Plamen Zaykov

Version 1.0 | 17/11/2019

Business Proposal

Football statistics

Contents

[1. Business and data understanding 2](#_Toc24895945)

[1.1 Data understanding 2](#_Toc24895946)

[1.2 Business problem 2](#_Toc24895947)

[1.3 Goals 2](#_Toc24895948)

[1.4 Data Science solution 3](#_Toc24895949)

[2. Exploratory Data Analysis (EDA) 4](#_Toc24895950)

[3. Data preparation 9](#_Toc24895951)

[4. Modelling 12](#_Toc24895952)

[5. Evaluation and Deployment 13](#_Toc24895953)

[6.1 Win the league acceptance model 13](#_Toc24895954)

[6.2 Predict goals acceptance model 13](#_Toc24895955)

[6. Conclusion 14](#_Toc24895956)

# Business and data understanding

In this chapter, it will be discussed how data is understood, what is the business problem and how it can be solved and what are the goals.

## Data understanding

Statistics is a collection which usually represent gathered data from an institution. In this case the statistics are from Football institution. The dataset contains 570 rows which represent 5 leagues: Bundesliga, EPL, La Liga, Ligue 1, RFPL and Serie A. For each league there is data for years 2014 until 2018. The dataset also consists of name of the league(League), year from which the statistics are(Year), position/place on which the team finished the current season/year (position), name of the team(team), number of played matches(matches), number of won games from the played matches(wins), number of draws from the played matches(draws), number of lost matches from the played matches(loses), number of scored goals(scored), number of missed goals(missed) and number of total points the team gained at the end of the season(pts). The original dataset holds additional columns which contains some predictions for the specific year like expected goals which were predicted before that specific season. These predicted columns won’t help my predictions because they do not contain real and accurate data, that’s why I decided to leave them out.

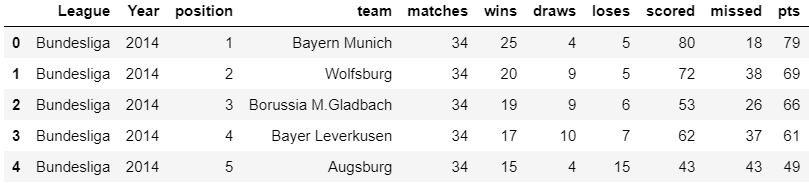


Figure Overview of the first 5 rows of the dataset

## Business problem

Football statistics datasets are usually related to predictions about scores, position, goals etc. In order to perform better, football clubs go through their statistics of previous matches and analyze their performance to see what they can upgrade for better future results. Based on statistics, football clubs also try to predict what their chances for winning the league are, their position, how many goals they can score etc. The focus of my business problem will be similar to most football clubs: predict how the teams will perform in the next season, or in other words, predict the statistics for next season based on statistics from previous years.

## Goals

For the business problem explained in the previous subchapter I have formulated the following goals:

* Predict on which position teams are likely to finish next year.
* Predict how many goals teams are likely to score next year.

## Data Science solution

For Data science problems there are 4 types of Machine Learning(ML) tasks: Supervised learning, Unsupervised learning, Semi-supervised learning, Reinforcement learning. From all 4 types mainly 2 of them are used, Supervised and Unsupervised. The difference between them is that Supervised ML algorithms use data which is labeled or in other words it already contains the right answers. In Supervised ML important thing is to define the output of the problem, for example we want to know if we should go left or right, if the type is A or B etc. then classification is the answer of our problems, because classification algorithms deal with limited number of values, so the answer of the examples will be either left or right and A or B, no new values, only answer. On the other hand, we have Regression algorithms which are also of type Supervised ML. Regression is used when the predict new values, for example how many or how much. The answer from the algorithm will be a value that wasn’t present in the predefined values (training sets). Having explained about Supervised learning, Unsupervised learning algorithms are used when the data is not labeled or in other words there are no correct answers yet. This is when for example we want our algorithms to find the differences between dogs and cats, but we don’t specify what are the specification of both animals. The algorithms will try and distinguish by itself what are the different features and after that predict if the new value (test data) is cat or dog.

Having statistics means the data I am working with is already labeled data because I have all features of the teams like wins, draws, loses etc. Having said that I think Supervised machine learning algorithms are best solution for my problems (goals).

# Exploratory Data Analysis (EDA)

Laying on assumptions may lead to false calculations or misleading results, that’s where graphs come in use. EDA is very useful for the start of a project, graphs give very clear and useful overview of the wanted features.

