Data warehousing and data mining report

# student details

### Name(s): cathal hughes & russell brady

### Programme: case

### Module Code: ca4010

### Submission Date:

### Module Coordinator: Mark roantree

# Plagiarism declaration

We declare that this material, which We now submit for assessment, is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of our work. We understand that plagiarism, collusion, and copying are grave and serious offences in the university and accept the penalties that would be imposed should I engage in plagiarism, collusion or copying. We have read and understood the Assignment Regulations. We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the source cited are identified in the assignment references. This assignment, or any part of it, has not been previously submitted by me/us or any other person for assessment on this or any other course of study.

We have read and understood the referencing guidelines found at http://www.dcu.ie/info/regulations/plagiarism.shtml, https://www4.dcu.ie/students/az/plagiarism and/or recommended in the assignment guidelines.

Name(s): Russell Brady & Cathal Hughes Date: 24/10/2018

table of Contents

[student details 1](#_Toc528517339)

[Name(s): cathal hughes & russell brady 1](#_Toc528517340)

[Programme: case 1](#_Toc528517341)

[Module Code: ca4010 1](#_Toc528517342)

[Submission Date: 1](#_Toc528517343)

[Module Coordinator: Mark roantree 1](#_Toc528517344)

[Plagiarism declaration 1](#_Toc528517345)

[Introduction 2](#_Toc528517346)

[Idea and dataset description 2](#_Toc528517347)

[Idea 2](#_Toc528517348)

[Dataset Description 3](#_Toc528517349)

[Data Preparation 4](#_Toc528517350)

[Algorithm description 4](#_Toc528517351)

[Results and Analysis 4](#_Toc528517352)

# Introduction

The following report provides a detailed explanation of our idea and dataset, how we prepared our data, the algorithm we used to make our prediction and our subsequent results and findings. Also included is our insights on how we felt the process went and our general thoughts on the results and predictions with which we obtained.

# Idea and dataset description

Soccer is the most popular sport in the world and is played by 250 million players in over 200 countries and dependencies. The premier league is arguably the most popular league in the world with huge financial strength, TV coverage and entertaining matches. There are also vast amounts of data available in datasets online relating to the premier league with everything from match results to the number of throw-ins in a game. Because of this we decided it would be an appropriate area to apply data mining techniques and make predictions.

## Idea

Our idea was to take the statistics and results from the last number of years in the premier league, clean the data to obtain only the attributes we wanted, pass this data through an appropriate algorithm and generate a prediction for the result of a given game based on this. Our hope was to predict whether a game would end in a home win, away win or a draw.

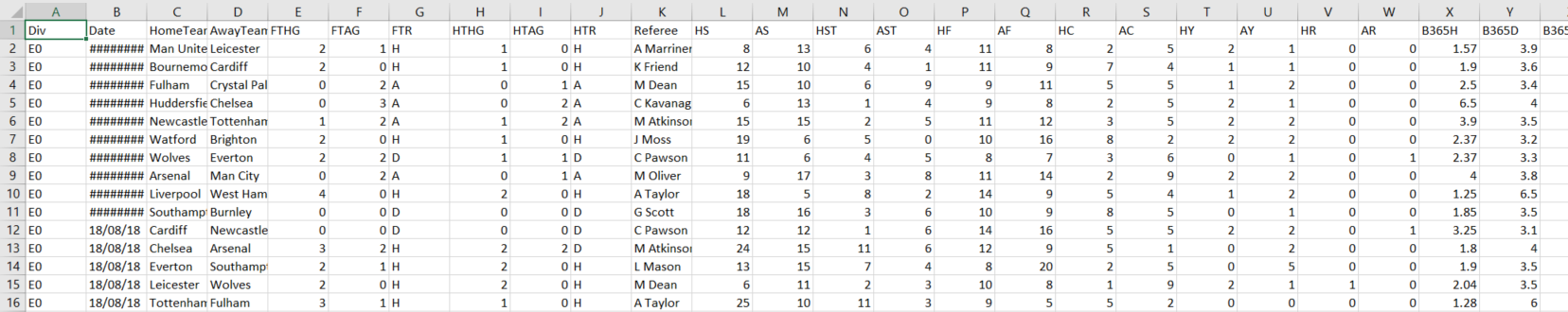
The reason we decided to take on this idea is that there is a real-world application. Betting is a huge part of all sports, soccer being no different, with up to one billion euros being placed on each premier league games globally according to the Tribune. Our idea was that if we could generate predictions with a reasonable accuracy there would be a possible monetisation opportunity. With a reasonable prediction accuracy for example, you could have a bot place smart bets across a host of games to try and make money. You would only place bets across games where the odds, along with our predictions would indicate you are going to at least make your money back. Obviously, this is all theoretical but we were curious to find out the possibilities and that is why we went with the idea.

