Final Capstone Submission

Predicting PL Match Outcomes

Submitted by Matt Han

Submitted on March 29, 2020

The world of sports and data have long been intertwined and this relationship developed from an attempt for teams to gain a performance edge to bookmakers creating models for which their entire business structure relies upon. As an aspiring data scientist, my goal was to create a model that would successfully predict around 50% of the Premier League match outcomes using open source data for seasons between 2009-2019. Sports fans from around the world are always sharing their take on which teams would win their matchups but their opinions were only validated by their knowledge and "gut feeling". However, large sets of data are available in repositories at a click of a button. Bookmakers have spent huge sums of money on dedicated data teams to develop the best possible prediction models and I was interested in exploring this facet of data science as my capstone project. Many different techniques have been employed in the past and often range in complexity that is outside the scope of this project. I decided that the poisson regression model was appropriate to use in this particular case since it generated decent results based on the available data and was challenging to undertake since the course did not cover this topic.

Two data sets were used in the application of the model: one sourced from Kaggle that included in-game data (e.g. goals, cards, shots, etc) for seasons between 2009-2019 and the other from FiveThirtyEight that included SPI and expected goals for seasons 2016-2020. The datasets were downloadable in CSV format and required limited cleaning other than removing certain null values, concatenating individual CSVs into final dataframes, dropping unwanted columns, and renaming/reordering columns for legibility purposes. You can view the cleanup process in the *Capstone - Data Cleanup* Jupyter notebook and the cleaned datasets were also exported as CSV files in the "Clean Data" folder.

The poisson regression model described in this report relied on two important features: the teams and the goals scored by each team per match. The model relies on the data fitting on a poisson distribution which measures the probability of goals scored within a specific time period against the average rate of goals scored. We assume that the goals scored are independent of time, that is to say, that the outcomes of scoring do not depend on goals that have already been scored in a match. Mathematically speaking:

$$P(k) = \frac{e^{-\lambda}\lambda^{k}}{k!}, k > 0$$

Where *P(k)* is the probability of the event occurring (in this case goals scored), λ is the average goals scored, and *k* is the number of occurrences (e.g. k = 1, means 1 goal scored).

In order to ensure that the poisson regression model is valid, I had to plot the poisson and skellam distribution as part of the EDA process.

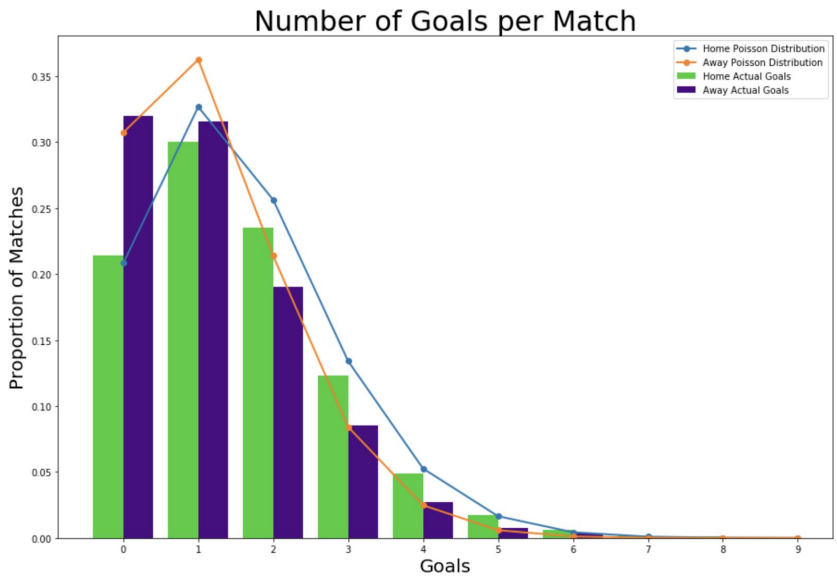
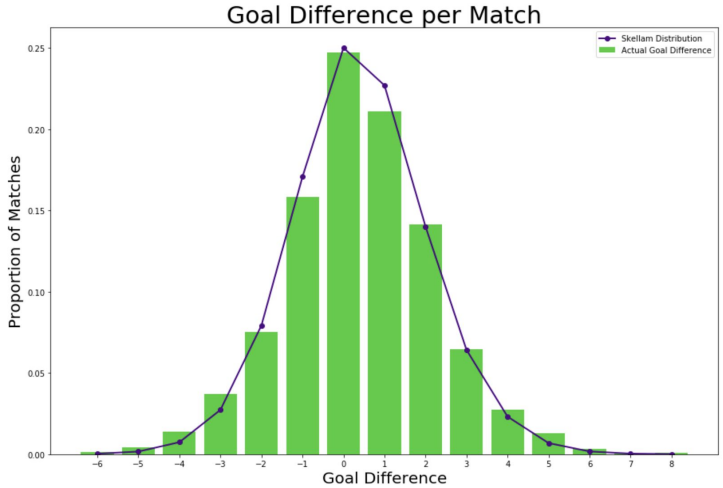


