Report

Matthew Doherty - doherty.ma@northeastern.edu
Cody Ho - ho.co@northeastern.edu

Problem Statement and Background

For our project, the aim was to find out whether or not there was any validation in the saying ‘never back the early kickoff’, a phrase widely known among football (soccer) fanatics managing their fantasy teams and placing bets on games. The phrase has gained wide popularity across social media, and has become one of the first things taught to newbie users beginning to dip their toes in fantasy or betting. Many seasoned betting veterans don’t even bat an eye when following the phrase, revealing the religious-like extent that the phrase has gained. But what does it even mean? The phrase says that the early kickoff has a greater chance of having an unlikely result. Thus, if you had two teams with seemingly similar odds of a victory, one playing at the 12:30 kickoff and the other playing at the later 5:30 kickoff, following the advice from the phrase would lead you to avoiding the noon kickoff and trusting the comfortable 5:30 kickoff (and possibly help avoid a bad outcome early on ruining the rest of your day). But with so many people religiously trusting this pseudo-myth, there seemingly is no data or factual evidence backing it up. Thus, the motivation behind our goal to debunk the truthfulness of the phrase stems from our shared passion for both the English Premier League and competitively playing Fantasy Premier League (FPL). Revealing and analyzing the statistics behind early kickoffs can empower us fans to make decisions that are more data-driven rather than emotionally influenced.

Introduction to your Data

The data used in our project was collected from Football-Data.co.uk, one of the most well known and trusted football data resources. Football-Data provides datasets for all of the top leagues in the world, including the Premier League where match data can be found from the

2000/01 season and on. The data from each match is sourced from Opta Stats, the official data partner for the Premier League. We used the match datasets from the ten most recently completed seasons, which consisted of seasons from 2012/13 to 2022/23. Each data set includes the teams that played, kick off time, date, goal information, and betting odds collected from multiple leading betting platforms. Opta collects data each match using a three person analysis team. Two analysts are tasked with recording match actions, such as who touched the ball, where they touched it, and how. The third analyst is in charge of making sure what is recorded is accurate. They do this by rewinding the footage and double checking that every piece of information is correct. When the match is finished, it is checked again to make sure the records are totally correct before the data is published. A potential source of bias in the way of collecting data is that some counts may be open to interpretation, like the number of big chances or slight touches. To counteract this, Opta provides a list of definitions for the analyst collecting the data to make sure they can be consistent. Football-data also collects betting data from thirteen different betting companies. The betting odds for games are always collected Friday afternoons for weekend fixtures, and on Tuesday afternoons for midweek games. A possible consideration that may apply could be potential privacy violations when collecting betting data from the variety of betting companies. Since football-data is not the party producing the odds, it may take incorrect odds if the betting sites make last minute changes after the data collection.

To analyze the relationship between kick-off times and results, we grouped matches by the hour of their kick-off. Premier league matches have kick-off times between noon and 8pm. It is important to note that these times are in the local English time. It is also important to note that the greatest number of matches are played at the 3pm kickoff time, and many also at the 8pm kickoff time, resulting in more data for those two times compared to the other kickoff times.

Figure 1 portrays how the number of games per kickoff times is distributed. It is important to note that kickoff time data was only provided from the 2019/20 season and on in the datasets.

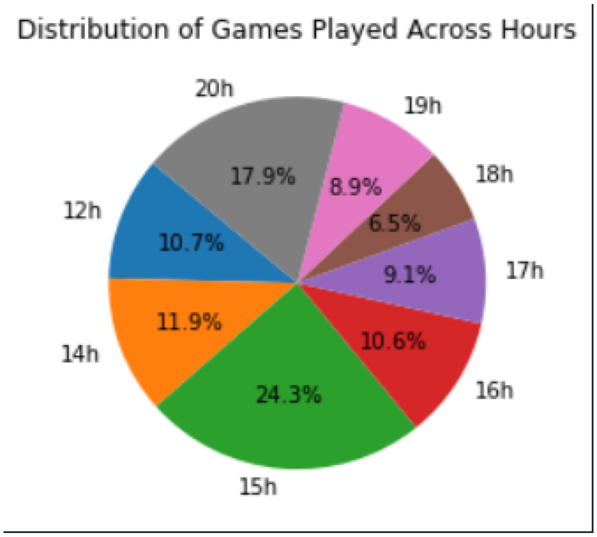


Figure 1: A pie chart depicting the distribution of kickoff times using data since the 2019/20 season.

