Comparing the performance of machine learning and statistical models in predicting football games

Runxing Kenneth Fu

Chinese International School, Hong Kong

Introduction

In the realms of sports analytics and betting, the ability to accurately predict the outcome of specific games has been a particular subject of interest that has undergone intense research and development. In recent years, with the widespread distribution of technology and online resources, the utilization of algorithms and models has gradually become more prevalent in sports forecasting as they are able to consume and interpret high volumes of data, far beyond the capacity of humans. Algorithms such as random forest classifiers and Poisson regression have stood out in particular, each with its own strengths and limitations.

Thus, this essay seeks to answer the following research question: **To what extent is the random forest classifier more accurate than the Poisson regression model at predicting the outcome of football games?** Through the investigation of this question, this essay will offer insight into the comparative performance of two widely used predictive models at predicting the 2021/22 season of the English Premier League.

These two models were selected due to their wide adoption and successful application in various domains while being fundamentally different. The random forest has been a conventional machine learning algorithm since its development in 2001. It has automatic learning capabilities to bolster human performance and handle complex data to achieve results with notable accuracy (El). On the contrary, the Poisson regression model, first introduced in 1837, is a statistical function used to model relationships between variables. It uses mathematical theories developed by human expertise to develop a long-standing history in sports analytics after being employed for decades now (Poisson).

Through the comparison of random forest classifiers and the Poisson regression model, this essay seeks to determine the extent to which machine learning algorithms have already surpassed traditional mathematical and statistical models in predictive performance. Thus, we can assess whether machine learning techniques can indeed outperform human intuition and domain expertise in an often unpredictable field. Beyond the scope of sports forecasting, this essay also sheds light on potential areas of improvement for machine learning algorithms in general, outlining paths for future growth in a rapidly developing field.

To answer the research question, an experiment predicting the outcome of every game in the 2021/22 English Premier League season will be conducted on both the random forest classifier and the Poisson regression model. The data from these experiments will be analyzed, with a particular focus on the accuracy of each algorithm for different teams and games.

Poisson Regression

Poisson regression, named after and introduced by French Mathematician Siméon Denis Poisson, can be classified as a conventional algorithm such that its behavior is predetermined, and the same set of inputs will always result in the same set of outputs. It is a statistical technique used to model count data (non-negative integers). These integers represent the number of occurrences of an event in a fixed interval of time or space, such as the number of goals scored in a football game. Thus, the model can be used to predict the number of goals scored by each team and forecast the outcome of the game. Poisson regression makes the assumption that the integers follow a discrete probability distribution called the Poisson distribution, which expresses the likelihood of various counts happening across a predetermined range (Roback and Legler).

Poisson regression is fundamentally designed to model the relationship between predictor variables and the aforementioned count data. It expands on the standard Poisson distribution by allowing us to understand how the predictor variables will impact the count data. It models this relationship using a log-linear equation as follows:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \ ... \ + \beta_i x_i \qquad (1)$$

where $log(\mu)$ is the natural logarithm of the expected value of the count data, $x_1, \ ... \ , x_i$ are the predictor variables, and $\beta_0, \beta_1, \ ... \ , \beta_i$ are the regression coefficients, representing how much of an impact each predictor variable has on the count data. A positive regression coefficient implies a positive correlation between the predictor variable and the count data, while a negative coefficient implies the opposite (Roback and Legler).

To determine the most fitting values for the regression coefficients (β_i), a statistical method known as maximum likelihood estimation is used to estimate parameters such that the likelihood function is maximized. The likelihood function measures the probability of attaining the actual count data given a set of parameters. By maximizing this probability, it is able to estimate the best set of coefficients for the function. This maximization is done through several numerical optimization techniques such as gradient descent or Newton's method, which continuously adjust the coefficients until it converges to the best-fitting values (Roback and Legler).

Random Forest Classifiers

The random forest classifier, designed by Professor Leo Breiman in 2001, is a type of machine learning algorithm. Machine learning algorithms differ from conventional algorithms as they enable computers to learn from data and improve their performance without being explicitly programmed. Thus, its behavior isn't predetermined, and the same set of inputs could result in different sets of outputs depending on the computer's training method. Random forest classifiers are derived from the random forest algorithm to specifically tackle classification tasks, where the goal is to categorize data into different classes. For example, when given the data for a football game, the algorithm can predict whether the game belongs in the home win, away win, or draw class. The random forest classifier uses an ensemble of many decision trees together to make its final prediction (Yiu).

