MSc Project - Reflective Essay

Project Title:	Predicting Liverpool FC's upcoming 2021-2022 English Premier league season using Bayesian statistics and Machine Learning
Student Name:	Harry Agyemang
Student Number:	200112345
Supervisor Name:	Professor Thomas Roelleke
Programme of Study:	MSc Electronic Engineering Big Data Science

Overview

The project was successful as I made it in line with my plan and got the outcome from my milestones, predicting the outcome and league tables using Bayesian and machine learning prediction techniques. I used Jupyter notebook to write all my code in python and produce an outcome where I recorded my findings and displayed them in the project thesis. The critical part was that the data I used match results and statistics, extracting the essential information to build a new data frame. Designed to have every team's average attack vs defence and used this information to predict the outcome in the model. In the Jupyter Notebook cells, I also conducted data analysis but primarily focused on Liverpool, creating visualisation plots on the team.

Strengths and weaknesses

Strengths

- Critically analysed both prediction methods and gained a better understanding of football.
- Used Poisson distribution to predict Football match outcomes and scores for future researchers.
- Used machine learning techniques on football data, can apply this to different seasons, and similar projects may follow this.
- Added an optimisation function to improve the model's performance during training and testing to achieve the best results.

Weaknesses

- I couldn't add more match statistics that affected the game, and the focus was on each team's attack and defence stats, but more should be on other features, e.g., shots, possession, and Free kicks.
- The method should have been training, validate and testing as this is the correct option.
- I should have used hyperparameters turning as I should have done more research into it and improve/optimise the model performance enable to get the best possible result.

- Adjusted the train/test split to different values, e.g., 30/70 and saw how the data performed from there.

Future work

- If I had more time, create a dashboard showcasing my predictions for Liverpool FC for their upcoming season, a visual showing the expected outcome and the league table
- I believe this was a good idea to focus on one team. The project can have different domains, such as Manchester City and predicting how many points they will get in their upcoming games for the new season.
- Using either machine learning or Bayesian to predict in-game stats from players, e.g., how many passes will Paul Pogba make in a game.

Critical analysis of the relationship between theory and practical

During this project, I have spent hours studying prediction methods and different ways of classifying match outcomes. One area I found to increase more in was multi-class vs binary, and my project showed both sides of this argument. In terms of prediction methods, the focus was on using Bayesian statistics to predict a future event. Creating a model was tricky as there was not much research on football match outcomes, but there was for Medicine. I used what I've learnt from Bayesian for this project as medicine uses a similar binary classification approach, but for football, I would have to class win as one but draw and lose as zero. The theory behind Bayesian was interesting as I learnt more about the wide range of Bayesian statistics and how Bayesian can predict future events. In this project, I wanted to implement this. Still, practically it wasn't easy because existing models for predicting football matches and the available ones used a python library named pymc3, which wasn't easy to learn. I had tried many ways to adapt this model to the project but to not much success.

Compared to Machine learning, there are hundreds of materials on the internet that supported my project, and researchers had done this before but by using four leagues or on different teams. The theory behind machine learning is exciting and took me a bit more time to understand how to implement machine learning to predict these outputs. Because of my previous knowledge of machine learning, I was already equipped and had ideas for approaching this. Knowing this was a classification problem, I would need to investigate supervised learning techniques to predict a label from my input. Practically this was very difficult as, firstly, I struggled to know which inputs I needed for this task. I studied a bit more on what others had done and found a way to solve the problem of my project by calculating each team's attack and defence, then scaling for the model.

Predicting the premier league table task went well as I produced predictions on both multi-class and binary. I utilised the input parameters associated with each team, giving their best value in assessing their mean score and conceded value and calculating attack and defence. I also changed the FTR categorical feature into a numerical one. The target variables were either Win/Lose/Draw or Win/Not Win. I attached the prediction to the data frame and created a function to count the number of times a team had won/lose/ draw and then developed a standing table of all teams and their points. I would change,hyper-tuning, the models to get the best performance by changing the parameters as they affect the model dramatically. The results showed that Liverpool

would come second behind Manchester city but secure top four champions league positions in another way.

After this period, I trained the model on three machine learning algorithms and computed the f1 score for the training and test set. Later the model was used on the data, and the model made predictions on last season's dataset, which was good results. But the main task for this project was to see how Liverpool Fc would do in the new season. I found a way of doing this by manually putting Liverpool's upcoming fixtures for the new season in a test.csv file and ran the model on the file to the predictions. The model was a success as it predicted the new unseen data and gave probability rates on the likelihood of a win, lose draw outcome for Liverpool. I then exported the results and attached them to the dissertation. The only way to see if my predictions were correct is to watch Liverpool's new season and compare it against the forecast to see how close the model was to the real thing and compare to a bookmaker's predictions.

If given more time for this project, I would have done it differently by discovering more online resources with Bayesian predictions. I would also change the input and output parameters to predict scores and predict player performances in tournaments such as the Euros or the World Cup. I believe this task can be adapted to different scenarios in football and come out with different results which may benefit its target audience, e.g., bookmakers. The timescale affected my project as I balanced four modules each term and had exam periods, which made me pause the process. Having more time on this project would help me spend more time implementing as I ran out of time and started writing my report.

