Exploring Machine Learning Prediction for Soccer Outcomes

Ryan Baker
University of Utah
u0724394@utah.edu

Nathan Wilkinson
University of Utah
u0295388@utah.edu

ABSTRACT

We attempted to create a predictive model for soccer games based on detailed statistics scraped from the web. Our hypothesis was that by the use of detailed statistics we could make more accurate predictions that those based on simple statistics like win percentage for each team. We found that... . Limitations. Future Work.

1 INTRODUCTION

Sports betting is a common activity among sports enthusiasts around the world. In this setting, it is desirable to be able to predict the outcome of matches between two given teams. Typically, sports analysts and experts gather rich sets of data on each and every game for individual teams, and sometimes gather data on individual players. In this project we will be focusing on the problem of predicting the outcome of soccer matches, which can be used in the sports betting arena. While it can be difficult for humans to accurately predict the outcome of a match, we explore whether or not it is possible to predict outcomes accurately using machine learning techniques. While it is applicable to sports betting, it is also interesting to consider whether or not machine learning can predict outcomes, even in the presence of intangible elements of the game, such as team chemistry, home-field advantage, winning- and losing-streaks, player transfers and injuries, and other factors.

The methods utilized in this project explore a variety of algorithms, such as support vector machines (SVM), logistic regression, bagged forests, and perceptron. A large portion of this work is exploration with features, such as comparing raw statistics against derived features. We also explore the divisions of training and test data, such as comparisons between chronologically split data and randomly split data. The main lessons we learned from this project are the following:

Feature selection and the relevance of data is extremely important and can significantly impact machine learning model performance. The selections of concept class and hypothesis space are important aspects machine learning pursuits and can heavily impact the performance of a model. That is, the bias introduced by the best hypothesis in the space can be a significant limiting factor and must be monitored and controlled carefully.

2 RELATED WORK AND BACKGROUND

Sports betting is a common activity among avid sports fans in all sports, including soccer, basketball, football, baseball, and many other sports. To facilitate betting, many organizations publish public betting lines that describes which team is favored to win a game, and how much they are favored to win a game. However, many, or perhaps even most, of these models are strictly statistical and probabilistic. We seek to apply machine learning. Our goal is to focus on soccer, particularly in the United States, and to explore how different algorithms may produce better results, and if different train/test splits on data have an impact on learning results (e.g. explore if it is more useful to split data randomly or linearly within the course of a single season).

We have explored several different studies of machine learning with sports betting. These studies explore data from NCAA (college) basketball in the United States [4], NFL (professional) football in the United States [3], and EPL (professional) soccer in England [2]. These reports have utilized Bayesian methods [2, 4], decision trees and random forests [4], and Gaussian processes [3]. An important common aspect of all of these studies is that they recognize that using simple, raw statistics is not nearly as useful in machine learning algorithms. That is, they recognize the need for alternative features to help represent matches. This includes factors like home

field advantage, winning- and losing-streaks, and player injuries. We will continue to explore these conclusions as we seek to form our feature representations and identify our desired algorithms.

3 DESCRIPTION OF DATA

The data used in this experiment was scraped from the web from the Audi soccer data store [1]. The data from each match is identified by a unique id specified in the url. The data was compiled by the company Opta, which gathers detailed statistics on sports matches. After being scraped from the web, an SQLite database was used to aggregate and consolidate the data to allow it to be transformed into per-match feature vectors.

Once the data was in feature vectors as a CSV file, more high-level feature extraction was performed. We extracted summary statistics for each game that averaged each team's statistics over the past games from each season. Then, we generated statistics that summarized a team's performance during the season like total wins, ties, and losses and average points per game. In soccer, a win is worth 3 points, a tie is worth 1 point, and a loss is worth 0 points. We also added some columns that could be used for the output of the prediction like "did the home team get a result (win or tie) or not?" and "was the outcome a win, tie, or loss for the home team"?

We seperated the features into home and away because we suspected that home-field advantage would play a large role in determining who won games. We also subdivided the data into files by season and league, for example, Major League Soccer for 2016 or English Premier League for 2017. In total, there were 3312 games and X (403?) features from which the train and test sets could be made. However, we explored fitting models to different subsets of the data such as games from a specific year or leagues.

4 METHODS

5 EVALUATIONS

6 FUTURE WORK

7 CONCLUSION

ACKNOWLEDGEMENTS

This report was completed for the course project in CS 5350: Machine Learning at the University of Utah in the Fall 2017 semester. We would like to acknowledge the help of Prof. Vivek Srikumar who taught us the principles and tools necessary to complete this project. Much of the implementation of different algorithms for this project was completed as part of assignments throughout the semester, and was modified to fit our needs for this project.

REFERENCES

[1] Audi Soccer Data Store. audi-player-index.com/en/getMatch/<id>/latest/. (????).

[2] Anthony C Constantinou, Norman E Fenton, and Martin Neil. 2012. pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems* 36 (2012), 322–339.

[3] Jim Warner. 2010. *Predicting margin of victory in nfl games: Machine learning vs. the las vegas line.* Technical Report. Technical Report.

[4] Albrecht Zimmermann, Sruthi Moorthy, and Zifan Shi. 2013. Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. *arXiv preprint arXiv:1310.3607* (2013).