To build a random forest, the algorithm first uses a procedure known as bootstrapping, where several subsets of the training data are produced. In order to build these various data subsets, bootstrapping uses random sampling with replacement from the initial training dataset. The training data for each decision tree in the random forest is provided by these subsets of data (Sruthi).

Only a portion of the available features (columns) is taken into consideration by random forest for each decision tree. This process is known as feature subsampling. The number of features considered for each decision tree is usually a hyperparameter, inducing randomness into the model. Random forest guarantees that the trees are diverse and lowers the possibility of becoming overly specialized at its training dataset, employing different subsets of characteristics for various trees (Sruthi).

After obtaining the bootstrap samples and randomly selected features, the random forest grows multiple decision trees. Each decision tree is trained independently depending on its bootstrap sample. The decision tree algorithm recursively divides the data into subsets depending on the selected features, with each division intended to maximize information gain. The process keeps going until a stopping requirement is satisfied, such as when the maximum depth is reached, when there aren't enough samples in a leaf node, or when certain hyperparameters are reached (Sruthi).

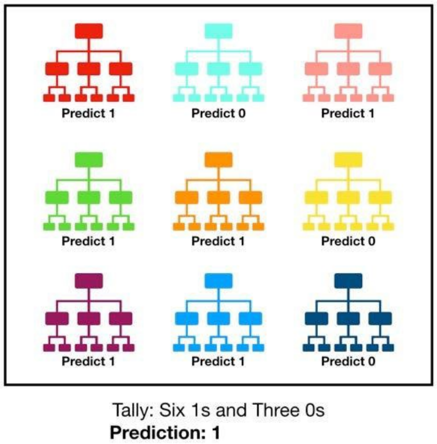


Figure 1: Visualization of a random forest classifier making a prediction ("Visualization of a random forest model")

The random forest classifier then uses the trained decision trees to make predictions on new data points. Each decision tree independently predicts which class the incoming data should belong in. A majority vote among all of the decision trees' predictions yields the final prediction, based on a simple concept - the wisdom of crowds. The class with the most votes becomes the final outputted class (Sruthi). For example, in Figure 1, out of a total of nine decision trees, six predicted 1 while three predicted 0, meaning the final prediction is 1 as it represents the majority.

Hypothesis and Applied Theory

The theory of both models has already been described in sufficient detail, so which one of the two algorithms is more accurate needs to be determined. The nature of football games and sports, in general, is very multifaceted, and many different factors come into play when predicting the outcome of a game. Additionally, form and fitness can fluctuate on a game-by-game basis, meaning the interplay between these factors is also largely varying and unpredictable. The random forest classifier's approach using the aforementioned ensemble learning process allows it to handle nonlinear relationships and can capture more intricate dynamics and patterns that influence the outcomes. However, Poisson regression, as aforementioned, assumes a linear relationship between the predictor variables and the outcome, meaning it is less likely to pick up on these intricacies. Thus, the central hypothesis is that the random forest classifier will outperform the Poisson regression model and should obtain a higher prediction accuracy.

Investigation

Data

As the goal of the experiment is to compare the performances of the random forest classifier and the Poisson regression model in predicting the 2021/22 season of the English Premier League, all games from the five seasons preceding the 2021/22 season were used for the training dataset. Data was taken from Football-data.co.uk, which contains match data from all games in the Premier League ranging from the 1993/94 season to now (2022/23). The data of each season is stored in a separate CSV file, so all the data across these five seasons were read and stored in one file to obtain the desired dataset. The key data points stored from each game were the home and away teams, the number of goals scored by each of the two teams in that game, and the result of the game (home win, away win, or draw). Additionally, the average number of goals scored and conceded by each team when playing at home and playing away are also used as data points. These data points serve as an indicator of a team's offensive and defensive "strength" and are calculated by summing the total number of goals scored for a respective team and dividing by the number of games played.

