Gordon McIntire

Final Report:
What it Takes to Win in the EPL

Problem Statement

We are starting a sports betting business! With some interest in european football and some experience with data science as well, my friends and I have decided to bet on the outcomes of upcoming football matches happening in England. Since we are just starting out, we will restrict ourselves to simply betting on the final result.

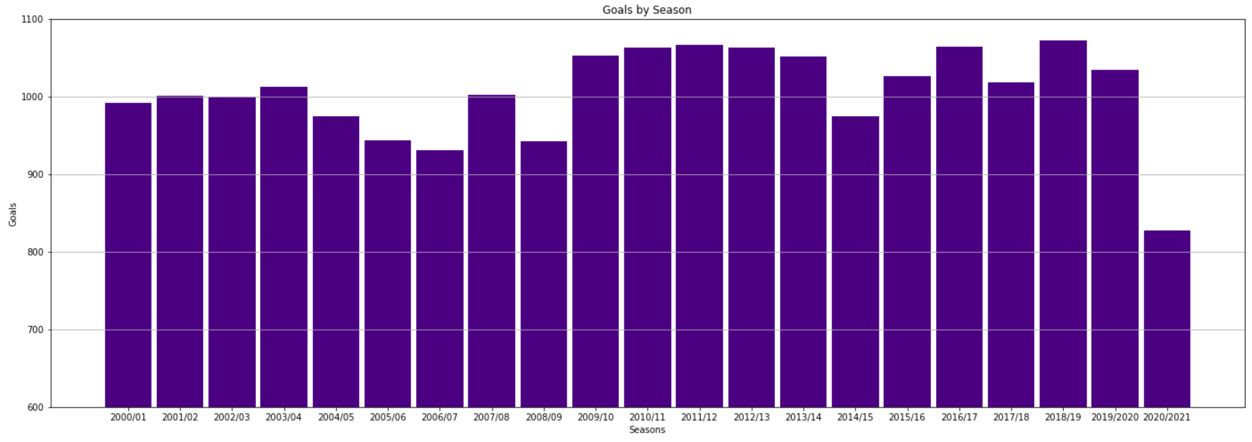
By using the data from almost 8000 matches played over the past 20 years, I utilized Supervised Machine Learning and created a Logistic Regression model and a more accurate Random Forest model to predict the winner in any football match.

