analysis using Machine Learning

The English Premier League continues to grow in popularity due to an increase in competitiveness across the league. New richer and bigger ownership groups continue to make the league more competitive and as such, data has become instrumental both internally for clubs and externally for media outlets, sports betting, and match analysis. What this project intends to do is have a look at data from the view of both sides of that conundrum.

Research Question

The purpose of this project is to answer the following questions:

- Which match day data points were most important in predicting a match-winner?
- What is the age Profile of a Premier League champion?

The significance of these questions are as follows: Being able to reliably predict a match-winner allows for sports betting companies to set odds and are instrumental to setting those odds. From a betting perspective being able to predict a match-winner allows for one to make more educated decisions before placing a bet on a match. From the club and analysis perspective, having an understand of the squad make-up of previous champions or high-performing teams can give some insight on where a team is at in their project or in their recruitment efforts.

Project Methodology

Data Import

The data used for this project came from two sources:

- https://fbref.com/
- http://football-data.co.uk/englandm

Data Cleaning/Feature Engineering

The data used for machine learning and making predictions came from football-data. 10 datasets were downloaded from the website and pulled into a jupyter notebook environment. The data was scrubbed, and several features were added to the set that measured: goals conceded, goals scored, team form which is their performance over the past 5 games, and total points accumulated. Winning a game result in 3 pts, a draw results in 1, and a loss is 0. So total points accumulated is particularly valuable in making predictions on league matches because they are an indicator of the team's league position. Additional efforts on the data preparation front included rearranging the columns to follow a format easier to understand and aggregating the 10 tables into one large data frame.

After the final data frame was created a few more features were added: goal difference and point differential. Goal difference is a metric that measures the difference of total number of goals scored and the total number of goals conceded. Point differential measures the difference between the points that determine the standings in the league table. A positive value indicates that a team is better, a negative value is denoted by the weaker team. The larger the gap, the bigger the difference between the two teams in league position.

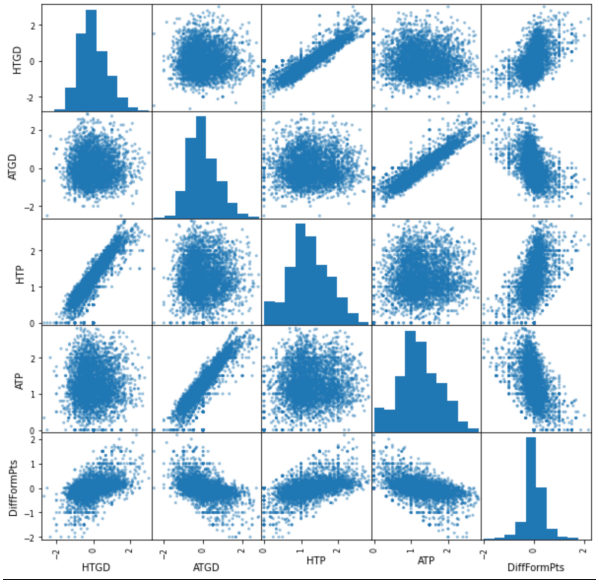
Additional data preparation that occurred was the removal of data points that will affect the models being built. Date had no positive or negative effect on the model so it was removed, additionally final scores were removed since the model would know the actual outcome of the match. Machine learning models require that data is numerical format, so data that wasn't numerical like strings that denoted a team's form needed to be converted to dummy points.

The second part of the project called in data from FBref and required minimal cleaning. There was one feature engineered to create the data visualizations which was a scatterplot pitting share of minutes played (the feature created) against Age of the player.

Exploratory Data Analysis

The first bit of data analysis was to look at the total number of matches and determine how likely it was for the home team to win the match. The reasoning behind this is simple, home teams are more likely to win matches for various reasons that include: morale, not needing to travel, and familiarity with a ground. The analysis was conducted, and it was found that the home team wins the match 44.79% of the time. This is high value considering a match can go one of 3 ways: a win, a draw, or a loss.