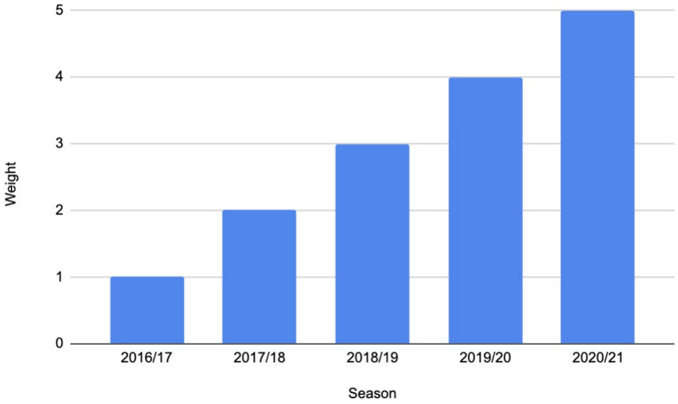


Figure 2: Weight values across seasons from 2016/17 to 2020/21 for predicting 2021/22 season game outcomes; *Google Sheets*, 5 August 2023

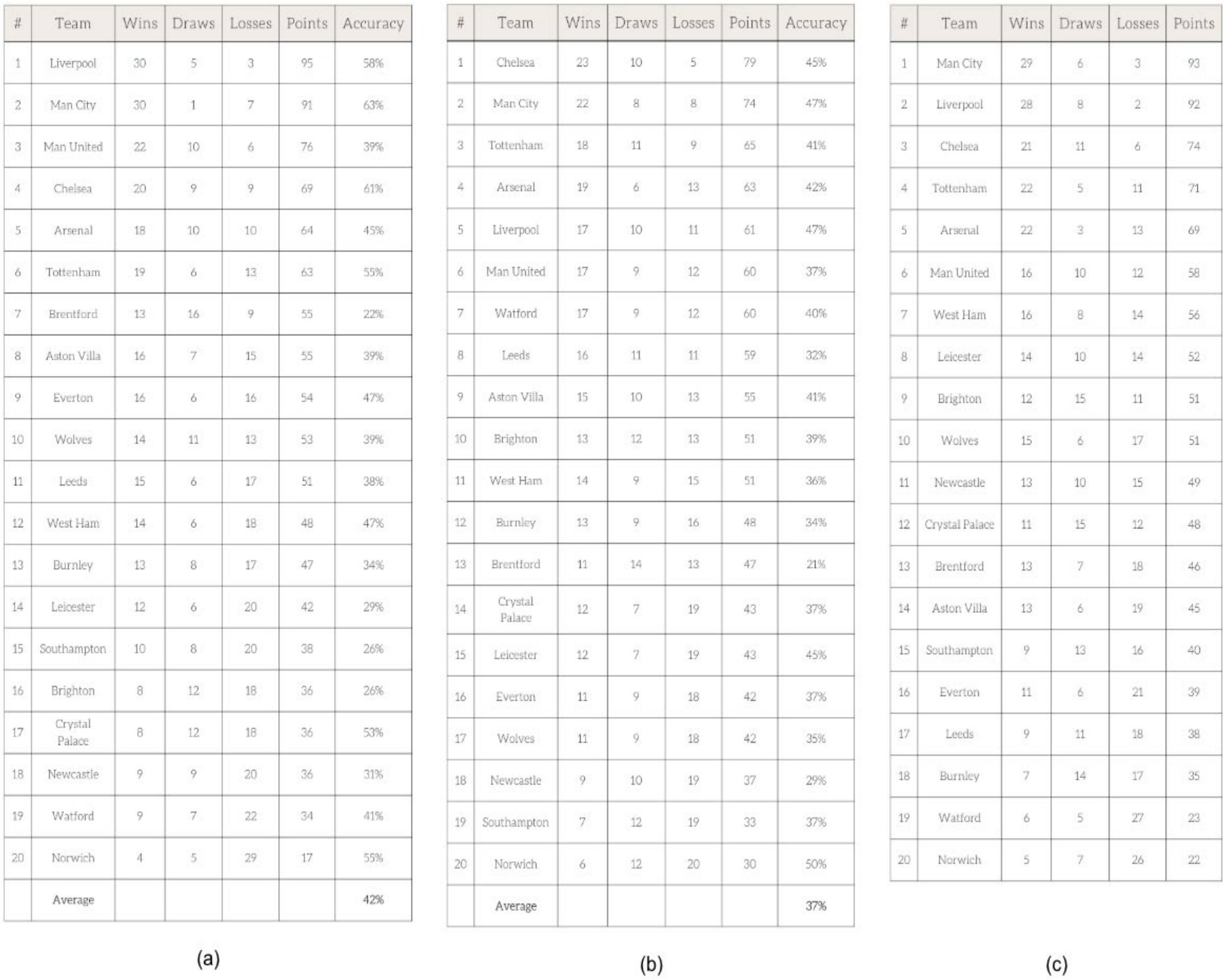
However, treating each of the five preceding seasons equally would be inappropriate, as form, personnel changes, and other factors mean that a team is likely to perform more similarly to the previous season as opposed to five seasons ago. Thus, the five seasons are weighted differently, as shown in Figure 2. The 2016/17 season starts off with a weight of 1, and each following season's weight is incremented by one, lending more importance to seasons closer to the one being predicted. To allocate the weights, the data points of each season are duplicated a number of times according to the weight. Thus, there are a total of 6080 data points used to train the two models.

