Project Dataset Selection AIT-580-P01 (Fall 2019)

By Kamil Ismailov

Dataset description

1. The dataset contains statistics and information about twenty soccer clubs of England soccer tournament which calls Premier League. The dataset created by the Football-Data betting portal [1]. This website provides historical results and odds by the web-links in the CSV format files for football betting players [2]. As a result, betting enthusiasts can analyze the dataset and make some predictions.

I chose the dataset of Premier League results of the season 2018/2019 [3]. The size of the file is 94KB. The data has 62 different items including the betting odds of several bookmakers. The dataset contains numerical and categorical types. There are nominal and ratio measurement scales in the dataset. The Football-Data also provides a meaning of abbreviations of all items [4].

2. The purpose of the Football-data makes money by offering calculated bets and bonuses for online sports bookmakers. Through the dataset, they assess odds and betting coefficients for each soccer match.

3. The dataset is provided as an open-source. The issue might be with the credibility and quality of information. The web-portal doesn’t provide any reliable analytic organization which collected this data. Even the Football-data created this dataset by themselves, they should also specify analytics who gathered all information and how they collected it. The challenge was to compare the dataset with the official statistical portals because these organizations don’t provide whole statistical information for free.

4. By analyzing the selected dataset, I can answer some specific questions related to the soccer clubs in England. Based on the specific results of matches I want to extract the potential patterns which can explain, for example, why Manchester City soccer club won the England championship last season. The different items of the dataset might give useful information about the coach’s tactic preferences. For instance, many tackles and foals point to the defensive style of play, etc. Another purpose of the analysis is to compare the result of the previous season with the current season and make predictions about a possible champion.

5. I am planning to use RStudio [5] and Python [6] software to analyze the dataset. As a dataset has already represented by CSV-file, it is easy to extract and implement it to the chosen software. Moreover, my current skills allow me to make some statistical analysis now.

6. The Football-Data betting portal made some similar study which I try to obtain. It calls “Rating System for Foxed Odds Football Match Prediction” [7]. This study describes “how rating analysis using computer-ready results and betting odds data can help one to establish a betting edge, as in the chart above right” [8].

The reason why I interested in is that I watch the Premier League. So, I would like to satisfy my curiosity about soccer statistics. I hope that my outcomes may help to other soccer fans to understand the game more clearly.

References

1. Football-Data, 2019, https://www.football-data.co.uk
2. Football-Data, 16 October 2019, https://www.football-data.co.uk/englandm.php
3. “Premier League, Season 2018/2019”, Football-Data, 16 October 2019, https://www.football-data.co.uk/mmz4281/1819/E0.csv
4. “Notes for Football Data”, Football-Data, https://www.football-data.co.uk/notes.txt
5. RStudio Desktop [Computer software], Version 1.2.5001, 2019, https://rstudio.com
6. Python Software Foundation. Python Language Reference, version 3.7, 2019, http://www.python.org
7. “Rating System for Foxed Odds Football Match Prediction”, Football-Data, 2003, http://www.football-data.co.uk/ratings.pdf
8. “Historical Football Results and Betting Odds Data”, Football-Data, http://www.football-data.co.uk/data.php