Personal development

I am thankful for this opportunity to study for a postgraduate degree at this university; I have achieved my goals and overcome challenges while doing my masters. Coming from a computer science background, I knew I'd have to maintain my standards and work twice as hard. Over the year, I've made tremendous progress, become more familiar with new technologies, and widen my knowledge with the courses I've learned. This project is particularly challenging not having a Mathematics background didn't keep me from succeeding as I was able to learn new areas which I would never have thought I would. This project has improved my research, programming, and critical thinking skills. Lastly, I'd like to thank my supervisor Thomas for his support over this past year.

Legal, social, and ethical issues and sustainability

Legal issue: For legal issues, Clubs often use devices and data analysis to improve the performance of their players and achieve better results. The recent involvement of technology poses a risk in the use of player data. Football clubs collect large amounts of data this records players movements, runs and stats during the 90 minutes. There are legal issues that may happen with technology monitoring these players. Clubs can watch players, heart rate, distance and sprints made in a game; data collected over a season then it is analysed. This technology gives players the chance to perform better in games. Another analysis is performance-based for scouts to make better decisions on signing a player who has excellent potential. One example of a Legal issue is GDPR, which regulates personal data and can lead to harsh penalties if found to violate. Sports teams will have to obtain consent through player contracts, and clubs must state their purpose

of using the data, e.g., for performance-related reasons. Clubs analyse players biometric and health data, thus requiring consent to process this. Other legal issues include when a player retires or transfers to another club to play. [2]

Ethical issue: Companies like Zone7 have developed a data-driven AI system that works with football teams. The difficulty is players not allowing them to store and collect data on their data. Many football organisations are investing in data; the reasoning includes improving the player's performance, reducing injuries, and assessing the player's talent. Additional data is collected, e.g., during matches and biometrics such as blood pressure. Clubs are trying to create revenue streams by improving their connection with fans and are not considering more about their data and how private it is. Another ethical issue is that clubs put in players contacts that they will have to wear a wrist strap to track their performance, health, and welfare while away from games and training grounds. If clubs are collecting sleeping patterns from players, there is a potential for a line crossed when collecting private data on players. The data presents an ethical issue as clubs could write contracts where it allows them to track biometrics on plays while they are away from the sport.[1]

Social Issue: The existence and accessibility of football data could be argued to encourage social issues such as gambling. This issue affects finances, interpersonal relations, health and wellbeing, and impact vulnerable people the most. For example, this article states that the Scottish premier league made up to three million selling their data to betting firms. With the recent rise in sports betting, and big firms buying data from clubs to fuel this: People can bet on the next goal scorer to goals scored in a game. Betting is a concern as there is a growing trend in encouraging gamblers to continue betting. Charities spread increased awareness of the problems with football betting and its damage to people's finances and lives. While betting companies promote their brands and increase adverts. Betting companies such as Ladbrooks are paying to be sponsored by professional football clubs, generating extra revenue streams. A company named FootballDataCo sells data on teams' behalf. They sell on data rights of clubs in England and Scotland. The betting companies would buy football stats from FootballDataCo to use. Their high-quality data is sold to the big betting firms for turnover. Its two main markets are one supplying stats to media and bookies, and the other collects information using scouts around the world.

Sustainability issue: For sustainability, football clubs follow strict regulations and provide evidence of waste management, plant-based and providing sustaining transport for fans. This data is essential as clubs need to follow guidelines and give this information. It works by clubs being awarded points per category if they have achieved this, whether in their stadium, training grounds or offices. Each club in the premier league is ranked against this. At the top is Arsenal, Manchester City, Manchester United and Tottenham. The aim is to help reduce the impact on the environment and spread more awareness about these problems and the solutions to combat them. [4]

Conclusion

Overall, I am delighted with how my project went. The success of the project came from my way of planning and setting milestones and achieving them. The project helped me a lot as I created mini-goals and worked towards them to achieve tremendous results in my project. One successful outcome was using a machine learning algorithm that performed well on the data to predict Liverpool's upcoming games, and hopefully, my

predictors were suitable. The project includes Bayesian statistics and machine learning which influences predictions with any domain. Still, these techniques are more than adequate for this project and have given good results in predicting match outcomes and points accumulated by each team to see who wins at the end of the season. I have offered a new way of studying a team's upcoming results and using prediction methods such as Machine learning to do it. The only way to determine if my predictions were correct is to watch Liverpool matches in the new 2021-2022 season and compare the results against my forecasts.

References

[1] G. Blackstock, "Football bosses making millions selling the match data used by live bet bookies," Sundaypost.com, 04-Nov-2018. [Online]. Available: https://www.sundaypost.com/fp/the-numbers-game-how-football-bosses-are-making-millions-selling-the-match-data-used-by-live-bet-bookies/. [Accessed: 17-Aug-2021].

[2] A. Nolan and L. Steele, "Sports technology and the GDPR: data privacy concerns in sports analysis," Lexology.com, 24-Jun-2020. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=38355c57-547b-4e76-b575-e600d9519d2a. [Accessed: 17-Aug-2021].

[3] O. Daskal, "Data tracking of athletes tests ethical boundaries," Ctech, 31-Jan-2021. [Online]. Available: https://www.calcalistech.com/ctech/articles/0,7340,L-3890711,00.html. [Accessed: 17-Aug-2021].

[4] D. Lockwood, "How green are Premier League clubs? Tottenham top sustainability table," BBC.com,25-Jan --2021.[Online].Available:https://www.bbc.co.uk/sport/football/50317760. [Accessed: 17-Aug-2021].