Methodology

To forecast the outcomes of every game in the 2021/22 season, two Poisson regression models were fitted to estimate the scoring rates of each team, one when playing at home and the other when playing away. As for the random forest classifier, an ensemble of 100 decision trees was used to predict match results. Both the Poisson regression and the random forest classifier used the same set of training data aforementioned.

To implement the random forest classifier, the 'scikit-learn' library on Python was used. It provides tools for data mining and data analysis using a variety of machine learning algorithms, such as Random Forest. It also provides various evaluation metrics including accuracy to assess the performance of the trained model ("Ensemble methods"). As for the Poisson regression model, the 'SciPy' library on Python was used. The 'SciPy' ecosystem as a whole provides different libraries for scientific computing. For the Poisson regression specifically, the 'scipy.stats' module under this ecosystem is used, which is specifically used for statistical analysis, such as the Poisson model ("Statistical functions").

Data presentation



#	Team	Wins	Draws	Losses	Points	Accuracy
1	Liverpool	30	5	3	95	58%
2	Man City	30	1	7	91	63%
3	Man United	22	10	6	76	39%
4	Chelsea	20	9	9	69	61%
5	Arsenal	18	10	10	64	45%
6	Tottenham	19	6	13	63	55%
7	Brentford	13	16	9	55	22%
8	Aston Villa	16	7	15	55	39%
9	Everton	16	6	16	54	47%
10	Wolves	14	11	13	53	39%
11	Leeds	15	6	17	51	38%
12	West Ham	14	6	18	48	47%
13	Burnley	13	8	17	47	34%
14	Leicester	12	6	20	42	29%
15	Southampton	10	8	20	38	26%
16	Brighton	8	12	18	36	26%
17	Crystal Palace	8	12	18	36	53%
18	Newcastle	9	9	20	36	31%
19	Watford	9	7	22	34	41%
20	Norwich	4	5	29	17	55%
	Average					42%

(a)

#	Team	Wins	Draws	Losses	Points	Accuracy
1	Chelsea	23	10	5	79	45%
2	Man City	22	8	8	74	47%
3	Tottenham	18	11	9	65	41%
4	Arsenal	19	6	13	63	42%
5	Liverpool	17	10	11	61	47%
6	Man United	17	9	12	60	37%
7	Watford	17	9	12	60	40%
8	Leeds	16	11	11	59	32%
9	Aston Villa	15	10	13	55	41%
10	Brighton	13	12	13	51	39%
11	West Ham	14	9	15	51	36%
12	Burnley	13	9	16	48	34%
13	Brentford	11	14	13	47	21%
14	Crystal Palace	12	7	19	43	37%
15	Leicester	12	7	19	43	45%
16	Everton	11	9	18	42	37%
17	Wolves	11	9	18	42	35%
18	Newcastle	9	10	19	37	29%
19	Southampton	7	12	19	33	37%
20	Norwich	6	12	20	30	50%
	Average					37%

