Title: (T7) Football Bets
Members: Artjom Valdas, Eduard Rudi

Business goals

- ## Identifying our business goals

  - ### Background

Today, bookmakers provide many different bets on sports, including football. Such companies always give out three (usually different) bet odds - one team wins, the other team wins or both teams play a draw. Since these companies will never work to their own detriment, sometimes they can specifically increase the coefficient for a strong team and lower it for a weak one, thereby a person who does not understand teams and statistics is likely to lose his money.
Some people know how bookmakers work and therefore they do not trust any coefficients, but simply analyze football teams manually (including previous games, the number of goals, points in the standings). Based on their own analysis, people can make their own decisions and place a bet on a team that, according to the company, should lose.But such an analysis takes a lot of time, especially if a person wants to bet on several different teams.

  - ### Business goals

Our main goal is to automate the analysis of football teams in the English Premier League, to bring out interesting and important analytics for schedules and to create a model that will most likely predict the outcome of future meetings between the two teams. It would also be interesting to find out what factors influence the victory or defeat of a team and whether these factors exist at all.

  - ### Business success criteria

The result will be considered successful if the model can predict the outcome of 8 out of 10 games. In this case, we can assume that the created model works very well. If the program can still find some factors that affect the outcome of the game, then the goal will be achieved at 110%.

- ## Assessing our situation
  - ### Inventory of resources

For this project we need - the programming language Python (specific libraries for statistics, machine learning and plotting), Jupyter notebook, data on today's and previous seasons of the English Premier League, bookmaker company on which we can compare the results(https://www.betsafe.ee/et).

Requirements, assumptions, and constraints

Since the data is in the public domain, we don't have to sign any documents for nondisclosure. Therefore, we do not need any permission to use the data and we can begin to analyze it now.

Risks and contingencies

The project may not be completed on time, because there may not be enough computer resources to calculate the probability of victory of one team over another. This problem can be solved by several computers by distributing the load between them. Also, not enough data can become a problem, but this is not dreadful, because the same data can be taken from other football leagues.

Terminology

Home Team - the team that plays in the home stadium
Away team - team that plays in the guest stadium
Half Time Home/Away Team Goals - number of goals scored by the home/away team in the first 45 minutes
Full Time Home/Away Team Goals - the number of goals scored by the home/away
Home/Away Team Shots on Target - number of successful shots on goal
Home/Away Team Shots - number of shots delivered towards the goal
team for the entire match (two periods of 45 minutes)
Probability of victory - percentage, with what probability team will win

Costs and benefits

This project does not have a specific goal to earn, but it can be easily fixed. If the model will really predict very well, then for the sake of interest it will be possible to try to replenish the account at some kind of bookmaker and bet on teams that will have a high percentage of victory.

Defining your data-mining goals

Data-mining goals

One of the goals is the creation of a model that, based on football data, can predict the victory or defeat of a team. For this we need to analyze all the previous matches of both teams and also analyze the previous meetings of these teams. Also one of the main tasks is to translate data into a human-readable form using various graphs and try to find the pattern between victory and defeat and the factors that affect them. If such a factor exists, then one can go further from it and make predictions based on the found reasons.

Data-mining success criteria

Data-mining will be successful if the model can predict the outcome of the match with an accuracy of 80%. And of course, it would be nice if the model was still able to find signs that interfere or, on the contrary, help the team win.

Gathering data

Outline data requirements

The most basic data we will be needing are following:

1. The first essential data is of course, what teams are playing. If we do not know which teams are playing, how are we going to separate, who won and which statistics belong to.
2. The second list is, who is the home team and the away team. We have looked at some datasets and more precise datasets, also tell who is the home team and the away team. It is not that crucial, but it is nice, because with this, we can tell how much home stadion affect winning, because your home and everything is familiar and thus it is more relaxing or is it the opposite, because your home, you have pressure on your shoulders to not disappoint the home crowd.
3. We will be needing also, who won the match. If we do not know who won the match, how are we going to make any decisions.
4. We also would like to have the odds for the match, it can coefficient or just home team wins, away team wins or draw. With that we may find some interesting correlations.
5. For last are some statistics about the match, i.e. referee, shots taken by teams, penalties and yellow/red cards given. The are for finding some interesting correlations.

For the most part, all of the datasets we have seen, are all csv format files. Also, the datasets can be found till 90s and even later. So we do not have to worry about not being enough data. But we have decided to take initially three seasons (three years) worth of English Premier League data.

Verify data availability

As I have said in the last bullet point in the end, there is a lot of data for football betting. It is because betting has been always part in human history. And since the Internet has become part of our life, someone has gathered and combined data. Because of that, we can see for example https://www.football-data.co.uk/englandm.php here, that we can get data as early as 1993.

Define selection criteria

We will be using https://www.football-data.co.uk/englandm.php datasets and initially 2017 - 2018, 2018 - 2019 and 2019 - 2020 season for England Premier League as there are England league matches are quite colorful and there many matches.

Describing data

The source of the data is https://www.football-data.co.uk/englandm.php and the file formats are csv. There are 893 different cases and each case has 63 columns. There are 5 fields that describe the data, those are league division, date, time, the home team and the away team. Other columns are match statistics and different kind of odds. All of the fields what we need are there so we do not have to look for other datasets. The most important columns are home team, away team, who won and odds.

Exploring data

Every season of data have beginning the same. They have date and time, teams, goals, who won and also game statistics. Where the difference comes, are in odds section. At the moment we are not sure if the columns are completely different or is it just the name. If it is just a name, then we can easily change it, if not, we have to think about it, how we will implement them or just completely drop them. One minor thing, that I have not mentioned, is that the column names are mostly in code or shortened. We have thought about it and we are most likely to change them into something that is more understandable. There are a lot of extra columns we do not need them at the moment, so we are very likely to drop them as well.

Verifying data quality

The data should be good and accurate, because they are getting their results from www.xscores.com and their statistics from websites such as BBC, ESPN Soccer, Bundesliga.de, Gazzetta.it and Football.fr.

Planning our project

Tasks

- Data processing - there are more than 60 columns in our three csv files, of which we will not need at least half. To do this, we need to understand exactly what we need, in order to remove everything that is not needed. (Each member ~2h)
- Creaty various schedules - in order to conduct some kind of analysis, starting from our data, it would be nice to visualize them. In such cases, many things will be easier to see and understand. (Each member ~5h)
- Create a prediction model - one of the main goals of this project is to predict the outcome of a football match. To do this, we will have to translate the data into numbers and use one of the training models that we took on the course. (Each member ~10h)
- Using statistics and probability theory to make own prediction for the match - this way we can find out how accurate the prediction model is. (Each member ~10h)
- Find which factors influence the outcome of the match - if in fact something has an effect on winning or losing, then this dependence can be used for predictions (Each member ~3h)

Methods and tools

We are going to use the Python programming language in Jupyter notebook. Most likely we will use matplotlib, pandas, numpy and sklearn libraries.