Prediction of Premier League results

IDS2021_D9

Oliver Savolainen, Kaur Vali, Märt Tender

Project GitHub repository:
https://github.com/OliverSavolainen/IDS2021_D9

Business understanding

Identifying your business goals

Background

Football is a sport whose match outcomes many people inside the field or in the betting industry want to predict accurately, but no one has found an adequate solution.

Business goals

In this project, we will try to achieve accurate predictions based on data from the most popular league in the world, the English Premier League. Our goal is to find the best model to predict Premier League results, specifically the number of goals scored by a team and additionally find out which factors impact results the most.

Business success criteria

We will likely determine the success of this project by comparing our results to some popular sports betting sites and finding out if our model can outperform these sites and make a profit.

Assessing your situation

Inventory of resources

- Datasets found on GitHub and *fbref.com* and Kaggle (see links in Verifying Data Availability section below)
- Time, which will be spent on the project
- Knowledge about the subject of the project (football)
- The data science skills acquired through the Introduction to Data Science course
- Jupyter Notebook, Python, Pandas etc.
- Laptops

Requirements, assumptions, and constraints

Schedule:

- 30.11 FBRef data gathered
- 04.12 Dataset completed, features engineered

- 08.12 Models trained, best performing one chosen
- 10.12 Goals scored calculated into probabilities and odds and compared to actual odds
- 13.12 Poster and video completed

The only requirement for acceptable finished work is that the created model makes reasonable predictions about the goals scored in future matches, where reasonable is defined by Oliver Savolainen as he is the most knowledgeable of the project team members in this field.

Risks and contingencies

- A model takes too long to train.
    - Train the model on a faster computer.
    - Modify the model to one which can run on a regular computer.
- Some relevant factors are too complex or even impossible to include in the model.
    - Create less specific features and combine said feature with another one.
    - Presume said factor will not affect the result of the model too drastically.
- Covid-19 pandemic affects the results of the model.
    - Leave out the previous year from the model.
    - Include match attendance in the training data.

Terminology

- Football - team sport, where the target is to kick the ball into another team's goal more times than the opponent.
- English Premier League - the most popular football league in terms of viewers in the world, includes 20 teams from England or rarely Wales.
- Goal - ball kicked into the net.
- Goals scored (GF) - the number of goals by a team.
- Goals conceded (GA) - the number of goals by the opposing team.
- Pass - one player kicking the ball to another player on the team
- Assist - pass that results in a goal scored
- Penalty kick - a situation where rules are broken inside the team's penalty box (rectangular box with a length of 16 meters), and an opposition player can shoot a shot from a spot 11 meters away from the goal with only the goalkeeper allowed to stop the shot.
- Wins, draws, losses - win if goals scored by the team is more than goals conceded, draw if those numbers are equal, loss, if goals scored, is less than goals conceded.
- Expected goals (xG) - Very simply, xG (or expected goals) is the probability that a shot will result in a goal based on the characteristics of that shot and the events leading up to it. Some of these characteristics/variables include:
  Location of shooter: How far was it from the goal, and at what angle on the pitch?
  Body part: Was it a header or off the shooter's foot?
  Type of pass: Was it from a through ball, cross, set-piece, etc.?

Type of attack: Was it from an established possession? Was it off a rebound? Did the defence have time to get in position? Did it follow a dribble?

- Non-penalty expected goals (npxG) - expected goals minus penalty kick expected goals (penalty kick is worth 0.75 xG)
- Goal difference - goals scored minus goals conceded.
- Expected goals scored, conceded, difference - see goals scored, goals conceded, goal difference with excepted goals instead of goals.
- Also, see https://fbref.com/en/statsbomb/
- Home and away team - home team is the team, who's stadium the match is played at, away team is the opposition.
- Attendance - the number of people at the stadium (usually supporting one of the teams).
- Odds - the probability of something happening converted to decimal values of 1.00 and above; see calculator here https://www.aceodds.com/bet-calculator/odds-converter.html.
  More advanced statistics and terminology from fbref.com could be used but are not currently defined here.

Costs and benefits

Costs:

- Time.
- Energy.
- Nerves.

Benefits:

- A completed final project for the Introduction to Data Science course, hopefully, accompanied by a passing grade.
- A potential advantage when betting on future Premier League match results.
- More knowledge in machine learning and model prediction.

Defining your data-mining goals

Data-mining goals

The goal is to have easily analysable data to create different models based on which we can make predictions about future match results. We want to make a presentation that compares our predictions to those of other companies and the actual results of matches.

Data-mining success criteria

Data mining will have been successful if our model can accurately predict the results of matches to rival the predictions given by betting offices and companies.

Data understanding

Gathering data

Outline data requirements

Everything defined in the terminology except for football and English Premier League will be used as features in the data, with both data from the season before (since 2016/17 season) and data from a number of previous matches from the same season included as all of these factors should play a role in deciding the outcome of a football match.

