Capstone Project Summary

Sports Match Outcome Prediction Modeling

MSc in Business Analytics, Central European University

Shahana Ayobi

June 2023

Table of Contents

1. Introduction

The client for my capstone project is an emerging sports business intelligence consultancy that specializes in football. The consultancy is aiming to become the leading provider of comprehensive insights, analysis, and benchmarking for football clubs, and leagues. Therefore, for this project, I am aiming to build a prediction model that would accurately predict football match outcomes to ensure that the consultancy provides its clients with a competitive edge and help them make dynamic data-driven decisions. This project seeks to develop precise prediction models that consider potential financial risks associated with skewed odds.

The main predictors used are home and away team expected goals, team total market value, home and away team attacking strength, audience attendance, B365 betting odds, and time of the match, and the outcome variable is match outcome with home win, draw, and away win values.

2. Data Engineering

The raw data for this analysis come from three different websites including Football.co.uk team statistics, FRref for expected goals and form of the team statistics, and Transfermarkt for team market values. The project takes into account three Premier league seasons from 2019/2020-2021/2022, the total number of matches played are 1140. The datasets go through several data preparation processes such as dropping empty rows, removing symbols and abbreviations, ensuring measurement uniformity, and standardizing team names. The datasets are then merged using unique identifiers, and new variables such as home and away attack strength are created, missing values are imputed, and flag variable is added for the attendance variable, variable correlations are examined and correlated variables are dropped from the model.

3. Modeling

To ensure accurate predictions, the data is split into training and test sets using a 20% to 80% ratio. This allows the model to train on a larger portion of the data and evaluate its performance on unseen test data, preventing overfitting and assessing its generalization capabilities. Accuracy is chosen as the performance metrics for this project to which serves as an objective benchmark to measure the models' predictive accuracy and their ability to minimize erroneous predictions.

Then three types of machine learning models are run that are multinomial logistic regression, random forest, and gradient boosting (GBM). Two types of models are run for these algorithms, one with only the attendance as explanatory variable and the other with all the above mentioned variables. The multinomial model predicts multiple outcome classes (home win, draw, and away win) simultaneously, while random forest combines decision trees to make predictions. The random forest model is trained with 500 trees, and the stopping rule and maximum depth are set to prevent overfitting. The GBM algorithm combines predictions from multiple decision trees and adjusts for unimportant variables. These modeling approaches aim to optimize prediction accuracy and account for complex relationships in the data.

4. Result Analysis

The complex multinomial logit with all of the 12 variables gives the best result with 64% accuracy on the test set followed by random forest with 59% and GBM 57%. The variable significance analysis showed that expected goals scored by the home and away teams had the

strongest impact on the predicted probabilities explaining 40% of the model performance, followed by factors such as market values, attack strength, and match time. The complex multinomial logit model provides better generalization and interpretable coefficients, and the comparison of predicted odds with the odds provided by the B365 bookmaker showed reasonable performance, particularly for home odds. However, some discrepancies were observed in the away team odds, which could be attributed to the model's omission of current market dynamics considered by bookmakers.

5. Recommendations

These recommendations are provided in order to better improve the accuracy and reliability of predictions, and enable more informed decision-making in the field of sports betting. The consultancy can leverage automated Natural Language Processing (NLP) or fuzzy matching techniques to ensure consistent team name standardization and facilitate data merging from multiple sources since the data is not available in one source. Exploring additional data sources, such as weather and referee statistics, player injuries, and market betting dynamics can capture unforeseen factors that impact match outcomes. Comparative analysis of different leagues and identifying league-specific predictors will enhance the generalizability of the models.

6. Learning Outcomes

Through this project, I have gained valuable knowledge on utilizing machine learning algorithms for predicting football match outcomes in a business setting. It has allowed me to navigate the complexities of handling diverse data sources, address inconsistencies in the data, and conduct in-depth evaluations to assess model performance. Additionally, it showed the dynamic nature of the betting market, demonstrating the need for continuous monitoring and adaptation.

7. Summary

In conclusion, this capstone project aimed to build accurate prediction models for football match outcomes in a business context. By using machine learning algorithms, including multinomial logistic regression, random forest, and GBM, the project achieved an accuracy of 64% on the test set with the complex multinomial logit model. The project highlighted the significance of selecting relevant predictors and considering the dynamic nature of the betting market.