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 The Dataset

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

2.1 Types of Data

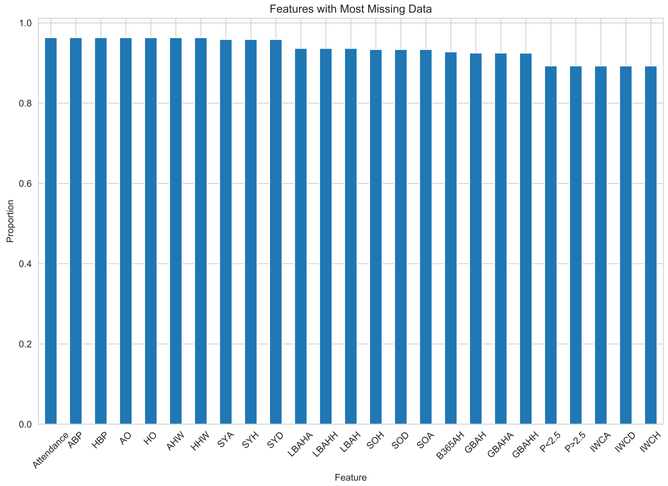
- **General Match Details**: This includes general information such as the team names of those playing, which team is home/away, the date and time of the game, who the main referee is, spectators in attendance etc.
- **Full-time and Half-time Outcomes**: This includes what the scoreline was like at half-time (0-0, 0-1, etc.), what the scoreline was like at full-time, and who won (which can also be deduced from the scorelines), number of each team's shots on target, offsides, corners, freekicks etc. This is information obtained after the match ends, but is not known prior to the match when betting occurs.
- **Odds**: The standard games odds come in the form of a real number for wins, draws, and losses. 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. The dataset also includes odds for Total Goal scored, Asian Handicap odds (a more nuanced betting system), and averages/maximums across all websites.

2.2 Metadata Description

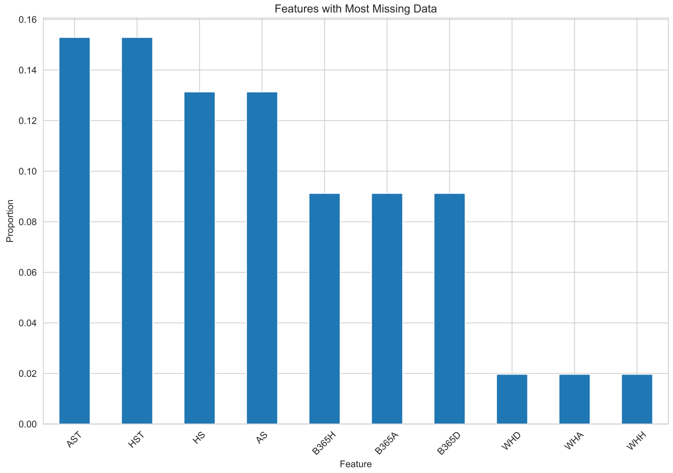
- **Timeline**: The data for most leagues goes back to about the early 2000s but some leagues have records as far back as the early 90s.
- **League Count**: The data includes leagues from 27 different countries, with most countries only having a single league, but some as many as five.
- **Amount of 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 scraped the football-data website for 22 years worth of games from 5 different leagues (England, France, Germany, Spain, Italy), giving us a total of 37,400 examples. There are approximately 190 columns in the dataset with the majority of features pertaining to betting odds from different betting websites. As well, many of the odds are missing for older matches, 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. Specifically, of the approximate 7 million entries in the dataframe, there are about 4.5 million NaN values. To further visualize this, we can plot the proportion of missing values for each of the original features:

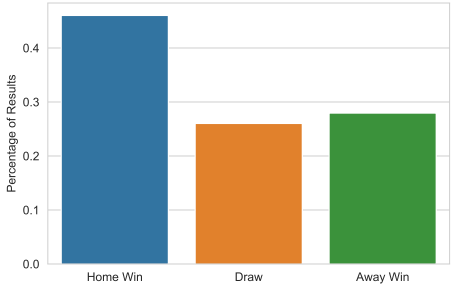


Features with very rates of missing values will not be useful to us. For now, we will subsample about 20 features that we feel are predictive of the outcome of a match, and that have a relatively small percentage of missing values. These include the home team, away team, goals for each team, shots and 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.

Upon further analysis of the data, we can also inspect how the match outcomes are distributed:



As we can observe, home-field advantage is quite prevalent in the data, with 46% of games resulting in a home team win, versus only 28% for the away team. A hypothetical baseline model that always predicts a home win (46% accuracy) gives us a bare minimum accuracy that we can aim to beat, and will also indicate whether our model is underfitting.

3 Model Fitting and Evaluation

To combat overfitting, we intend to utilize regularization to prevent producing overly-complex models that will generalize poorly to our test data. We also intend to use K-fold cross-validation with our more complex models for hyperparameter tuning and model selection. We will likely experiment with a few different models (e.g. Logistic Regression, Bagging, Neural Networks), and for each, we will make sure to limit the model complexity to a certain extent in order to reduce variance. For example, we will constrain the max depth and number of features for the tree ensembles.

Regarding metrics for the evaluation of our models, we will use accuracy scoring to determine the quality of our predictions.

