4741 Project Midterm Report

Tomas Alvarez (ta352), Benjamin Yeh (by253), Rafael Chaves (rvc29)

1 Introduction

For our project, we plan to use several datasets to create a model that can aid in the prediction of soccer game results. To begin the process we explore the Football Data dataset. This dataset contains basic game information and odds sourced from major betting companies.

2 Dataset Description

The Football Data websites, splits its data into csv files by league, and season (year). First we explore the type of data in a typical csv file that you can get off of the Football Data website. Next, we talk about the metadata and where this information is coming from, and finally we discuss some potential issues with the data.

2.1 Types of Data

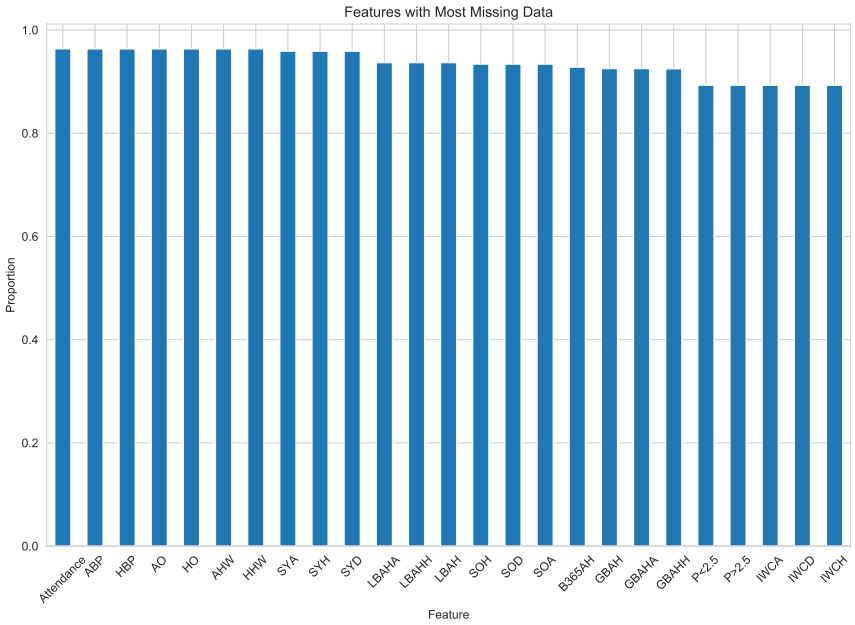
- **General Match Details**: This includes general information such as the team names of those playing, which team is home/away, the date and time of the game, who the main referee is, spectators in attendance etc.
- **Match Outcomes**: This includes what the scoreline was like at half-time (0-0, 0-1, etc.), what the scoreline was like at full-time, and who won (which can also be deduced from the scorelines)
- **Mid-Game Statistics**: This includes the number of each team's shots on target, offsides, corners, freekicks etc. This is information obtained as the match happens, but is not known prior to the match when betting occurs
- **Match Odds**: The standard games odds come in the form of a real number for each of win, draw, and loss. As an example, (4, 3.4, 1.95) means that for every dollar a person bets on a home win, draw, or home loss, they get back 4, 3.4, and 1.95 dollars respectively if their result is correct
- **Total Goal Odds**: There are also sourced odds for the total number of goals scored being over or under 2.5
- **Asian Handicap Odds**: There are Asian Handicap odds which are similar to the win/draw/loss odds but allow for more refined score-line predictions (win/loss by 1, 2, 3, etc.)
- **Market Averages and Maximums**: Since the data encompasses many different company's odds, there are also averages and maximums for every type of odds given

2.2 Metadata Description

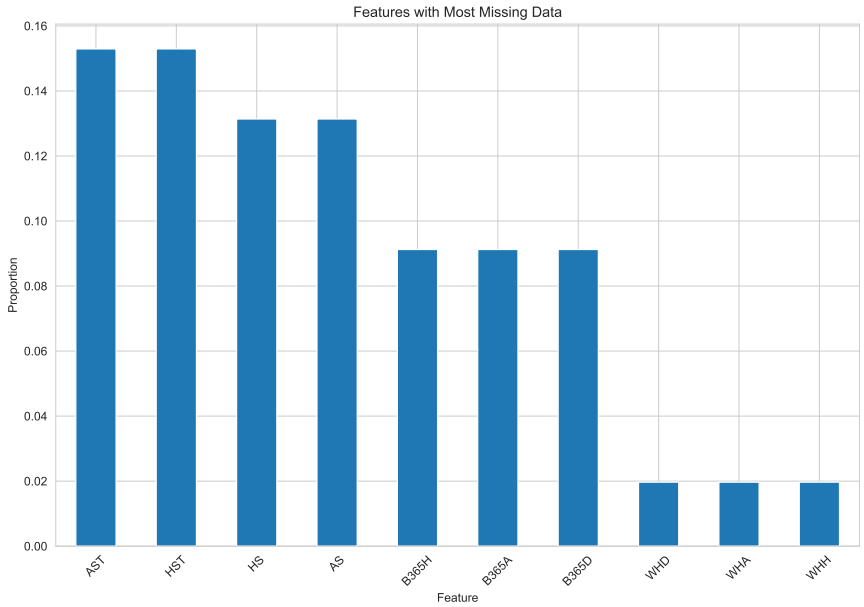
- **Timeline**: The data for most leagues goes back to about the early 2000's but some leagues have records as far back as the early 90's
- **League Count**: The data includes leagues from 27 different countries, with most countries only having a single league, but some as many as five
- **Amount of Odds**: The Match odds are sourced from at most 13 different websites, while the Total Goal and Asian Handicap odds are sourced from at most 4 different websites

2.3 Notable Features of the Data

Right now, we have collected 20 years worth of games from 5 different leagues (England, France, Germany, Spain, Italy), giving us a total of 37,400 examples. If needed, we can pull more data from other leagues or older data. There are approximately 160 columns in the dataset, although the majority are betting odds from different companies. The majority of these are completely missing for the older data, for reasons such as less data collection in older times, access to historical odds from companies, or companies simply not taking bets for a game. To see this, we can plot the proportion of missing values for each of the original features:



Features with very rates of missing values will not be useful to us. For now, we will subsample about 20 features that we feel are predictive of the outcome of a match, and that have a relatively small percentage of missing values. These include the home team, away team, goals for each team, shots and shots on target, and the odds from the betting companies that had the most (oldest data). With this subsampling, our features are much less sparse:



Later, we can add more features from the original dataset, and we can also engineer some features. We will also need to find a way to impute these missing values, which we can do by looking at other data for the home team and away team and estimating it: for example, it takes on average 3 shots on target to get a goal, so we can impute this for examples that are missing the HST/AST features.

3 Further Work

We would like to gather a few more useful features through feature engineering methods. We would like to add the following features before further developing our model:

- **Form**: The average number of points a team obtained in their previous n games (will probably do $n = 5$). Computed by win = 3 points, draw = 1 point, loss = 0 points.
- **Potency/Solidity**: Potency can be calculated by the average number of goals scored over the last n games, and solidity is the average number of goals conceded over the last n games.
- **ELO**: The ELO rating of the home and away team, obtained from here. This will give us a sense of each team's quality rather than just their recent form. We can also consider adjusting the last two features based on the opposition ELO.