(b)

#	Team	Wins	Draws	Losses	Points
1	Man City	29	6	3	93
2	Liverpool	28	8	2	92
3	Chelsea	21	11	6	74
4	Tottenham	22	5	11	71
5	Arsenal	22	3	13	69
6	Man United	16	10	12	58
7	West Ham	16	8	14	56
8	Leicester	14	10	14	52
9	Brighton	12	15	11	51
10	Wolves	15	6	17	51
11	Newcastle	13	10	15	49
12	Crystal Palace	11	15	12	48
13	Brentford	13	7	18	46
14	Aston Villa	13	6	19	45
15	Southampton	9	13	16	40
16	Everton	11	6	21	39
17	Leeds	9	11	18	38
18	Burnley	7	14	17	35
19	Watford	6	5	27	23
20	Norwich	5	7	26	22

(c)

Figure 3: Final premier league standings for (a) random forest classifier; (b) Poisson regression; and (c) real standings; *Canva*, 5 August 2023

As shown in Figure 3, the random forest classifier achieved an average accuracy of 42% across all games in the 2021/22 Premier League season while the Poisson regression model only achieved an average accuracy of 37%.

Anomalies

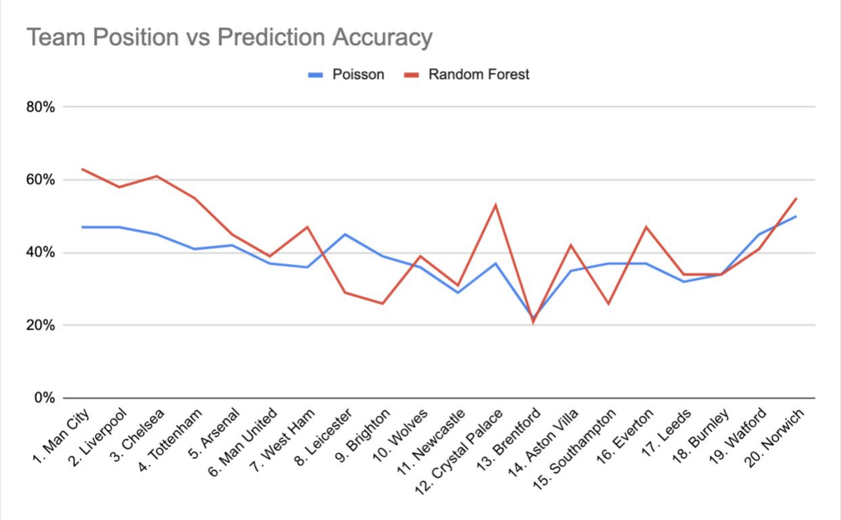


Figure 4: "Team Position vs Prediction Accuracy"; *Google Sheets*, 5 August 2023

Significant external factors impacting clubs must be taken into consideration, and two clubs in particular, Newcastle and Brentford, can be regarded as anomalies. This is reinforced by Figure 4, as the prediction accuracy of both models dropped significantly for these two teams in comparison to their adjacent teams.

As denoted in Figure 3, Newcastle's prediction accuracy was merely 29% and 31% for the Poisson regression and random forest classifier respectively, both performing worse than if someone were to randomly guess their results (33%). This can be attributed to external factors impacting Newcastle during the season which weren't present in previous seasons that the models were trained on. In recent years prior to the 2021/22 Premier League season, Newcastle had been a mid-table to poor team. This held true at the beginning of the season, as they were winless after their first 14 games played (Emons). However, on October 7th, 2021, Newcastle was successfully taken over by the Saudi Public Investment Fund (Morgan et al.). This change in ownership brought a significant influx of money and a change in the coaching staff. Furthermore, during the winter transfer window in January 2022, halfway through the season, the change in ownership allowed them to spend 85 million euros on signings, purchasing extremely expensive and key players to the team ("Newcastle United Transfers 21/22"). Thus, they turned their season around in the second half and won 12 out of their last 18 games. They were able to finish in a higher position than they had in the previous 3 seasons when no previous team had even managed to avoid relegation after going winless in the first 14 games, as Newcastle had (Emons). Though player transfers, coaching changes, and ownership changes are regular occurrences in the Premier League, ones to this degree all in the same season are extremely rare and incompatible with the previous seasons that the models had been trained on. Thus, Newcastle can be regarded as an anomaly.