Verify data availability

Examining the first chosen Github repository, it seems that the data is almost purely player based, while we want data about each match. For the last 3 seasons, this data about each season exists, but it seems pretty difficult to work with and it does not exist for the first 2 seasons we want to include in our predictions. This means we need to find new data sources. The first one is https://www.kaggle.com/josephvm/english-premier-league-game-events-and-results, which has data up until this month's games with information about the match, the attendance, scores and players that played. Also, we will try out data from https://www.kaggle.com/louischen7/football-results-and-betting-odds-data-of-epl, because it includes betting odds from each match, but this dataset is not complete as it ends in the middle of last season. At the same time, last season might be an outlier, so not using it might become useful.
Another problem has occurred examining data from fbref.com, as it seems data about expected goals does not exist from the 2016/17 season, btu we can substitute in data from https://understat.com/league/EPL/2016, but according to previous knowledge the expected goals model from this site might not be as accurate as the one on fbref.com.

Define selection criteria

We will be using the matches.csv file from the first Kaggle link mentioned in the last paragraph, the files marked by the season numbers from the second Kaggle link and data about teams combined from https://understat.com/league/EPL/2016 and fbref.com. The data will be about seasons ranging from 2016/17 season to 2021/22 season and fields included will be defined in the next paragraph.

Describing data

One row of data includes matchweek, date, team name, whether the team is the home or away team, attendance, statistics from fbref.com from the previous season described in terminology, same statistics about 5 and/or 10 last matches of the team, the opposing team and the same statistics about that team, a column about, whether the team's best player is playing (according to awards given by the team the previous season) and finally goals scored by the team aka the target value we are predicting. If we are not happy with results on our validation dataset, then additional columns from chosen datasets could be used. If

the described row is about the home team of the match, next will be about the away team from the same match.

Exploring data

Looking at each relevant dataset, no empty fields can be found with the exception of attendance in the first Kaggle dataset matches.csv file, where some fields are empty because attendance was 0 due to covid-19 restrictions. Other problems have been solved in the previous tasks (or mentioned in the next task). According to Oliver's previous knowledge about the subject everything seems to be in expected ranges, so all the data can be used to train models.

Verifying data quality

We have more data than we think is necessary, but currently, it is scattered in many different files. We need to choose the required features and put them together into one dataset. There does not seem to be any missing data aside from the odds data from last season (and this season), but this can be solved by using a site like https://www.oddsportal.com/soccer/england/premier-league/results/, which provides odds from each match.

Planning your project

Task 1

Title: Create this pdf
Description: As part of the 10th homework of the Introduction to Data Science course, create this pdf file to be a starting point of the project and to outline the project's goals and strategy for reaching those goals
Assignee(s): Oliver, Kaur, Märt
Estimated time spent per person: 6 hours
Tools / methods used: Google Docs, Facebook Messenger

Task 2

Title: Prepare data for analysis
Description: Combine the multiple datasets into one big data frame with all of the needed features and not too many unnecessary features.
Assignee(s): Oliver, Kaur, Märt
Estimated time spent per person: 12 hours
Tools / methods used: Jupyter Notebook

Task 3

Title: Train models model on the dataset to make predictions
Description: We start by using some learning model algorithms and testing/validating them. We need to do Task 4 also and then try to see if we can find a better model.

Assignee(s): Oliver, Kaur, Märt
Estimated time spent per person: 5 hours
Tools / methods used: Jupyter Notebook

Task 4

Title: Use the model to predict results and check their accuracy
Description: We take the predictions from the model and test them on a validation dataset to see how good our predictions are. If they are satisfying, then we use them, otherwise, we will try task 3 again to find a better model.
Assignee(s): Oliver, Kaur, Märt
Estimated time spent per person: 5 hours
Tools / methods used: Jupyter Notebook
Notes: Repeat tasks 3-4 until the created model is satisfactory.

Task 5

Title: Compare our model to the predictions by prediction sites
Description: The task is to determine whether our model is comparable to the models used by prediction sites and how our model predictions differ from other predictions. To do this, it is necessary to find the odds by the prediction offices.
Assignee(s): Oliver
Estimated time spent per person: 3 hours
Tools / methods used: Jupyter Notebook
Notes: The predictions of prediction sites are included in one of our datasets.

Task 6

Title: Create the poster and video
Description: Make the poster and the video about the project. This includes video editing, the design of the poster and generating the graphs.
Assignee(s): Kaur
Estimated time spent per person: 10 hours
Tools / methods used: Jupyter Notebook, Adobe Illustrator, Adobe Premiere Pro, a video camera.
Notes: The estimated time is subject to change as we don’t know how much data we will need to show and how.