The goals of using EDA in this project are to find any trends between the teams like if a specific team usually wins the league, how many goals do teams usually score each year, are same teams tend to finish in top positions each year.

My first goal of using EDA was to find a connection between goals and final position.

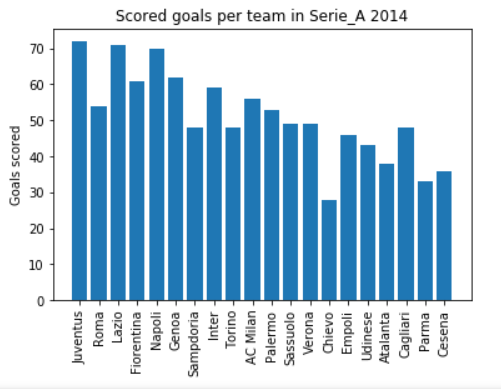


Figure Scored goals per team in Serie A 2014

Figure Overview of Serie A 2014

Figure 2 shows the scored goals per team from 2014 in Serie A league, if we also see the actual position on which the teams finished this season Figure 3 we can say that there is no clear relation between the amount of scored goals with the final position because ‘Roma’ who finished second that year have less goals than the next 4 teams but they still finished above them.

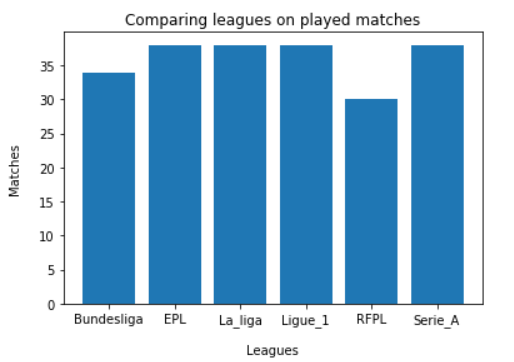


Figure Overview of played matches in every league

Figure 4 shows an overview of the amount played matches in every league in the dataset from which a conclusion can be made that every league has its own predefined number of matches which is followed every year. This could affect on the team performance in league level and at European stage.

Some leagues have trend which shows that every year almost same teams win the league or are least are at top 3 places. Lets see if that is correct.

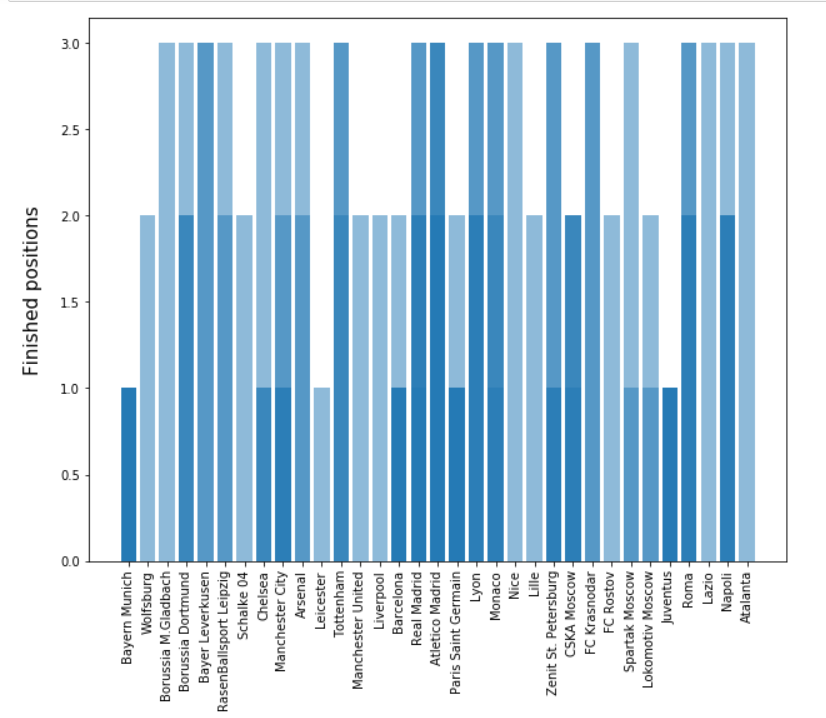


Figure Overview of all teams finished position

Figure 5 shows the place on which teams finished through all seasons in the dataset.