Data Science Approaches

To be able to create our visualizations, we first had to use data manipulation and cleaning techniques to organize and set up our data in the way we wanted. To do this we had to combine multiple files of data into one dataframe because every season of results had its own data file. We then needed to add multiple columns so that we could use them later on, such as average betting results columns, an upset column that found if an upset had occurred, a kickoff hour column, and a month column. To create the hour and month column we also had to convert the time data to timedate format as well. One data science algorithm we used was the Decision Tree Classifier. We used the decision tree to find what features were most influential for the computer to predict an upset. The decision tree was given the month, time, and both goal betting odds for over/under 2.5 goals in a game. It then would predict if an upset had occurred or not. Then from those findings, we used feature importance to find what features had the most accurate influence on finding whether an upset had truly occurred as seen in figure 2. We also used linear regression to find the relationship between the month a game was played and upset frequency, and we also found the relationship between the time a game kicked off and upset frequency.

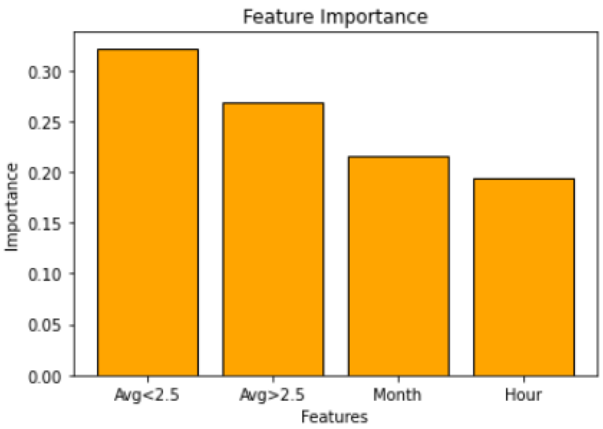


Figure 2

Results and Conclusions

Going into this project we were curious to find out if the saying ‘Never back the early kick-off’ was true and had any validity. This statement essentially means, never bet on the favorite for the early kick-off because an upset will likely happen. The early kick-off is the earliest match played on the game week, typically played at 12:30 on Saturdays. We also wanted to find out whether what month the match was played in had any connection if an upset would occur. We expected that matches played in August, December, and May would yield the greatest frequency of upsets because they mark the start of the season, the most compacted part of the season, and the end of the season respectively. The final aspect we were interested in prior to beginning work on the project was the relationship between goal upsets and kick off times.

When analyzing the relationship between upsets occuring and kick-off time we were surprised. We defined an upset as any time the difference between the betting odds for the real result of the match and the odds for the favorite result was greater than 3. Our findings showed us that the late kick-off (games played 5:30 pm or later) yielded the greatest frequency of upsets. With games kicking off in the hour of 5:00 pm showing the greatest frequency of upsets and games at 6:00 pm showing the least. In terms of frequency of upsets, the early kick off was relatively normal and does not show any reason to not ‘back the early kick-off’, as shown in figure 4. Figure 3 showed that the early kickoff, on average, took more games to produce an upset compared to the other kickoff times, further suggesting that the phrase may be incorrect.