As presented in Figure 3, Brentford's prediction accuracy was even worse at 22% and 21% for the Poisson regression and random forest classifier respectively, significantly worse than even Newcastle. However, this can be attributed to Brentford being promoted to the Premier League for the first time in club history during the 2021/22 season ("A History of Brentford FC"). Thus, Brentford wasn't present in the training process for both models, meaning that they were unable to give any predictions for Brentford on a real basis. Therefore, all games involving Brentford can be regarded as invalid and treated as anomalies.

After filtering out all the anomalies, the average accuracy of both models improved with the Poisson regression model achieving 40% accuracy and the random forest classifier achieving 44% accuracy.

Data Analysis

The 4% difference in average accuracy between the two models still demonstrates the slight advantage of the random forest classifier overall in accurately predicting football match outcomes.

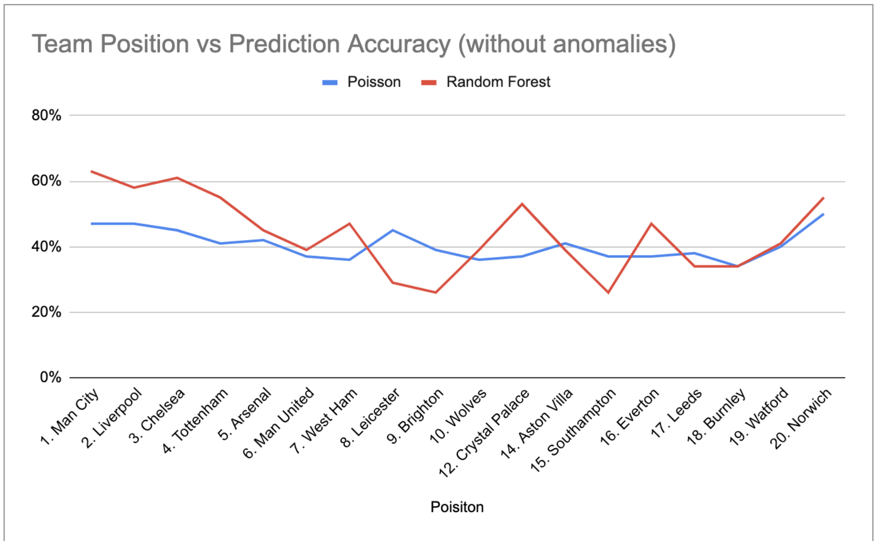


Figure 5: “Team Position vs Prediction Accuracy (without anomalies)”; *Google Sheets*, 5 August 2023

The overall trend of the prediction accuracy displayed in Figure 5 indicates that both the random forest algorithm and the Poisson regression model tend to predict the results of the best teams with higher accuracy. The random forest classifier was able to capture the consistent performance patterns of these dominant teams more accurately than the Poisson regression model, but they both displayed a similar negative correlation, meaning the lower the teams placed, the less accurate the predictions were.

Additionally, both models also predicted well for the worst teams in the league as shown in Figure 5, and they both displayed a similar positive correlation, meaning the lower the teams placed, the more accurate the predictions were. Thus, the models' accuracy progressively got better with each team at the bottom, and both models performed exceptionally well for the worst team in the league, Norwich, achieving accuracies greatly above their average accuracy. Overall, the two models performed at a relatively similar accuracy, with the Poisson regression model slightly eclipsing the random forest classifier for teams such as Watford, while the random forest classifier outperformed its counterpart for teams such as Norwich.

However, Figure 5 demonstrates how both models tended to struggle with mid-table teams, and the accuracy of the predictions for these teams generally trended below the model’s average. The average performance of the two models for these two teams was relatively similar. Interestingly though, the Poisson regression model was much more consistent than the random forest classifier. Even after removing the anomalies, the random forest classifier’s accuracy still varied significantly on a team-by-team basis, ranging from accuracies of up to 53% for Crystal Palace to 26% for Brighton. By contrast, the Poisson regression model’s variance was significantly smaller, ranging only from 45% for Leicester to 36% for Wolves.

