3.1 Objectives

The study aims to discover the relevance of odds provided by betting companies when used as means of prediction. The first objective is to analyse whether it is possible to use machine learning algorithms and odds to predict the results of matches. Secondly, this study aims to discover if the odds provided by multiple betting companies have become more accurate in predicting the end result of a match over the course of eighteen years and eighteen seasons of football. Another objective of this study is to determine in which of the five major European leagues, the odds were the most accurate thus highlighting which league is the most predictable or unpredictable out of the five leagues present in the dataset. To achieve the results required to determine each of the objectives mentioned, Machine learning techniques such as decision trees and neural nets were used to obtain each of the objective mentioned prior.

3.2 Dataset

The dataset for this study consists of the five major leagues in Europe (English Premier League, French Ligue 1, German Bundesliga, Italian Serie A, and Spanish La Liga). The first season to be included in this study is the season of 2000/2001 while the last season is last years’ 2017/2018 season for each of the leagues mentioned. The dataset consists of all the matches, results and odds (Directions of Odds being the standard, Home, Draw & Away directions) on the matches provided by 10 different betting companies spanning over the 18 seasons mentioned (Number of betting companies whose odds are available for a match is not always the same but is dependent on the season and/or league). ELO ratings of clubs - a rating that was developed to assess the skills of a chess player created by Arpad Elo and then adapted to football by considering factors such as a team’s old rating, goal difference, and home team advantage, were also added as extra variables in order to further aid in the model. The ELO rating of a team was not of a fixed value (rating could both increase and decrease) for a whole season and differed depending on team performance in that season.

3.2.1 Dataset Collection

Match results and odds were downloaded manually from football-data.co.uk. A file structure was set up in a manner that a folder was present for each league. Having done so allowed separation of the files according to their league. Using MS SQL Server, a database that consists of five tables (League, Country, Match, Team, and ELO) was created to store all the relevant information from the previously downloaded files. Multiple stored procedures to handle diverse operations were created. Such procedures include searching a team to see if it is present in the database, insertion of a team if it is not found by the procedure previously described, insertion of ELO rating and the insertion of matches and the odds related to it. ELO ratings were downloaded from api.clubelo.com. An application to parse the data consisting of the odds and matches was developed using C# and Visual Studio. The application first looped through a given directory to find and store the folders containing the odds. The folders found were then checked and depending on the name, the country was set accordingly, as this variable was to be used later. Using a foreach iteration, the application then loaded a file into a string. Since the data downloaded was in CSV format this was converted into JSON using choCSVReader a method within ChoETL, a framework that is specifically designed to read, load and transform CSV data. After the data was transformed into JSON, using NewtonSoft’s Json.Net, a framework in .NET that is used to handle JSON the data was parsed. First, a check whether both home and away teams existed in the data was done using the stored procedures previously mentioned. When a team was not found It was added into the database by calling the relevant stored procedure. After that, the match would be inserted into the database with all relevant information such as the date of the match, season, relevant league and the odds provided by each of the betting companies. Another application very similar to the one discussed was developed to fetch, convert and parse the ELO ratings. The teams present in the database were fetched using a stored procedure and were stored in a list. For each of the teams present in the list, a call to api.clubelo.com was made with the teams' name. The response to such call returned the all the available ratings for the club. Response from the API then went through the same procedure as the previous application, storing all the relevant information in the database by calling the relevant stored procedure. In the end an SQL query was produced to create the full data model. This query selected all the relevant information including the season, teams involved in the match and odds available on the match for a full season and/or league based on the parameters given.

### 3.2.1.a

Csv File

3.2.2 Ethical Considerations

Since the aim of this study was to highlight how accurate the odds provided by betting companies are, and which of the betting companies discussed in the study is the most accurate, the names of all the companies involved were replaced with fictitious ones. This step was taken in order to avoid any possible business harm that could be caused to any of the companies that were included in the dataset should the results from the study present any high disparities in the level of accuracy between said companies.