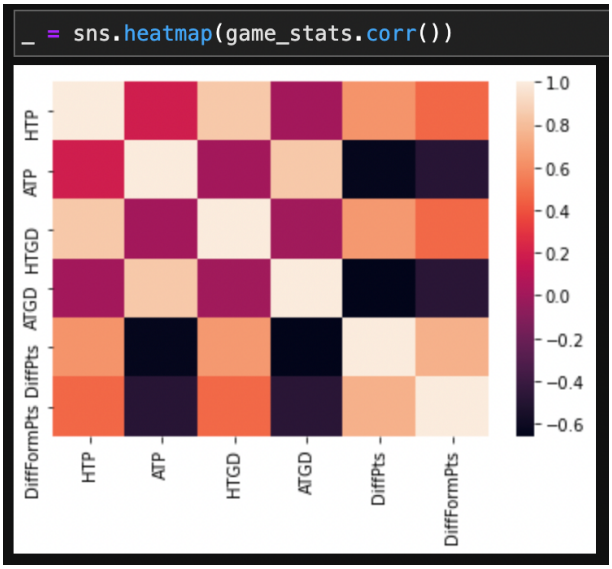
The next part of data analysis used a scatter matrix to observe the correlation between various features of the dataset.



The interpretation of the scatter matrix was that we could see some linear relationships between some of the variables. Home team goal differential and Home team points had a strongly positive linear relationship. The same could be said for the Away team goal differential and away team points. There seemed to be a very loose negative relationship between the

difference in form points and the away goal team differential. There was a loose positive relationship between home team goal differential and the difference in form points.

From there, I thought I should have a further look at these correlations through a seaborn heatmap which is useful in depicting the correlation between features of a dataset. This heatmap would allow me to answer one of the research questions which was which data points of match day data are most useful in predicting a match-winner?



From the heatmap above the following points seems to have the strongest correlation to the variable full-time result (FTR) on the 0 index. These features were

- DiffPts- difference in the league pts between teams
- DiffFormPts- difference in the form between teams from the past few matches
- HTGD- the goal differential for the home team
- HTP- the total number of points for the home team.

As we can see, the information relating to the home-team was highly correlated to the FTR and thus adds real value to any model that may be built to make predictions of a match.

Model Building

The problem that I was trying to solve using machine learning was a classification problem. More specifically, it was a binary classification problem. I used 3 different models and evaluated their results. These models included: logistic regression, a support vector machine classifier, and a random forest classifier. With the help of a tutorial on the topic I was able to perform this quickly in a few commands, 3 functions were created that allowed for me to measure the time to train the model, fit the model, and measure the score.