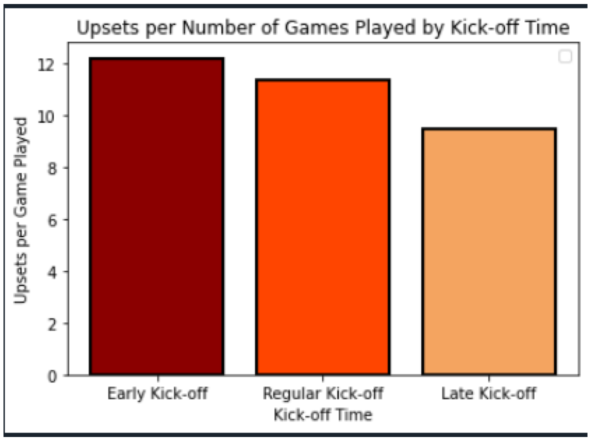


Figure 3

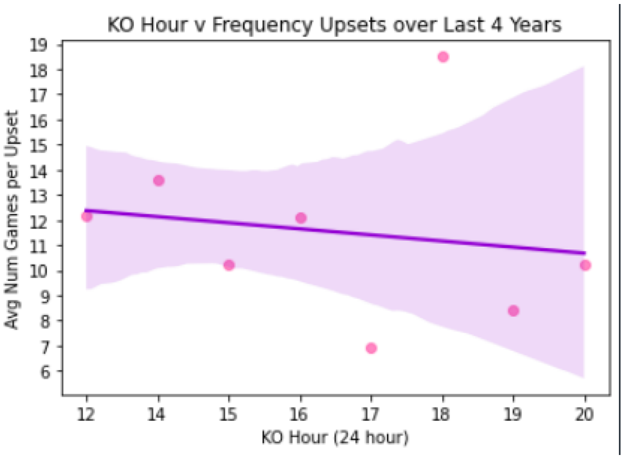
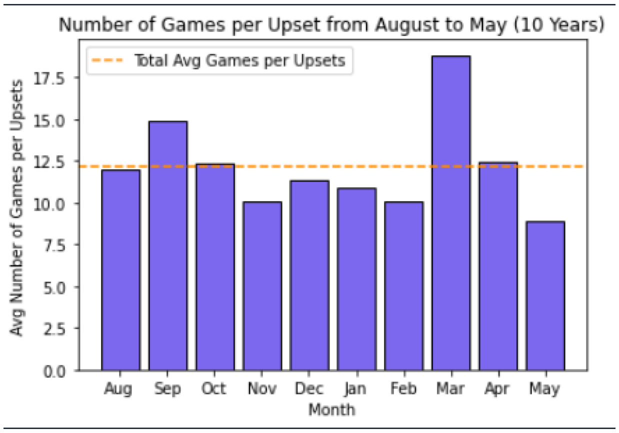


Figure 4

We then looked at the relationship between the month of the match and the frequency of upsets. Our results revealed that August yielded an average frequency of upsets, but May had the highest frequency of upsets: as shown in figure 5 revealing that May produced upsets approximately every 8 games on average. We also found from figure 6 that November to February all had a below average number of games per upset, meaning an upset was more frequent in the winter than in the fall. Our predictions lined up with these findings, except for August which ended up being a somewhat predictable month.



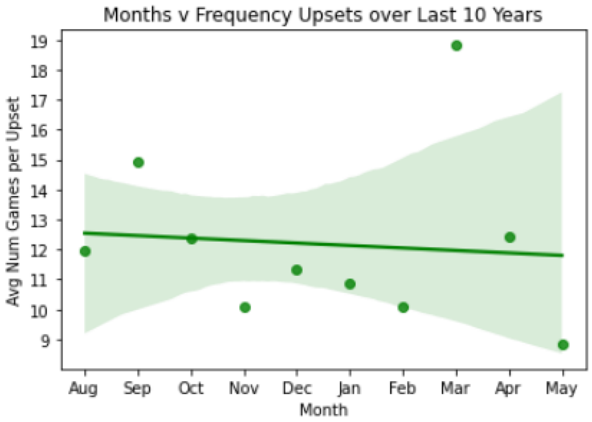


Figure 5 Figure 6

The final relationship we analyzed was kick off times and goal upsets. It is important to note that the data we used for goal betting all came from the 2020/21 season and on, since 2020 was the first full season that Football-Data had recorded extensive goal betting odds. We defined a goal upset as anytime the difference between goal betting odds is greater or equal to 0.8, and when the less likely event occurred in terms of the betting odds. The odds for goal betting are both odds that there will be greater than 2.5 goals in the game, and less than 2.5. A difference of 0.8 shows that there is very clearly a strong favorite, either it be over or under 2.5 goals. For example, if the odds are heavily favoring the game having over 2.5 goals but the match ends up 0-0, that would be a goal upset. We first found the frequency of goal upsets for each time category, then plotted them against each other in a barplot format, as shown in figure 7. The results showed that, on average, there was a goal upset every 7 early kickoffs, very low compared to the values from midday and late kickoffs that yielded over 12 and 14 games. We also looked at the types of goal upsets that were occurring at each kickoff time, and plotted the raw number of overscoring games and underscoring games, as shown in figure 8. For the early kickoffs, there were far more under scoring games, 11 compared to 1 overscoring games. This would lead us to concluding that the early kickoffs produce goal upsets more often than other times, and that there is a high likelihood of the game producing less goals than expected. These results do support the argument to ‘never back the early kick-off’ since it shows that the score lines are often lower than what most would predict. This would advise fantasy managers to use caution when captaining a player playing early on, and for betters to be cautious when betting over 2.5 goals in an early game.

