NBA-predictor

Autorid: Hendrik Saliste, Gert Kalmus, Karl Aleksander Kaplinski

NBA-predictor on ülikooli kursuse "Sissejuhatus andmeteadusesse" valmiv rühmatöö.

Esimene ülesanne - "Setting up"

The project repository has been set up.

Teine ülesanne

Developing a business understanding within CRISP-DM consists of four tasks: identifying your business goals, assessing your situation, defining your data-analysis, data-mining or machine learning goals and producing your project plan. For this exercise, please develop a business understanding of your project.

• Identifying your business goals:

o Background:

We are three students from the introduction to data science course. To complete the course, we must do a data science project. We chose NBA since some members are already very familiar with it and it has a lot of good data for analysis.

o Business goals:

We have two main goals with this project. First is completing the introduction to data science course. Second is creating a predictive model for NBA games to learn more about both the data science and sport aspects that go into it.

o Business success criteria:

The main success criteria is to create a model that through predicting NBA games, can give insight into the most important factors that decide the outcome of a game.

• Assessing your situation:

o Inventory of resources:

We have three people working on the project, who all have personal computers powerful enough to train different models. The data comes from Kaggle. Code will be hosted by GitHub.

o Requirements, assumptions, and constraints:

There are two important deadlines. 9.12.2024 when a poster introducing the project must be completed, and 13.12.2024 when the project has to be finished and the poster must be presented.

o Risks and contingencies:

The only real risk is that we due to some unforeseen circumstance, a member can’t produce their part of the project. By having good communication between members, should such a problem arise others can easily take over, making sure the project is completed on time.

o Terminology:

♣ NBA – national basketball association

♣ FG% - field goal percentage

♣ 3PT FG% - 3 point field goal percentage

♣ SPG – steals per game

♣ BPG – blocks per game

♣ APG – assists per game

♣ RPG – rebounds per game

o Costs and benefits:

There are no monetary benefits since this is a school project.

• Defining your data-mining goals:

o Data-mining goals:

Main goal is the creation of a model that, based on data available before a game, can predict the outcomes of NBA games. Presentation has to be done in the form of a poster, that showcases and explains the model and its output.

o Data-mining success criteria:

The model can be considered successful if over a season its prediction accuracy is over 75%.

Kolmas ülesanne

Data understanding within CRISP-DM consists of performing four tasks: gathering, describing, exploring, and verifying data quality. For this exercise, please develop a data understanding of your project.

Gathering data:

For this project, the base data used in the model originates from NBA datasets from Kaggle. Mostly "play_by_play.csv" is used, which describes every game event from every game and has 13+ million rows. Since the last season it contains is 2022-2023 season, the idea was to use the data from that season, to predict the following season. For that, the data from that season specifically needed to be extracted. Since the next season has already happened in real life, we can compare our eventual predictions next to what really happened.

Describing data:

The data of season 22/23 contains 516540 rows. There is a total of 30 columns. Specific games are identified by game ID-s, events with event ID-s and described in home or away descriptions, whether it is a foul, steal or a missed shot and et cetera. If a player is involved, their name will be in another column and so will be their team name. Each event can be extracted and used for calculations to eventually distinguish teams and players. There is also a score column which keeps track of the score. Game timer has its own column and quarters of the game are also separated.

Exploring data:

After extracting stats like points, assists and rebounds per game, star players can easily be distinguished. Also total points scored on the season tended to be in positive correlation with the amount of games played. To try to predict the contribution of each player in the following season, maybe the percentage of games they missed last season or maybe in their career should be involved into modeling. However, injuries, in principle, are not predictable, but tend to play a big role in the outcome of each game. After calculating some of the team stats, it was clear that the teams that did more winning, usually shot the ball more efficiently. Since the game of basketball is very turn based, defence needs to also be taken into account. Stats like steals, blocks, rebounds and defensive rating can help evaluate the defensive performance of teams.

Verifying data quality:

Going through the data with some calculations it seems, that 124 games from the season are not included. These games seemingly have no similarities and appear to be completely random. That means, that instead of 1230 games (default regular season), we currently have the data of 1106. Otherwise the data is seemingly very clean, precise and consistent, allowing for good analysis.

Neljas ülesanne

Hendrik Saliste (30 tundi)

Hendrik Saliste tegeleb andmetöötlusega. Peamiselt andmete eraldus algandmetest, korvpallialaste mõõdikute arvutus ning mudeli algandmete koostamine.

Karl Aleksander Kaplinski (30 tundi)

Karl Aleksander Kaplinski treenib ennustava mudeli. Analüüsib mudeleid, mudeli parameetreid ja nende mõju. Analüüsib kasutatavat statistikat ja selles sisalduvate faktorite mõju. Teostab ennustusi.

Gert Kalmus (30 tundi)

Gert Kalmus interpreteerib saadud tulemusi ja koostab rühmatööle plakati.