Due to the promotion and relegation system of the English leagues, the three worst teams in the Premier League each season are relegated and replaced by three teams in the second league. Thus, teams such as Leeds and Aston Villa, who had only recently been promoted to the Premier League, are only a part of one and two of the previous five seasons respectively, meaning there is much less data for them to be trained on. The Poisson regression model performed slightly better than the random forest classifier for these teams with less training data, achieving two percent higher for Aston Villa and four percent higher for Leeds as seen in Figure 5.

Discussion

The higher accuracies for the best and worst teams are likely because top teams are usually favored in games and win more often than they lose, while the opposite can be said for bottom teams. Thus, both models can predict wins for the top teams and losses for the bottom teams with confidence, resulting in higher accuracy.

The contrasting consistencies of the two models for mid-table teams are likely a result of the stability of coefficients and assumption alignments with the data.

Firstly, the Poisson regression model offers easily interpretable coefficients, directly signifying the effect of predictor variables on the expected tally of goals. These coefficients demonstrate constancy over various model runs on the identical dataset. The openness of the coefficients provides lucidity in comprehending the influence of each predictor on the total number of goals scored, thereby having more coherence in elucidating the associations between factors and results. By contrast, the random forest classifier lacks immediate interpretability owing to its ensemble of decision trees. The model does not present a plain interpretation of how each feature shapes the forecasts. The absence of coefficient steadiness in the random forest poses a greater hurdle in consistently explicating and deciphering the model's predictions.

Secondly, the Poisson regression model's premise of a Poisson distribution for goal counts can be justified under specific scenarios where the data closely conforms to this distribution. In cases where this premise is met, the model is more prone to generating predictable and precise prognostications. In contrast, the random forest classifier does not hinge on rigid distributional assumptions, rendering it more adaptable in managing assorted data types. While this flexibility can be an advantage, it may also lead to varied predictive performance across datasets, particularly if the data distribution deviates significantly from what the model expects.

The Poisson regression model’s ability to predict better for teams with less training is likely a result of its simplicity and alignment with assumptions.

Firstly, Poisson regression's elegance lies in its simpler structure, requiring fewer parameters to estimate, unlike the intricate framework of the random forest classifier. This complexity divergence becomes particularly significant when dealing with limited datasets. Complex models like the random forest, with their intricate decision trees and numerous hyperparameters, are more susceptible to being overly specialized in their training dataset, as they are more likely to memorize noise in the dataset. However, simpler models like Poisson regression allow it to generalize the situation better when dealing with these smaller datasets.

Secondly, the Poisson regression works under the assumption that the number of goals scored matches a Poisson distribution. It relies on specific distributional assumptions to predict the outcome, and the data tends to match these assumptions. Thus, even when dealing with less data, if the model follows these assumptions, it can still predict the results with decent accuracy. However, the random forest classifier doesn’t follow these strict assumptions, meaning it might struggle to capture certain patterns effectively and predict correctly for these teams with fewer data points.

Limitations

The first limitation of this study lies in the dataset. During the peak of the covid pandemic in England from 2020 to 2021, fans weren’t allowed into the stadiums to minimize the virus spreading. The last third of the 2019/20 season and the full 2020/21 season, the two highest-weighted seasons in the dataset, were played with no fans (“How has the COVID-19 pandemic”). Thus, the significance of home advantage is largely reduced over this period as some games were played at neutral stadiums and the home team didn’t have their fans supporting them. Additionally, player performances and availability may have been influenced by the virus.

Future Work

To address the aforementioned limitations of the dataset, future research can focus more on seasons prior to the pandemic in order to negate this concern.

Additionally, the nature of both models means they are unable to take form into account, which refers to the recent performances and momentum leading up to the match. Both models rely on historical data and do not explicitly capture the dynamic nature of team form. While the weight assignments for the prior seasons attempt to account for the importance of team form coming into the season, these weights don't paint the full picture as form can also fluctuate in season on a game-by-game basis. Thus, both models were unable to accurately predict for teams such as Newcastle. If the models could take form into account on a game-by-game basis, it would be able to recognize that during the second half of the season, Newcastle's form significantly improved (due to the change in club ownership) and predict more wins for the team.