This overview is not quite clear how many times the specific team finished on the current position so let’s see a more in-depth graph:

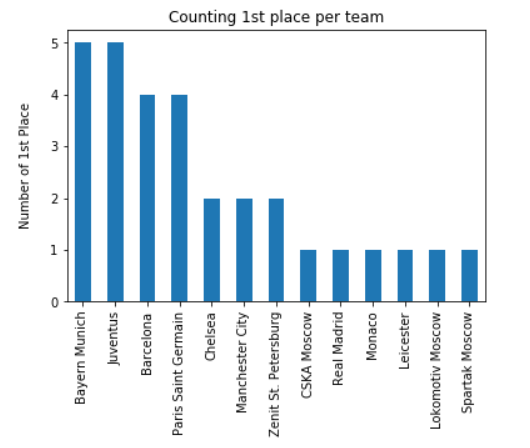


Figure Overview of teams finished on 1st place

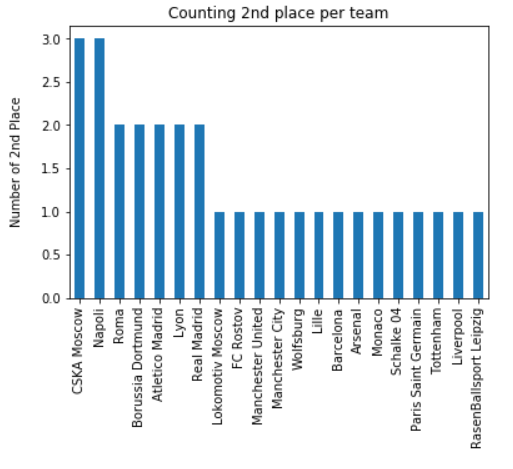


Figure Overview of teams finished on 2nd place

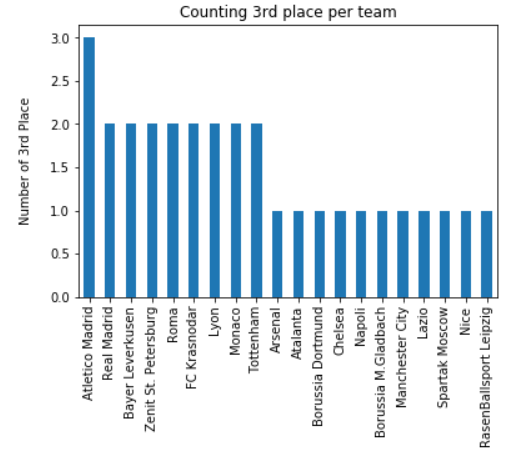


Figure Overview of teams finished on 3rd place

From these graphs it is clearer how many times a team has finished on certain position Figure 6 is for teams finished on 1st place, Figure 7 and Figure 8 are for teams finished on 2nd and 3rd place in the same order.

In conclusion, the findings from the graphs show that there is a trend for top 3 positions in almost every league. For example, in Bundesliga for all 5 years Bayern Munich are dominating the league with 5 first places, in RFPL top 3 is occupied by CSKA Moscow, Spartak Moscow and Lokomotiv Moscow, in Serie A the situation is same as in Bundesliga, Juventus are the team who has been at top position at the end of every season in this dataset. Having these examples, we can conclude that for future seasons the same teams are more likely to finish at top 3 positions than other teams. These trends will help the algorithms to be more precise and the accuracy should be higher.

# Data preparation

In order to have accurate results and predictions the dataset should be “cleaned”. The “cleaning” process means to find if every column has the correct data type (object, int, float etc.), if no null values appear – find why, are they at random or not, how to solve them also find if there is corrupted data – can it be decrypted and does this value matter a lot. In the dataset of Football Statistics, the type of columns is normal: all columns which contain string objects are of type object, all columns which contain integer values are of type int64, no values are missing and there is no corrupted data. However, extra columns, which were already discussed in chapter 1.1 Data understanding, are not needed so I left them out. So, from this:

