Predicting goal scoring probabilities for football matches

Capstone Project Summary

Zsombor Hegedus

8th July, 2021

Executive summary

This paper summarises the work done for a startup that operates in the sports betting industry, specialized in football. The client has a web-based product that offers real-time football match statistics and analytics for sports betters. As part of my Capstone project, I was tasked to come up with a model that could supply their current infrastructure with probability estimates concerning the end score of matches (e.g. how likely is it that the home team will score less than 2 goals). One crucial challenge that I needed to tackle is updating the probability estimates in real-time considering the progress of a given match.

I experimented with three approaches in which the first one is based on goal event probability modeling, the second experimented with survival regressions. The last one combines unsupervised algorithms with statistical distributions to deliver probability estimates. The client deemed the third approach to be the best for the use case after which I developed a model that executes this in Python.

This project summary discusses the posed challenge and the solution provided. It briefly introduces the data used for model building and highlights how I contributed to the client's goals. It summarises the key outcomes and briefly touches on the areas where I improved thanks to this opportunity.

Task introduction

The client asked me to develop a predictive model that can provide real-time probability predictions for the following outcomes:

- The home team will score less than 0.5 goals, 1.5 goals, 2.5 goals, 3.5 goals
- The away team will score less than 0.5 goals, 1.5 goals, 2.5 goals, 3.5 goals

Pre-match probabilities were calculated by leveraging a commonly used principle which is to use the Poisson distribution to predict the probabilities of discrete events. However, the

challenge with this task was to factor intra-match statistics into the calculation. For example, if one of the teams receieved a red card, it should deteriorate their chances of scoring goals.

Two data sources supported my work:

- **Historical intra-match data** - this holds data that is specific to each football match. The granularity of this is in minutes.
- **Team profiles** - these are statistics on the historical performance of a given team. Every observation in this dataset is a team at a specific time.

Modeling

The main challenge of my capstone was modeling. I first cleaned, transformed, and ordered the data to conserve the time-series-like property of match observations. After that, I carried out exploratory data analysis to find interesting patterns that would be crucial in the model building phase. Eventually, I started experimenting with the preprocessed data.

First of all I employed a machine learning approach and viewed the use case as a binary classification problem where the outcome variable used for training is a dummy that states whether a goal will be scored or not. I trained the following models: k-nearest neighbor classifier, logistic regression, random forest and XGBoost. I used 5-fold cross-validation with a randomized CV search of 10 iterations. I also did hyperparameter tuning and tried to tackle the problem of unbalanced classes (as the ratio of positive and negative cases was around 1:30). The approach allowed me to identify attributes like corners and shots on goal to be important from a goal-scoring perspective.

Furthermore, I experimented with survival regression models. The goal was to develop a meaningful model that indicates the time required for a goal to be scored. Cox Proportional Hazard Model was used which has a Python implementation in the `lifelines` package. Not only can the model factor in aggregate survival curves of the population but the features of individual observations as well. Even though progress was made, the approach was not pursued further due to its inablilty to produce estimates on the end-result distribution.

The key final solution

Both approaches provided an essential experience that was leveraged for the final solution. Without revealing confidential details, the developed model provides a distribution estimate with the help of unsupervised algorithms and known statistical distributions. It uses match features such as the number of shots landed on goal, corners, red cards. It also leverages team profiles which are mainly descriptive statistics related to the historic performance of a given team (for example, how many goals they scored in their last 10 matches on average).

The EDA phase revealed that teams in different leagues have significantly different goal-scoring averages and that those goal events are not distributed uniformly throughout the games. The final model also accounts for these differences; the match minute variable is also used as a feature, and modeling is done separately for the different leagues.

In a more detailed technical documentation, I highlighted several ideas that could be used for quantitative (Chi-squared and Kolmogorov-Smirnov statistical tests, elbow charts) and qualitative validation (focus groups, deep interviews). Moreover, I highlighted the potential risks and fields for future improvements. The code that executes the above is packaged and pushed to the Github repository of the client.

Benefits for the client

My work has contributed to the client's success in two ways.

First of all, the client wishes to enhance their brand visibility with the help of content marketing. To help with this goal, I wrote and published an article on Medium.com using the data provided by the client. This article is an exploratory data analysis that also introduces the rich database that the client owns.

Secondly, the client wants to provide unique analytics on their website that gives them a competitive advantage over their competitors. To contribute to this goal, I developed a feature that provides real-time intra-match probability prediction for the potential number of goals a given team is expected to score. The goal of this feature is not to recommend strategies, but to show probability estimates on the outcomes. Those estimates can be used by sports bettors to tailor their betting strategy to their risk appetite.

Lessons learnt

Working on this Capstone project was a great experience for me as it showed something invaluable; the importance of understanding the usability of the created product. One of the cardinal factors for real-life business use cases is usability business value to the stakeholders. In this Capstone project, I needed to respond to a business need and dvelop my own solution/model that required a lot of research. This was a new type of challenge which helped me grow as a professional in numerous ways.

The project also helped me get better at presenting. I introduced complex statistical formulae or models to an audience who might be technical but still required time to adjust to my coding style and solutions. The client was very welcoming and provided me with a lot of support in this area as well.