Project report

Project B14: International football

to analyse predictive powers of Football ELO systems

Brandon Rauba, Karl Hans Kostabi

Project repository: https://github.com/khkostabi/Andmeteadus_projekt

Business understanding

Identifying our business goals

Background: Football is one of the biggest sports on the planet. Biggest football matches are seen by hundreds of millions of people. Because of that there is also a lot of interest in predicting the results of games. This is also tied to the sports betting industry, where people bet money on football matches. We are aware of the Elo score, which is used as a measure of showing the quantified skill level and helping therefore to make predictions. As such we are interested in discovering how may we use the Elo system to predict the outcomes of football matches better. However since we suspect that football matches can be influenced by a lot of different factors we are also interested to know if home turf is one of them.

Business goals: Our main and primary goal is to focus on predicting international football match outcomes with different Elo systems. Our secondary goal is to analyse whether and how home turf plays a role in the outcome of matches.

Business success criteria: The project is a success in our eyes if we manage to decisively show how much better than random guessing different systems are at predicting the results of international football matches. It is a bigger success if we can show how much, if at all, does hosting a major tournament help a country's chances in the tournament

Assessing our situation

Inventory of resources: The resources available to us for this project are the following: our dataset (https://www.kaggle.com/martj42/international-football-results-from-1872-to-201), the two team members, python and personal computers.

Requirements, assumptions, and constraints: The requirements for this are the following.

- Deadline on 16.12.2021
- The validity of the dataset
- The resulting analysis must be able to show how good are different types of Elo systems at making predictions and to show how much hosting a tournament plays a role.

Risks and contingencies: As it appears to us, there is not risk to the completion of the project other than ourselves. The contingency to minimize risk is to start the project early and to follow a detailed schedule, that is subsequently followed.

Terminology: Not relevant to our project

Costs and benefits: Not relevant to our project

Defining our data-mining goals

Data-mining goals: Our goals are analysing atleast three different models for predicting the result of a match between two countries football teams. We compare the different models based on their predictive power and accuracy. We also will as our secondary goal try to see, what is the pattern between hosting a tournament and the performance of the corresponding team to see if it is higher than it would be if the two would be uncorrelated.

Data-mining success criteria: The models successfully make predictions about a teams performance in a match and those predictions are better than random. We also have to find if performance in a tournament and hosting said tournament are correlated or not.

Data understanding

Gathering data:

We have gathered our data from online community kaggle. Since kaggle offers free datasets for everyone we have acquired access to the data. Both our datasets that we use are in the .csv format. The data indeed is real and we can use it. Link to our database is in business understanding point under resources paragraph. We will run our data in Python and with some tests it seems to work fine without any errors.

Describing data:

Our data comes in two datasets. One that is showing the normal time winners and games. Second dataset is for the games that went to penalties. Both datasets come in .csv format. First dataset consists of 9 columns and 43 000 rows making the size of the file 3.22 MB. Second dataset consists of 4 columns and 433 rows making the size of the file 16.17 kB.

Exploring data:

Our project main goal was to compare football elo and our secondary goal is to find out if playing at home stadium gives any advantage. Therefore our key features are going to be the scores for both teams and the stadium where the game took place. Since our data doesn't contain unnecessarily many columns we won't have to make any subsets from our current data. Although making subset that contains all games and winners even the ones where game ended with peanlties could make our work much easier.

Verifying data quality:

As far as we have checked the data should be complete. Meaning that all the cases that are possible in football are covered. In our case draws are not considered as a option. With this small ammount of time that we have worked with the dataset doesn't show any „miss values" aswell. Since that data is up to date we can get the newest data on the go. Overall the quality seems to be pretty good and it will be enough for us to find the elo system that is our main goal.

Project plan

Tasks:

Task 1: Managing our time

Although homework 10 is preparation for our project we are going to make agreements on how and when we wan't something to be completed in order to finish the project on time and reach the goal that we have set.

This task will take both members to commit for atleast 2 hours.

Task 2: Preparing the data

In this task we will consider if we want to make subsets or not, work through our data to find any possible errors and prepare it to use for finding the elo.

In this task both team members will spend approximately 5 hours of work.

Task 3: Starting the elo system models training

Since this is a bigger step towards our end goal we will have more time to complete it. In perfect case scenario in this step we should finish one model for our elo systems.

Both member work on this task for approximately 10 hours.

Task 4: Check ups

This is just to meet up and check if our project is going according to our plan or not.

Both member work here for an hour.

Since we wan't to train more than one model we will repeat task 3 and 4 as long as we have achieved the end goal.

Task 5: Final check

Here we will check our final elo systems and discuss if we have reached our goals or not.

Both members should work here for 1 hour.

Task 6: Preparing the project video and presentation

This task is after our project is ready to present. We will make a video that gives an preview of our project.

For this task we will both use about 5 hours of work.