Data Science
Project Retrospective
- Group 5 -

Mohammed Aljader & Andreea Șindrilaru

Table of contents

Introduction

During the last 6 weeks, we have been tasked to create a project where we have been given a dataset with all the football matches from the Dutch Premier League, so that in the end we are able to make different predictions about certain teams. Before making the predictions, we had to make sure that the data is well cleaned and aggregated according to what predictions we want to make, and visualize the data using graphs so that we have a better understanding of what is happening with it. After these steps are done, we can use one of the machine learning algorithms, more specifically decision trees, to make our predictions.

Were our predictions accurate?

We have made some tables to check how accurate were, in the end, our predictions.

Winning team:

Winning team Predictions			
Match	Predictions	Reality	Result
SC Cambuur Vs RKC Waalwijk	Draw	Draw	Correct
FC Twente Vs Fortuna Sittard	Fortuna Sittard Win	Fortuna Sittard Win	Correct
PEC Zwolle Vs FC Utrecht	PEC Zwolle Win	Draw	Not Correct
Willem II Tilburg Vs Heracles Almelo	Willem II Tilburg Win	Willem II Tilburg	Correct
FC Groningen Vs Sparta Rotterdam	FC Groningen Win	Sparta Rotterdam Win	Not Correct
NEC Nijmegen Vs Go Ahead Eagles	Draw	NEC Nijmegen Win	Not Correct
AZ Alkmaar Vs Ajax Amsterdam	Ajax Amsterdam Win	AZ Alkmaar Win	Not Correct
Vitesse Arnhem Vs SC Heerenveen	Vitesse Arnhem Win	SC Heerenveen Win	Not Correct
Feyenoord Rotterdam Vs PSV Eindhoven	PSV Eindhoven Win	Draw	Not Correct
Total Correct predictions = 3 out of 9			

Even though we got 3/9 predictions right, given the fact that we made predictions for round 32 out of 34, each team tried its best in the respective match so that they would not go into the second league. (e.g. Willem II Tilburg, Sparta Rotterdam etc.)

Which team is more likely to score the first goal:

More likely to score the first goal Predictions			
Match	Predictions	Reality	Result
SC Cambuur Vs RKC Waalwijk	Away score first	Away score first	Correct
FC Twente Vs Fortuna Sittard	Home scores first	Away score first	Not Correct
PEC Zwolle Vs FC Utrecht	Home scores first	Away score first	Not Correct
Willem II Tilburg Vs Heracles Almelo	Home scores first	Home scores first	Correct
FC Groningen Vs Sparta Rotterdam	Home scores first	Home scores first	Correct
NEC Nijmegen Vs Go Ahead Eagles	Home scores first	Home scores first	Correct
AZ Alkmaar Vs Ajax Amsterdam	Away score first	Away score first	Correct
Vitesse Arnhem Vs SC Heerenveen	Home scores first	Home scores first	Correct
Feyenoord Rotterdam Vs PSV Eindhoven	Home scores first	Away score first	Not Correct
Total Correct predictions = 6 out of 9			

Figure 2

Winning team based on which team scores the first goal:

Winning team after the first goal Predictions			
Match	Predictions	Reality	Result
SC Cambuur Vs RKC Waalwijk	Away scores first and won	Away scores first and tied	Not Correct
FC Twente Vs Fortuna Sittard	Away scores first and won	Away scores first and won	Correct
PEC Zwolle Vs FC Utrecht	Away scores first and tied	Away scores first and tied	Correct
Willem II Tilburg Vs Heracles Almelo	Home scores first and won	Home scores first and won	Correct
FC Groningen Vs Sparta Rotterdam	Away scores first and won	Away scores first and won	Correct
NEC Nijmegen Vs Go Ahead Eagles	Home scores first and won	Home scores first and won	Correct
AZ Alkmaar Vs Ajax Amsterdam	Away scores first and won	Away scores first and tied	Not Correct
Vitesse Arnhem Vs SC Heerenveen	Home scores first and won	Home scores first and lost	Not Correct
Feyenoord Rotterdam Vs PSV Eindhoven	Away scores first and won	Away scores first and tied	Not Correct
Total Correct predictions = 5 out of 9			

Figure 3

In data visualization we saw that most of the times, when the away team scores first, the home team has a higher chance of losing the game. This is why we decided to make a prediction about the winning team after the first goal.

Compared to the “Winning team” table, where we only predicted which team is more likely to win without giving any additional information, the above table is more accurate. Therefore, providing additional information turned out to improve our prediction.

Overall, we got 14 correct predictions out of 27, which we consider to be a good result for a first data science project.

Conclusions and future work

While working on this project, we have learned how important data cleaning is, especially for making accurate predictions and, also data visualization, for understanding what is happening with the dataset much better. Related to machine learning, we got acquainted with basic algorithms such as Linear Regression, Decision Tree, Logistic Regression etc.

Also, we got a better insight of different types of learning such as: *Supervised Learning* and *Unsupervised Learning*, and when it would be good to use them, depending on what we want our machine to do.

In conclusion, we enjoyed working on this project and checking if our predictions were right or not. Now that we have the basics of what data science implies, definitely we can go even further and apply what we learned in this course to different projects.