Premier League Team Performance Analysis – Machine Learning – Group 68

Team members: Killian Ronan: 18328687, Thomas Dixon: 17341291 and Dylan Storey: 17333135.

Motivation: What problem are you tackling?

We must take into account lots of factors when deciding our project, including accessibility of the raw data we will be using, real world applications of our analysis and our personal interest in the subject matter. We are all supporters of the Premier League which led us to choosing this project, as well as for the reasons above. We decided to analyse the impact of COVID and the absence of fans at Premier League games on different football metrics, including goals per game, predicted goals, impact of home field advantage, team wage bills, bookings and league position.

Dataset: What data will you use and how will you collect it?

There is a huge range of readily available raw data and statistics on Premier League matches all over the web. We plan to use a mixture of data gathered both through scraping and the use of open data sources. There is an abundance of statistics which we plan to analyse through our models as stated above. We plan to parse our datasets game week by game week, to hopefully observe an improvement in performance within our models as our training data grows in size.

Method: What machine learning techniques are you planning to apply or improve upon?

We plan to apply some techniques including Logistic, Lasso and Ridge Regression. We will also use the k-nearest neighbours regression model. We plan to use cross-validation and mean squared error to obtain performance metrics on our training data. These models will allow us to analyse and form predictions based on our gathered data. We will use a reasonable baseline to compare both of our models to.

Intended experiments: What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

Firstly we will analyse our gathered data from the season before COVID using our selected models. Then we will reapply these models using a second dataset from a season under COVID restrictions. We will analyse the accuracy of our models predictions each game week and expect that they will improve as the training data increases in size.

We will train our two models using data from two different seasons. Next, in order to analyse the impact of COVID on the league, we will use these models to further predict the outcome of a new season. Comparing these two models, we can investigate the differences of the impact of our specified input features. We expect the lack of fans during the season under COVID restrictions to have an impact on refereeing decisions and home field advantage.

Finally, we will compare our models from both seasons and analyse the impact COVID has had on team performances and the outcome of games within the Premier League.