## Dataset Description

Although there was plenty of data available, it became apparent that we weren’t going to be able to obtain a perfectly clean standardised dataset. This meant we were going to have to generate our own dataset using cleaning techniques on existing data for premier league games and generate our own data for premier league standings for the relevant years.

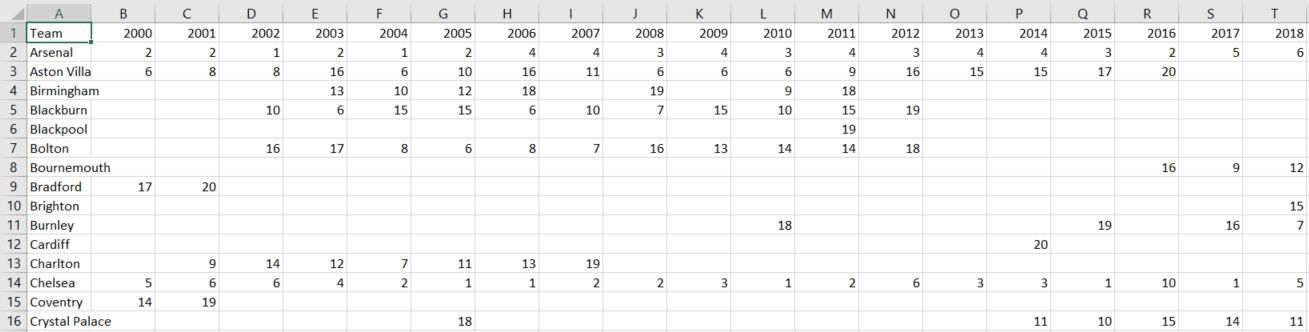
### Premier League Games Dataset

These datasets were obtained from www.football-data.co.uk/data.php. Each dataset was split up into its respective season and contained stats about all the matches in that given season. Each dataset contained a different number of attributes some contained 28 while others contained 65, some relevant and some not. Also, there was 380 rows per dataset since there are 20 teams, each playing each other twice during the course of the season. We collected the datasets from the 2000/01 season to the 2017/18 season. This meant we had 6840 rows of data to work with. The picture below is a snippet of some of the pre-cleaned data.



### PL Standings Dataset

We also used a dataset pertaining to Premier League standings for the relevant years. This is a dataset we generated ourselves. It was arranged alphabetically by team and contained the teams finishing place for the last 18 years in the Premier League. The dataset included the 46 teams which were involved over the course of the last 18 years. This dataset also contained multiple empty values for newly promoted teams and teams that haven’t been in the Premier League in years e.g. Coventry. This dataset would need to be incorporated with the above datasets and all empty values would need to be handled appropriately. The picture below is a snippet of our PL Standings Dataset.



### Generated Dataset

After concatenating all the datasets and incorporating the Premier League Standings dataset we then had to perform some major cleaning on the dataset. This will be discussed in greater detail further on in the report.

The quantity and differing numbers of attributes in the PL Games Dataset was difficult to deal with and was something that would not be feasible to work with. Some of these attributes related to in-games stats and bookies odds. We had no use for in-game stats as it would be pointless to make a prediction on a game in the middle of one. Although, the bookies odds could have proved useful and could have improved our prediction accuracy, we did not have the bookies odds for much of the dataset and thus we could not use them to make our predictions. As such we decided that we wanted to make our prediction solely based on the two teams and their form. The results would not be swayed by any third party and all predictions could be made using some basic stats of the two teams.

There are some attributes which weren’t relevant on their own but proved useful when used in conjunction with others meaning we had to generate some of our own attributes which will be explained in greater detail below. After cleaning, we were left with a dataset which contained 6840 rows and 12 attributes.

These attributes were the following - Home Team Points, Away Team Points, Difference in League Position, 6 columns containing team form over the last three games for both teams, home and away team goal difference as well as difference in league points. The target label for the dataset was Full Time Result which is one of these three values - Home Win (H), Away Win (A) or Draw (D). This was a clear classification problem.

Generating our final dataset to be passed into our algorithm was a challenge in terms of deciding which attributes to use, and then extracting these attributes. Our data contained categorical labels and only some of the data was standardised, meaning how we cleaned the data was very important. Once we generated the final dataset we then split it up into a training and testing set. Our training set contained 17 years of data from 00/01 to 16/17 and our testing set contained the most recent full season, the 17/18 season. Although this was not an optimal split it allowed us to make predictions for one whole season.

# Data Preparation

All the data was contained in CSV files and all the data cleaning was performed using python. Tools such as Pandas and ScikitLearn were used for manipulating and concatenating the data.

We separated the cleaning up into three different scripts namely - cleaning\_data.py, final\_dataset\_cleaning.py and data\_preperation.py.

## cleaning\_data.py

## final\_dataset\_cleaning.py

## data\_preperation.py

# Algorithm description

# Results and Analysis