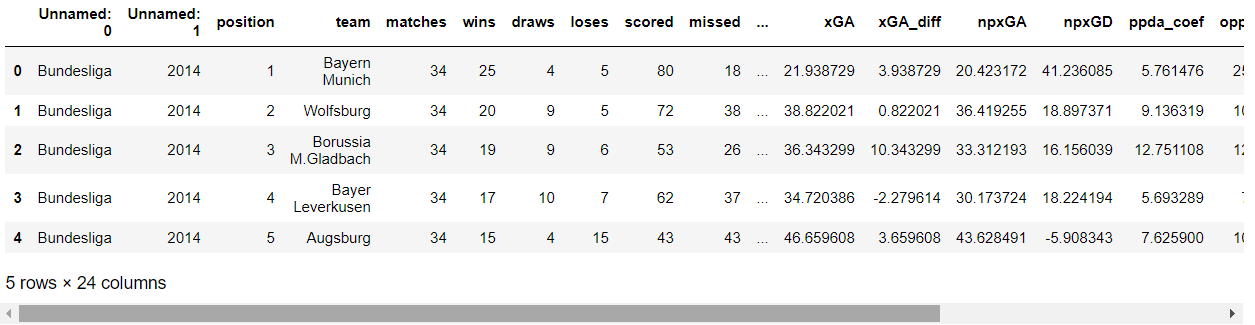


Figure Original dataset

I dropped all columns after column - “missed”

From Figure 9 other than the extra columns we can see that the first 2 columns, where the name of the league and the year are supposed to be have no column name - unnamed, so I changed their names and the dataset became like this:



Figure The dataset after changes

The types of the columns are:

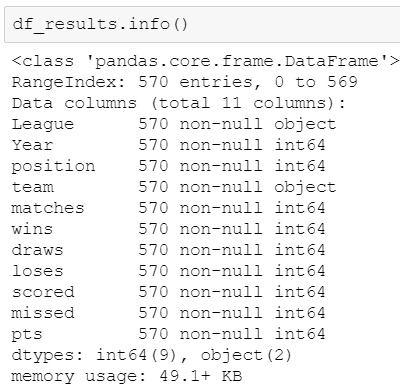
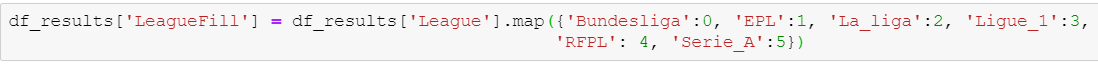


Figure Data types

There is one problem with the current situation, for some plots I need the league but graphs can be made only on integer or float columns but “League” column has an object type, that’s why I made an additional column called “LeagueFill” where I mapped all type of leagues with numbers



Other huge problem is that for predicting a team to win a league, algorithm can’t handle objects like Bayern Munich, Manchester United etc., algorithms use numeric data, so all teams need to be mapped with numbers. I achieved that with a simple loop which maps all unique values from column “team” with a numeric value – 0,1,2 etc. The final look of the dataset is:

