Predicting the outcomes of English Premier League matches
Computer Science – Part II – Progress Report

Joseph Marchant (jm2129)
Robinson College
30 January 2019

Project Supervisor: Helena Andres-Terre **Overseers:** Prof. A. M. Pitts

Director of Studies: Prof. Alan Mycroft Dr R. Mantuik

Project Aims

The aim of this project is to be able to predict the outcome of premier league results given various statistics and meta data about the teams before the match. It is a 3-way classification problem, predicting the outcome from {H,D,A}. A prediction accuracy of 40% would be a successful project, and to rival experts in their success rate an accuracy of 52%+ would be a very successful project.

Success Criteria

The initial success criteria was to reach a 40% prediction accuracy on the following problem: Given data that is only known before a specific Premier League (PL) match, predict the outcome from the three options: ”H” (Home Team Win), ”D” (Draw) and ”A” (Away Team Win).

My latest iteration achieves an accuracy exceeding 48%, thus I have achieved my success criteria. There were still multiple sources of delay, mainly involving indecisiveness on what data to use and how to represent it, as well as issues with messy data.

Work Completed

- **Data Collection** - 11v11.com and prem.com required scraping from over 10000 pages scraped per site (1 per match). football-data.com supplied 26 csvs covering lots of meta data. I also got data from various websites and wikipedia using scraping and manual collection.
- **Data Cleaning** - Main issues were missing data and disagreeing values between data sets. Made more difficult by chronological style of data where some data points depend on past data points, meaning data needed to be perfect.
- **Initial Dataset** - I wrote code to generate the PL standings given a set of games, giving me meta data about the teams in every match. The final dataset covered 5 statistics: Position, Goals Per Game, Goals Conceded Per Game, Points Per Game and Clean Sheets Per Game. Each was represented as the difference = (Home Team

Stat - Away Team Stat), giving 5 features. The target variable was the outcome from H,D,A.

- **Models and Evaluation** - Worked with Logistic Regression on a binary representation of the problem: "Did the home team win?". I.e. H = 1, D,A = 0. This achieved a 60%+ accuracy. I then built a Neural Network (NN) with Tensorflow. The best performing one achieved a 49% accuracy on the H,D,A problem. I also generated evaluation metrics for each model such as Accuracy, Precision, Recall and a Confusion Matrix.

I am around 2-4 weeks behind the initial schedule, however I hit my success criteria earlier than expected due to an effective first approach. Therefore I will be slightly altering my work plan.

New Work Plan

1. **Lent weeks 2–4: Jan 31 - Feb 13** Decide on around 10 new features for data. Collect data, clean and merge with current data. Test out data on a couple simple (non-NN) models.
2. **Lent weeks 4–6: Feb 14 - Feb 27** Build a more complex NN model for this new data. Apply tweaks and optimise both the model and the data set.

Rest the same as before (building a full evaluation report, then mainly focusing on the dissertation and extensions where possible).

New Extensions

After seeing the progress on the project, there are some alternative/ extra extensions I would like to work towards:

- Split data based on the difference in league position between the two teams. This will allow 2 different models to be created, picking up different patterns in the data depending on the game stats.
- Experiment with building two models that solve the binary problem "Did this team Win, yes or no?" for each team, then predicting the 3-class classification based on these.
- Experiment with Feature Selection Using Permutation Importance.
- Generate some more advanced Evaluation Metrics on the final system such as the ROC Curve and looking at Precision vs Recall.