3.3 Variables

3.3.1 Independent Variables

Since the dataset includes all the matches, their results and the odds of 18 full seasons, there were a considerable number of independent variables. The main structure of the data model consisted of all the matches of a season and all the bets (All companies) available and one for each season available was produced. A data model was also produced that contained every one of the leagues involved. To be able to determine the accuracy of each betting company individually, for every single one of the models produced (for each season), each of the company’s odds was tested. Every model was first tested without the ELO ratings. ELO ratings were added afterward as means of enhancement to the model, this allowed to determine the accuracy of odds alone, and if any variables alongside odds could be a help produce better prediction model.

3.3.2 Dependent Variables

Since the aim of this study is to determine whether odds have increased in accuracy over the years, the one and only dependent variable in this study was the Full-time Result (named FTR in the Dataset). The Full-time Result is predicted by using the independent variables previously discussed within machine learning techniques.

3.3.3 Constants

Models for this study vary in season, league and betting company being assessed. Therefore, it was not possible to have a fixed constant. For each of the data models produced (i.e. A model for all the leagues in the season 2017/18 and a model for Serie A for the same season), the matches within the model always remained the same, making them the constant of that specific model. This means that if the odds of the companies on the Serie A season of 2017/18 were being assessed for their accuracy, the same amount and exact matches were considered when the accuracy of a single betting company was being assessed for the same season.

3.4 Implementation

To obtain the results required to be able to achieve and complete each of the objectives listed, machine learning techniques such as Decision trees were used. These techniques were implemented using R studio and several relative libraries. First, all the relevant data model files were created (A data model file for each season, and for each league Available) in CSV format. An R file containing each of steps and tests made for each season and for each league was developed.

3.4.1 Data loading, Cleaning and Dummy Variables

Firstly, the data model file was loaded and odd related columns were converted into numeric values using R’s inbuilt function as.numeric(). Any NULL values were taking care of using by using na.exclude() a function within R studio that excludes any rows within the data set where a NULL value is present. Dummy variables were created to replace all the teams involved in matches for both the home and the away teams. This was achieved using the dummies library and the function dummy from within the same library. For every team, a column was produced using this function. Values within the columns created were 1s and 0s. A 1 represented that the team was involved in the match, while 0 meant that the team was not involved in that match. These dummy variables were then added to the data model. The original home and away team names were removed as they were no longer needed since the dummy variables replaced them.

3.4.2 Classification Tree

As previously discussed, to achieve the objectives, the accuracy of bets each season, and each betting company for each season needed to be discovered. To achieve this a classification tree was created for each of the variant possible (i.e. A classification tree for the entire season of 2001/2002 and A classification tree for a certain company for the same season). To sample the data, a single random sampling approach was taken where 60% of the model was taken to train the classification tree, achieved by using the sample() function. The remaining 40% of the rows within the model were used for testing. Using library C50 the classification trees were created using the C5.0() function. The C5.0() function requires the full training model (except the variable to be found), the variable to be found (in this case - FTR) and trails (amount of time the function is rerun. Trails are done to further enhance the models' ability to correctly classify the data). The process was done for each of the betting company present within the data model to be able to determine the accuracy of a single company. To achieve this accuracy, the training set was fed the bets of one single company and hiding the rest of the companies while retaining the dummy variables (Home and Away Teams) generated in the previous section. The classification tree was then plotted using the plot() function, where a diagram of the tree was given. The model that included all seasons had a rather large tree and was hard to read. To be able to read and understand what was happening within the classification tree, nodes within the tree had to be printed. This was achieved adding sub= in the plot and the node number, this produced a plot of the node and the nodes that are related. To further understand the plot of the classification tree, the function summary() was used. This function prints out on the R console several details. The structure of the classification tree, such as the nodes and classification from the resulting nodes. Attributes from within the data model passed that had the most affect on the tree are also displayed by this function allowing better understanding of the tree. To predict the dependent variable, FTR, the predict() function was used. Using the training set that consists of the 60% of the data created in the step prior and also the test set model (without the FTR column) the function tries to correctly classify the FTR into the 3 factors that exists within the model ( H – for Home Win, D – for Draw & A - for an Away Win). To compare the results obtained by the predict function and the actual results in the testing set a cross table was produced. Both the rows and columns of the cross table are the values of FTR (H, D & A). For each of the FTR values, one can see the number of correctly classified values as well as the wrongly classified. (I.E one can see how many matches whose FTR value is A were wrongly placed both in D and in the H value. The same applies for the other values). Percentages are also produced for each of the previously discussed.