To address this, more research can be conducted on the Dixon-Coles model instead of the Poisson regression model. The Dixon-Coles model is seen as an improvement to the traditional Poisson model as it assigns higher importance to recent matches when calculating team goal-scoring rates. This temporal aspect enables the Dixon-Coles model to better account for teams that are experiencing fluctuations in form during the season (Sheehan).

Conclusion

In conclusion, the dynamic and multifaceted nature of football games renders it challenging to predict the outcome, and both the Random Forest Classifier and Poisson regression models demonstrate strengths and limitations in their predictions. As the overall average accuracy of the Random Forest Classifier marginally supersedes that of the Poisson regression model, it can be concluded that Random Forest Classifiers are more accurate than Poisson regression models to a **small extent**.

The Random Forest Classifier does demonstrates significant potential for achieving high-accuracy predictions, as it reached higher performance peaks for top, mid-table, and relegation teams. Although it struggles with matching the consistency level of the Poisson regression model, this potential presents an exciting avenue for improvement. As more research is conducted and more refined algorithms are introduced, this study demonstrates how machine learning algorithms do indeed have the potential to elevate their predictive capabilities to incredible heights, far beyond that of traditional statistical models.

Works Cited

El, Jari. "Historical Developments of Random Forest." *Medium*, 24 July 2020, drjariel.medium.com/historical-developments-of-random-forest-41492deb6737. Accessed 4 Aug. 2023.

Emons, Michael. "Burnley 1-2 Newcastle: Callum Wilson Scores Twice as Clarets Are Relegated to Championship." *BBC*, 22 May 2022, www.bbc.com/sport/football/61453542. Accessed 5 Aug. 2023.

"Ensemble methods." *scikit-learn*, scikit-learn.org/stable/modules/ensemble.html. Accessed 5 Aug. 2023.

"A History of Brentford FC." *Brentford F.C.*, www.brentfordfc.com/en/history. Accessed 5 Aug. 2023.

"How has the COVID-19 pandemic affected Premier League matches?" *Premier League*, 15 June 2020, www.premierleague.com/news/1682374. Accessed 5 Aug. 2023.

Morgan, Tom, et al. "Newcastle United takeover confirmed as £305m deal with Saudi-backed consortium finalised." *The Telegraph*, 7 Oct. 2021, www.telegraph.co.uk/football/2021/10/07/newcastle-united-takeover-live-saudi-buyers-announcement-latest/. Accessed 5 Aug. 2023.

Newcastle United Transfers 21/22. *Transfermarkt*, www.transfermarkt.com/newcastle-united/transfers/verein/762/saison_id/2021. Accessed 5 Aug. 2023.

Poisson, Siméon-Denis. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Paris, Bachelier, 1837. *La Bibliotheque nationale de France*, gallica.bnf.fr/ark:/12148/bpt6k110193z/f6.item#. Accessed 4 Aug. 2023.

R, Sruthi E. "Understand Random Forest Algorithms With Examples (Updated 2023)." *Analytics Vidhyan*, 17 June 2021, www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. Accessed 4 Aug. 2023.

Roback, Paul, and Julie Legler. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in*. E-book ed., Boca Raton, CRC Press, 2021.

Sheehan, David. "Predicting Football Results With Statistical Modelling: Dixon-Coles and Time-Weighting." *GitHub*, 13 Sept. 2018, dashee87.github.io/football/python/predicting-football-results-with-statistical-modelling-dixon-coles-and-time-weighting/. Accessed 5 Aug. 2023.

"Statistical functions (scipy.stats)." *SciPy*, docs.scipy.org/doc/scipy/reference/stats.html. Accessed 5 Aug. 2023.

Visualization of a Random Forest Model Making a Prediction. *Medium*, towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed 5 Aug. 2023.

Yiu, Tony. "Understanding Random Forest." *Medium*, 12 June 2019, towardsdatascience.com/understanding-random-forest-58381e0602d2. Accessed 4 Aug. 2023.