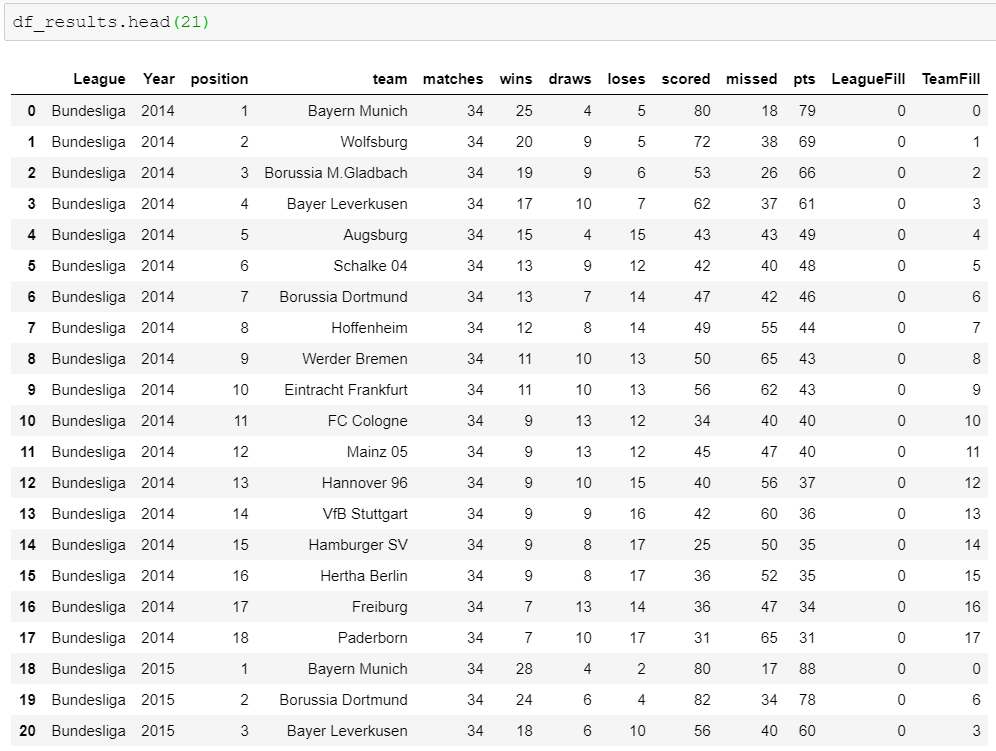


Figure Final look of the dataset

# Modelling

To make predictions, data first needs to be modelled which means to choose the appropriate algorithms that should be used for the specific problem. In order to find the right algorithm, I will use Figure 13 as cheat graph.

Before using any algorithms, I will use the SelectKBest algorithm to find the best columns on which my predictions should be based on.

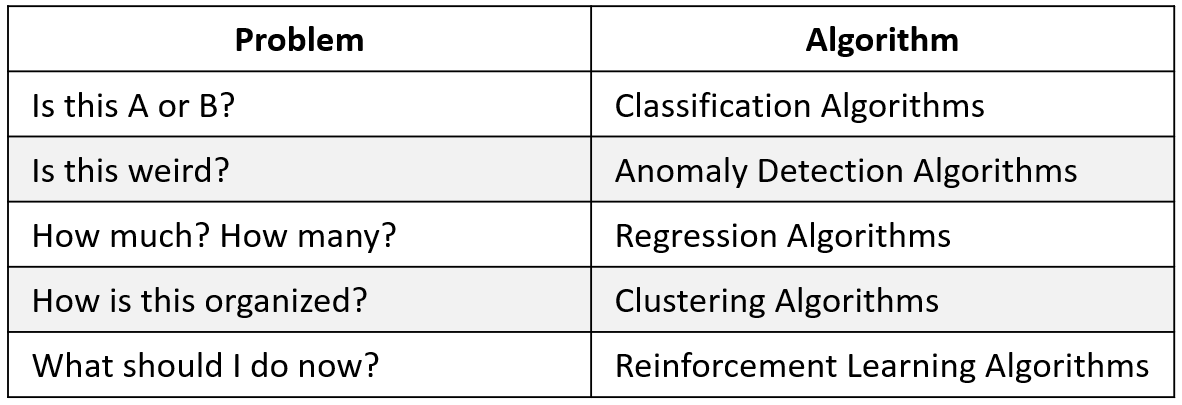


Figure Finding correct Algorithms based on the problem

* Predict on which position teams are likely to finish next year– After mapping all teams with numbers, the answer of this goal is on which place will a team be placed, so the problem is of type Is this A or B ? which corresponds to the Classification algorithms. Data from previous seasons like how many times has a specific team finished on 1st place or has been close to winning the league is important and would be used as a training data for the predictions.
* Predict how many goals teams are likely to score next year – predicting scored goals is dragging to the problem How many goals? which is solved by regression algorithms. Every team has amount of scored goals in the past years.

# Evaluation and Deployment

In this chapter it will be discussed the benefits of using the models that were specified in the previous chapter.

## 5.1 Predict position acceptance model

Predictions of type winning the league or final position are very hard to make in real life because of the many factors supporting it like player injury, manager swap, team not in condition etc. and that is what makes football a beautiful game, but in this case I will consider a perfect scenario where only pure statistics are handled. As the goal say, it will be predicted which team will finish on which position after all matches have passed. Having to predict specific list of values (positions) will be handled by classification algorithm which is the best option when the output should be of already predefined value (class).

## 5.2 Predict goals acceptance model

Predicting the possible number of scored goals in the following year can be target for the players, they should try their best and exceed the target goals. Based on scored goals in previous year prediction could be made for the next season. Using regression algorithm would be the best solution because the output will be an infinite number.

# Conclusion

The goal of this individual challenge is to prove the knowledge and skills I have obtained so far. Football statistics dataset was a good example of applying cleaning, which is necessary for best predictions, making an overview for specific part of the dataset which is of interest such as showing which teams have finished in top positions last years. EDA also help to find if there are any trends like there is in Bundesliga and Serie A where Bayern Munich and Juventus are dominating and chances of finishing again at first place next season are great. For the actual predictions choosing correct algorithm is necessary because not all will perform the same and not all are suitable for the goals I have.