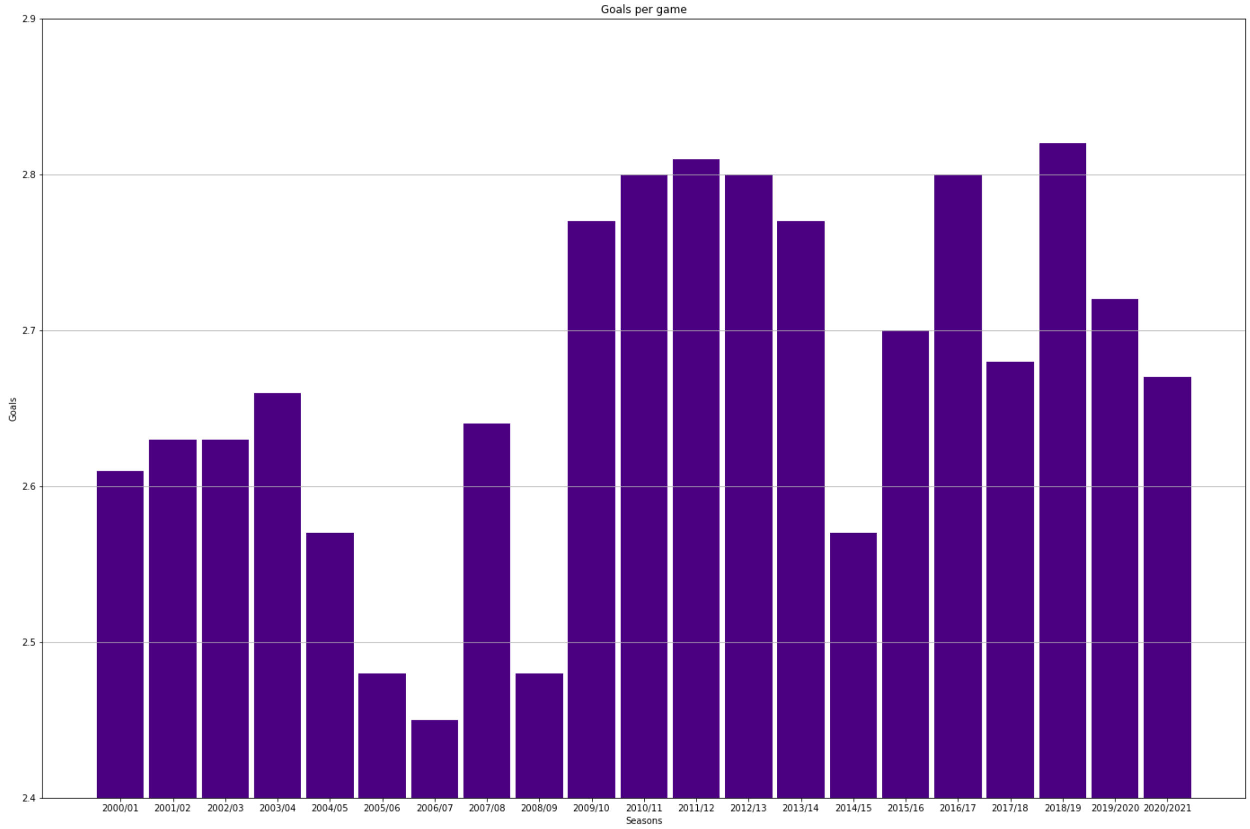
Data Wrangling

The initial dataset was relatively clean and required only a little reduction, having 10734 rows and 23 columns. However, for the seasons 1993-2000 we only have the score of each game. Starting in 2000, we have a lot more features from each match, and this is the majority of our data, so we remove those earlier seasons. Apart from that change, the remaining data is without any null values or incorrect entries. I moved ahead to the exploratory phase with 7910 rows and 23 columns.

Exploratory Data Analysis

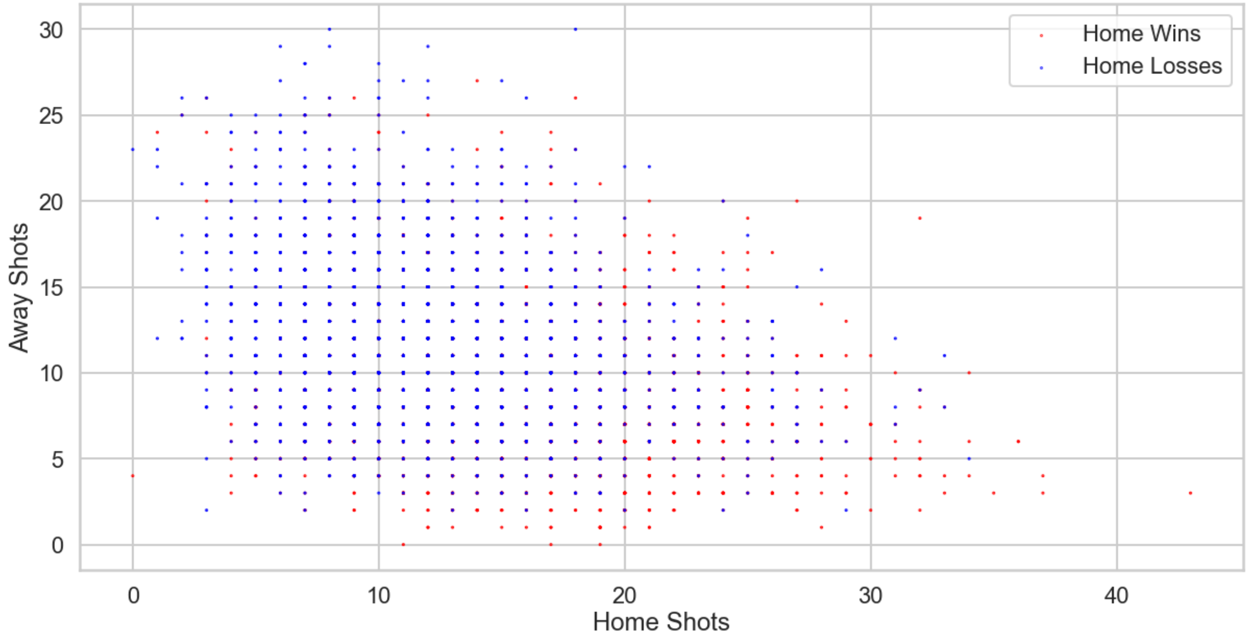
I looked at the level of goalscoring throughout the seasons, as well as the goals per game.





Ultimately, we want to determine what contributes to the final result of a football match to be a home win, loss, or draw so that when we have match statistics like these we can make a prediction on the outcome with the help of a model.

I used scatterplots to explore the weight that certain features have on the final result of a match.



Model Selection

With an aim to predict a match's outcome based on certain criteria, I tried out a couple different machine learning classification models: Logistic Regression and Random Forest Classifier.

Based on the figures above depicting home/away shot quantities and what final results they tend to produce, I knew I wanted to form a Logistic Regression model to see how well this feature would perform in prediction. It didn't perform very well, with only 0.59 accuracy, so I plugged shots on target instead. This produced an accuracy score of 0.67, which is better but certainly not enough to be satisfied.

Moving onto the Random Forest Classifier and looking to classify a match's result based on its shots, shots on target, fouls, corners, and cards, the model's performance scores were disappointing, just barely getting to an accuracy of 0.544.

Takeaways

Using Logistic Regression with on-target shots did produce model that one could use to predict the outcome of a match, though not one that would instill much confidence. On-target shots was still a better statistic to model on than shots, which we less important. Features like fouls, cards, and corners didn't reach a accuracy level close to on-target shots either.

Future Research

Working on this project gave me a lot of ideas for further research in football. I looked at data from 20 years in the premier league, but obviously there are many other leagues in the world, and it would be interesting to compare different regions. I'm also aware of datasets with more features on each match.

Another interesting project would be to look at the profile of individual teams or players, and perhaps look to aggregate players' winning tendencies to a team's chance of success.