International football results from 1872 to 2020

Pelle Holden Porila
Kristo Markov

Business understanding

Background:
This year should have taken place UEFA Euro 2020 but because of Covid-19 it was postponed until 2021. Every time when it has been a big tournament it is alway interesting to see how people try to predict the outcome of these tournaments with different methods . Who uses animals, who cards and so on. But we do not think that these methods are really good because in the end they do not have any logic behind them and they do not use any kind of data. We believe that it would be interesting to try to predict the results based on their previous games using Machine Learning.

Goal and success criteria:
Our goal is to try and predict so many games correctly as possible. For example we would be happy if we could predict the outcome of 15 games out of 51 games.
If we predict the game game winner and loser or if the game ends in a draw then we have been successful.

Inventory of resources, requirements, assumptions, and constraints:
For the project we would use the “International football results from 1872 to 2020”dataset. We got this dataset from Kaggle Open Datasets and the author of this dataset is Margus Jürisoo. We would use Jupyter to write our code and we would use Meelis Kull lecture slides to get more information on how to do something and what we should do.

Risks and contingencies:
Our only risk would be that one of use would do some work and then forgot to save it or wouldn’t push it up and the other one would do that too. To avoid these things we update our repository as much as possible and communicate with each other every time when someone does something.

Terminology:
- Victory - the winning side gains 3 points.
- Loss - the losing side gains 0 points.
- Draw - Both sides gain 1 point.

- Group stage - the first part of a UEFA Euro 2020 final tournament when teams are divided into groups and the teams in each group play against each other. The first and second of each group then move onto the playoff-s.
- Knockout phase- elimination tournament which starts with 16 teams who passed the group stages. If a team loses then it is going to get eliminated from the competition. If a team wins then it is going to move on to quarterfinals,
- Quarterfinals - There are a total of 8 teams and 4 winners are going to move on to the semifinals.
- Semifinals - There are a total of 4 teams and 2 winners are going to move on to the finals.
- Finals - The winner of the final is the winner of the whole tournament.

Costs and benefits:
We do not have any costs with this project because the dataset is free and the program where we are going to write our code is also free. But we would gain knowledge about machine learning, who could be the winner of UEFA Euro 2020, which team has scored the most goals, which team has won the most games, which team has lost the most games. But we would not gain any money from this so costs and benefits are even if we look at it from an economic point of view.

Data understanding

Outline data requirements:
For our data mining goals we need the following data: The date of when a match took place. Which team was in home country and which one was away. Home teams' score and away teams' score. Which the tournament took place. Which city it took place in. Which country the match took place in. Whether the match took place at a neutral venue or not.
Verify data availability:
The needed data exists and we can use it.

Define selection criteria:
We can make use of every feature in the following dataset:
https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017
Here we can find more variables if we need to increase the scope, especially the data on players and coaches:
http://www.rsssf.com/intland.html

Describing data:

So first of all we have to dates, the more recent the date the more matches there are in a couple of years period It follows the format YYYY:MM:DD so that's pretty good. Then we

have home team which shows which country was playing at home. Then the away team which shows the country of the away team. Home score which is a numeric values which shows the amount of goals made by the home team, away_score which shows the goals of the away team. The tournament which shows the tournament the teams were playing at, value friendly is used if there is no tournament. City shows which city the teams played at, country shows which country the teams played at, neutral is a boolean value that shows if the match took place at a neutral venue or not.

Exploring data:

There are 40619 entries. The mean home_score is 1,7438, The mean away_score is 1,1849. There are 7199 FIFA qualification matches and 2432 UEFA euro qualification matches, 900 FIFA World cup matches and 286 UEFA euro matches

Verifing data quality:

The data is on good format and seems to be complete.

Planning project

1. Deciding which tools and method we are going to use for each task (2 hours together)
2. Selecting data (Kristo 3 hours).
   We will decide which portion of the data that is actually going to be used for data mining.
3. Cleaning data(Kristo 3 hours)
4. Constructing data (Pelle 3 hours)
   We may need to derive some new fields or otherwise create a new form of data.
5. Integrating data(Pelle 3 hours)
   We need to merge some of those disparate datasets together to get ready for the modeling phase.
6. Selecting modeling techniques (1 hour when working together)
   We have not yet made up our minds which techniques to use.
7. Designing tests (Kristo 4 hours and Pelle 4 hours)
8. Building models(Kristo 4 hours and Pelle 4 hours)
9. Assessing models (2 hours when working together)
10. Making a presentation (Kristo 10 hours and Pelle 10 hours)

Repository: https://github.com/pellehol/DataScienceProject.git