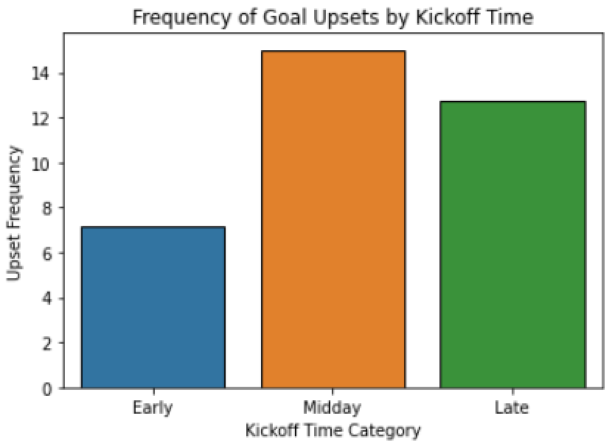


Figure 7

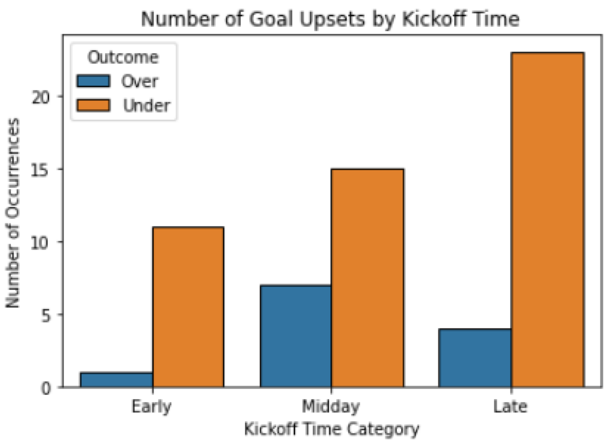


Figure 8

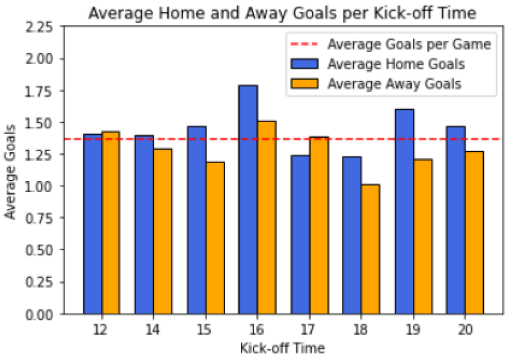


Figure 9

Overall, we learned a lot through this experience. We wanted to find out if there was any truth to ‘never back the early kick-off’ and we found there was some. While the result of the early kick-off did not produce a higher frequency of scoreline upsets, it did produce a higher number of goal upsets. The takeaway for betting enthusiasts and fans is that the scoreline result of the early kick-off may be predictable, but the scoreline that reaches that result may not be. Betting enthusiasts can also learn that the best time of the season to bet on upsets occurring is during the winter or in May. Fans of regularly favorited sides can also prepare for the winter months and May to be the hardest part of their season. In the end, we learned there was some truth to never backing the early kick-off and certain months do produce a higher frequency of upsets.

Future Work

The results of this project have been very interesting and open up many possibilities for future work. The most interesting results we found had to do with goal upsets. There seemed to be a true relationship between early kick-off matches and goal upsets occurings. In future work this is something we would like to continue to look into. Maybe diving deeper into the different degrees of goal upsets that occur and looking at when they occur. Also an interesting point we would like to look into for future work would be looking at the relationship between kick-off time and upsets, but using multiple parameters for different definitions of upsets. In this project, we defined an upset as any time the difference between the betting odds for the real result of the match and the odds for the favorite result was greater than 3. In future work we could change this definition to greater than 1, 2, 3, or 4 to analyze when larger upsets happen and when smaller upsets happen. The final topic of interest sprouted from this project would be to combine our knowledge from this project into analyzing betting data. Diving deep into what type of results betting or losing the most money in, or what upset results pay out the most frequently would yield interesting results and discover useful relationships. Overall, we are anxious to move forward to our next project now knowing what we do, and are excited about where this work could take us.