Project D21: KAGGLE-International football

Team members:
Mikk Viigand
Hergo Hansman
Illimar Laanisto

Github: https://github.com/MViigand/international-football

Task 2

Our clients would mostly be sports betting sites and companies that look to gain better insight into the possible results of international football games. This means that betting sites would be able to make more accurate predictions, because instead of just using their current prediction system they could also plug in our system that is mostly based on historic data. This would mostly reflect in using better coefficients for given games.

The business goal would be to improve the prediction accuracy for the companies by any noticeable amount. This would mean the accuracy should at least increase by 5% to know that this increase was not random but actually caused by using the data provided by us. This in turn would increase the profits for said companies by being able to make better and more profitable coefficients for international football games.

The project will be considered a success if the profits of a company using our data gains from international football games goes up at least 10% within the timeframe of 1 year after being implemented.

Resources available for the project are 3 beginner data miners who are also open to solve problems in any other areas should any problems arise. For data there is currently one Kaggle dataset consisting of over 40000 international football game results.

The schedule for completion is the following.
Dec 5 - Finishing the first 2 tasks posted about the project, finding out how big is the home court advantage and whether games played on neutral venue have smaller goal difference than rest of the games.
Dec 12 - Finish the model that would predict the score of future games
Dec 15 - Finish the poster regarding presenting the project
Dec 17 - Presenting the project

There is a risk that the data used is not enough to properly predict results of future games, in that case we would need to find another dataset to increase our accuracy and add variables to predict on.

There are no notable terms to define regarding this project.

The current project does not have any costs or benefits in this case.

The main deliverable from our project is the model that would predict results of the games, while also delivering reports about the other 2 goals of this project. Those being the effect of home court advantage and the goal difference between games played on neutral courts against games where one team plays at home.

The criteria for the future game score prediction model would be to at least achieve a model accuracy of 80%.

Task 3

The data necessary to address the data mining goals is quite simple. The project requires data across a large timeframe, but it doesn't have to be very detailed with a lot of different properties. A dataset containing a large amount of games from a long time period with simple properties as the date of the match, playing teams, scores of the teams and where the match was played will be suitable. Predicting the outcome of future matches likely requires more detailed data, but over a smaller time frame, as factors can change quickly.

We have access to the data required for the project. We are able to process it using conventional means as a dataframe without making many changes to the structure of the dataset or the data itself.

The data for the project is sourced from Kaggle. We will be using a dataset with information from more than 40000 games of international men's football. We will use the entirety of the dataset, as all of it is relevant to this project.

As mentioned before, the dataset we will be using is a public dataset from Kaggle formatted as a csv file. The data is divided into 9 properties, 5 of which are strings, 2 of which are integers, 1 is a boolean and 1 is a date. The properties stored as strings are the names of the playing teams (2), the country and city where the match took place (2) and the name of the tournament the game was a part of. Integers describe the scores of both teams. The boolean describes if the match was played on neutral grounds or not. The date field stores the date of the match, in a YYYY-MM-DD format. There are 40641 cases in the dataset, each describing one match. This data is suitable for our data mining goals.

The data quality is usable. There are minor quality issues, like missing or incorrect values, but these issues aren't common and only affect a very small amount of the overall dataset. Cases with missing or incorrect data can easily be removed, or in some cases corrected, as the incorrect data tends to be obvious misspellings.

Task 4

1. Data Preparation
    a. Selecting data- Mikk 3h
    b. Cleaning data - Mikk 3h
    c. Constructing data - Hergo 3h
    d. Integrating data - Hergo 3h
    e. Formatting data - Hergo 3h
2. Modeling
    a. Selecting modeling techniques - Illimar 3h
    b. Designing test(s) - Illimar 8h
    c. Building model(s) - Mikk 8h
    d. Assessing model(s) - Mikk 8h
3. Evaluation
    a. Evaluating results - Illimar 8h
    b. Reviewing the process - Hergo 9h
4. Deployment
    a. Planning deployment - Illimar 3h
    b. Planning monitoring and maintenance - Illimar 4h
    c. Reporting final results - Mikk 4h
    d. Reviewing final results - Hergo 8h
5. Preparation for presentation
    a. Making poster - all together 4h
    b. Presenting

We have still yet to decide what models we are going to use to reach our goals.