Figure 1. Poisson Distribution plotted with the number of goals scored against the proportion of matches. The Poisson Distribution is an **estimate** of the number of goals scored for both sides from 2009-2019.

Football is typically a low scoring match and more goals scored per match is less likely to occur. This is consistent with what is displayed in the plot. Furthermore, each goal scored is seen as an independent event from one another (i.e. a team scoring a goal does not lead to more goals being scored), goals from the home and away side do occur in the same instance. The Poisson Distribution is an **estimate** of the number of goals scored for both sides from 2009-2019.

We can provide further evidence to conduct a poisson regression by looking at the Skellam Distribution which is typical of poisson-distributed values. It is the difference between the two means for each goal scored in our dataset. The Skellam distribution will show a convolution of the two poisson distributions calculated above into a binomial distribution.



After gathering all the evidence to use the poisson regression model, the package used was the Generalized Linear Model (Formula) from statsmodel which allows the poisson formula to be inputted. This resulted in a summary table that lists 2 coefficients for each team; one for attacking strength where a larger and more positive number represents a higher likelihood of a team scoring a goal and one for defensive strength where a lower and more negative number represents the difficulty in scoring against the team. Both of these coefficients are taken into account when we attempt to predict the match outcome.

To predict the match outcomes and their probabilities, a function was created that allows the user to enter the home and away team as well as the maximum number of goals for which each probability is calculated. This function returns a probability matrix of each goal occuring and is important to calculate the probabilities for the match outcomes. A second function was created to take the probabilities in that matrix and sum the top triangle for a home team win, the diagonal as a draw and the lower triangle as an away team win. In order to evaluate the accuracy of the model, I ran the first function through a for loop using the test set which held data for 190 games of the final season and compared the highest predicted scoreline with the actual scoreline. If the result was the same irrespective of the actual score, then these outcomes were deemed correctly predicted. I ended up with 49% accuracy on the EPL data which is a hair short from my initial goal. In an attempt to boost the accuracy score, I added features using the SPI data such as SPI, match importance and the expected goals for each team but the accuracy dipped to 47%. It is important to note that professional models that have more time, resources and publicly unavailable data acquire predictive accuracy rates of 60-65%.

The key learning through this project is that it is notoriously hard to predict match outcomes. If it were easy and accuracy was higher, everybody would be getting rich from sports betting. One of the main reasons for its unpredictability lies in the fact that many features are unaccounted for or are unquantifiable such as game status, player transfers, starting lineups, weather conditions, injuries, player fatigue due to match schedules and manager changes. Choosing the correct amount of data is crucial as well- too much leads to inaccurate results since teams change all the time and are not always based on past success and too little does not allow the model to train properly. There is also an issue of teams getting relegated to lower leagues and they don't get promoted for years. If a team was awful during a relegation season and got promoted 3-5 years later, it is very difficult to predict their success in our model.

The point of the model was to explore whether we can predict match outcomes and to challenge myself with a model that was not taught in class. It's clear that we are able to predict these outcomes but adding the right features might boost the accuracy. Models for sport predictions are constantly being tweaked or discovered for bookmakers to get a competitive edge. My goal is to constantly optimize and utilize the best techniques to achieve the best results.