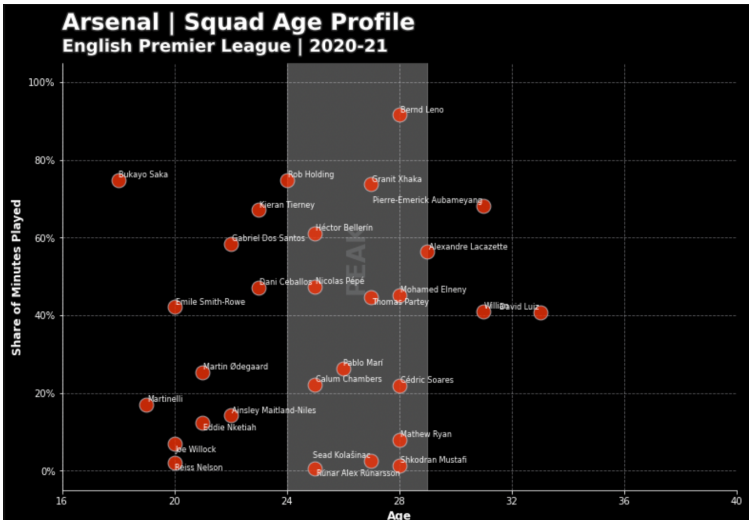
Results

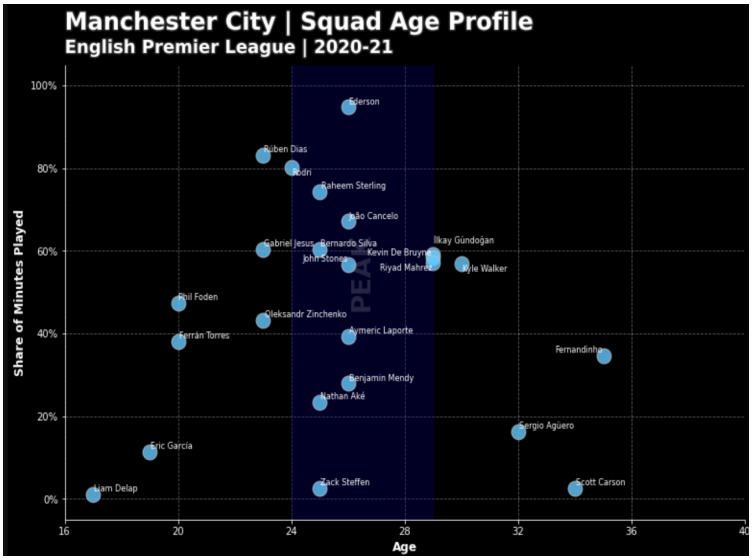
Logistic regression outperformed the support vector classifier and random forest classifier in both the f1 and accuracy score performance measures. The f1 score for the logistic regression model was .61 while the accuracy score was .66. However, I wanted to do further performance evaluation and rebuilt the models my way and used a classification report to summarize the models and their performance. Again, logistic regression outperformed the other models with higher scores in precision, recall, f1, and the accuracy scores. So, I opted to try and optimize that model by tuning the parameters with the help of grid search. What I found was that the grid search didn't do much as the models optimum tuning for the logistic regression model didn't improve the model by that much. Perhaps including further data points could raise the accuracy score of the model.

Data Visualization

The final part of the project included some data visualization exploration. I wanted to use data visualization to perform an age profile on the big 6 premier league clubs. The reasoning behind only choosing the big 6 was simple, of the previous 25 champions of the premier league title, only once has a team outside the big 6 won the premier league. To perform this analysis, I generated a scatter plot for teach team that measures a players share of minutes played

throughout the season to against the players' age. Additionally, a zone was drawn that represented the prime age range for a premier league footballer from 24-29.





This is a comparison of the best lowest ranking member of the Big 6 in the 2020-2021 season Arsenal vs. Manchester City the premier league champion of that season. Arsenal finished 8th. The findings from the plots were that teams with smaller squads and. More players in the peak age strip that contributed high amounts of minutes tended to perform better than team's like Arsenal. Arsenal received meaningful minutes from younger players and the player's in their primes tended to contribute less meaningful amounts game time. This indicates a

problem with recruitment because the players who are supposed to be in their primes are getting less game time than youth-players.

Conclusion

In conclusion, the machine learning models that were built to predict wins based on match day data were able to achieve respectable scores with logistic regression being the best model of the ones tested. The answer to my first research question which data points were most important in making predictions were: data points related to difference in league standings, difference in form, and goal differential and total points for the home team. The answer to the second research question relating to the age profile of championship caliber teams is: while age is important in squad performance, the percentage of players in their prime years and the minutes they contribute can give some insight on how succesfully a team has recruited. Teams with lower amounts of players in their primes playing meaningful minutes indicates an overreliance on experience, or teams at the start of a youth project.

References

https://www.youtube.com/watch?v=6tQhoUuQrOw&t=1264s

Metrics To Evaluate Machine Learning Algorithms in Python-

https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Tune Hyperparameters for Classification Machine Learning Algorithms-

https://machinelearningmastery.com/hyperparameters-for-classification-machine-learning-algorithms/

https://sharmaabhishekk.github.io/mpl-footy/main/2021/08/09/squad-age-profile.html