4 Features and a Baseline Linear Model

4.1 Feature Engineering

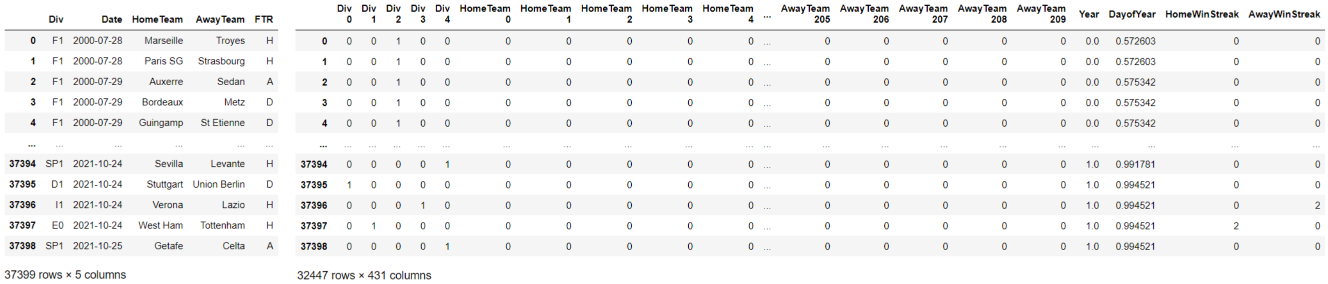
Preceding model development, we defined an even smaller subset of features to consider for experimentation: 'Division', 'Date', 'HomeTeam', 'AwayTeam', and 'FTR', the full-time result of the match and target response. We also chose to ignore any available mid-game details, as by our project objective we intend to predict the outcome of a game *before* the match begins, not after it has already started.

Because all the features in our selection were nominal and non-real values, it was necessary to encode the data into real values so that it could be appropriately passed into a model later. This was done by one-hot encoding the features 'Division', 'HomeTeam', 'AwayTeam'. As well, 'Date' was deconstructed into two new features, one indicating the year(e.g., 2000, 2012, etc.), and a second indicating the day of the year(1, 27, 365, etc.). The last transformation to encode our dataset to real values was label encoding our response from [Away, Draw, Home] ⇒ [0, 1, 2].

For feature development, one simple feature we created was 'Home/Away Team Win Streak'. Within a season, this feature would indicate the winning game streak that the home and away team were on up to the current match. An important note though, is that this feature only indicates the teams win streak for *home* and *away* matches, not the *consecutive* win streak including both. While simple, the reasoning is that the better performing team to date - be it home or away - is more likely to win the game. In this sense, we are measuring performance between the two teams by how many consecutive home/away games they have won at match time. In addition, as it was evidenced earlier, there also appears to be a home field advantage with respect to the outcome of a game. Therefore, capturing whether the home team is also on a home win streak may be a further indicator of the outcome of the match.

The final step in feature engineering before moving to build a first model and *after* splitting the dataset was to perform feature scaling. Following feature engineering, our resulting data frame for learning is sparse with the exception of the features 'Year' and 'DayofYear'. It is important to scale these features because during training, large magnitude features can dominate over other feature weights and adversely influence gradient update directions as well.

Below is the transformation of the dataset before and after feature engineering and splitting.



	Div	Date	HomeTeam	AwayTeam	FTR
0	F1	2000-07-28	Marseille	Troyes	H
1	F1	2000-07-28	Paris SG	Strasbourg	H
2	F1	2000-07-29	Auxerre	Sedan	A
3	F1	2000-07-29	Bordeaux	Metz	D
4	F1	2000-07-29	Guingamp	St Etienne	D
...	...	...	...	...	...
37394	SP1	2021-10-24	Sevilla	Levante	H
37395	D1	2021-10-24	Stuttgart	Union Berlin	D
37396	I1	2021-10-24	Verona	Lazio	H
37397	E0	2021-10-24	West Ham	Tottenham	H
37398	SP1	2021-10-25	Getafe	Celta	A

37399 rows × 5 columns

	Div 0	Div 1	Div 2	Div 3	Div 4	HomeTeam 0	HomeTeam 1	HomeTeam 2	HomeTeam 3	HomeTeam 4	...	AwayTeam 205	AwayTeam 206	AwayTeam 207	AwayTeam 208	AwayTeam 209	Year	DayofYear	HomeWinStreak	AwayWinStreak
0	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0.0	0.572603	0	0
1	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0.0	0.572603	0	0
2	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0.0	0.575342	0	0
3	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0.0	0.575342	0	0
4	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0.0	0.575342	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
37394	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	1.0	0.991781	0	0
37395	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1.0	0.994521	0	0
37396	0	0	0	1	0	0	0	0	0	0	...	0	0	0	0	0	1.0	0.994521	0	2
37397	0	1	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1.0	0.994521	2	0
37398	0	0	0	0	1	0	0	0	0	0	...	0	0	0	0	0	1.0	0.994521	0	0

32447 rows × 431 columns

4.2 Multinomial Logistic Regression

We developed a basic l2 regularized Multinomial Logistic Regression model fit over the 'entire'(considering all leagues and teams) dataset. The model was trained over 18 seasons from 2000 to 2018, and the remaining 3 seasons 2019 to 2021 were reserved for testing(an approximate 80-20 split). The resulting model produced an accuracy of 60.99% over the training seasons and 57.43% over the 3 test seasons, beating out the baseline model that only predicts home wins with 46.51% and 42.85% accuracy over the same split. The Logistic Regression model and its accuracy score will serve as a benchmark to beat for further model exploration and development.

5 Further Work

For remaining work, we would like to gather a few more useful features through feature engineering methods. specifically, we would like to add the following features before further developing models:

- **Last Match Result**: Indicates which team won in last facing.
- **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.