Ashish Vinodkumar
IDS 702
October 21, 2020

Final Project Proposal

Outline:

I have enjoyed playing and watching soccer since an early age. Specifically, watching Manchester United play in the English Premier League (EPL) became a weekend ritual. There are millions of Manchester United and soccer fans scattered throughout the world who are interested in knowing the outcome of every match or predicting which team would win the entire season. Placing bets on soccer teams has increasingly gained center stage leading to viewers and gamblers meticulously analyzing the odds/likelihood of their team winning the season, in order to place informed bets.

Research Question:

In my analysis, I aim to answer, what are the odds of a team winning the English premier league given the team's offensive and defensive statistics in a given season. The dataset I gathered from Kaggle contains data spanning from 2006/2007 to 2017/2018 seasons. There has been prior research conducted in this space with different datasets where the study sought to answer the odds of a team winning the season using an XGBoost model. However, I plan to answer this question using a Logistical Regression model in R and will further explore if there are distinct groups based on the "seasons" that warrant fitting a hierarchical logistical regression model.

Data:

The data I found on Kaggle consists of 2 csv files called "Results.csv" and "Stats.csv". The results file contains home team, away team, number of goals scored, and match outcome values spanning across all seasons between 2006 and 2018. The stats file contains metrics around the performance of a team in a given season, such as wins, losses, total goals, total yellow cards, total scoring attribute, clean sheet, saves, interception and so on. This data will be used to model and associate a team's performance on their likelihood of winning the season. There are 4560 observations in the results dataset and 240 observations in the stats dataset. Performing an initial assessment of the dataset, I am fortunate to have found a dataset that contains data across all observations. Here is the link to the dataset: https://www.kaggle.com/zaeemnalla/premier-league?select=results.csv

Project Plan:

To answer my research question, I plan to explore a logistical model. I aim to further explore if I need to enhance the model by grouping on seasons and use a hierarchical logistical model. My project roadmap is as follows: Week of 26th: EDA and In-depth analysis of the 2 datasets, including data cleaning and any necessary transformations. Week of November 2nd: Model selection, optimization and create final research paper draft. Week of November 9th: Model assumption validation, further model optimization, and create presentation plus record video. Week of November 16th: Incorporate feedback into model, and update/finalize research paper. Week of November 23rd: Submit final paper.