Info

Repo link: https://github.com/tiitvaino/ids_tiit_robert
Group: D18
Group members:

- Tiit Vaino
- Robert Reimann

Business understanding

Identifying business goals

Background

Powerlifting is a strength sport that consists of three attempts at maximal weight on three lifts: squat, bench press, and deadlift. As in the sport of Olympic weightlifting, it involves the athlete attempting a maximal weight single lift of a barbell loaded with weight plates. Powerlifting evolved from a sport known as "odd lifts", which followed the same three-attempt format but used a wider variety of events, akin to strongman competition. Eventually odd lifts became standardized to the current three.
We seek to understand the factors that go into a successful powerlifting competitor. We have a large dataset which allows us to measure the impact of features such as age, weight, drug testing etc.

People want to know their and others competitors possible results on competition.

Business goals

Providing hypothetical results to the customers, what their possible results may be on a competition.

Business success criteria

We can tell if this project was a success, if we can predict most of the time quite accurately.

Assessing situation

Inventory of resources

- For this project we have two people. Both of them are Computer Science second year students. They have some knowledge about data management, statistics and programming.

- We also have possibility to ask advice from:
    - Lecturer: Meelis Kull (meelis.kull@ut.ee)
    - Teaching Assistants:
        - Victor Pinheiro (victor.pinheiro@ut.ee)
        - Anna Aljanaki (anna.aljanaki@ut.ee)
        - Markus Kängsepp (markus.kangsepp@ut.ee)
- Data we have in csv-file from Kaggle, which contains over 1.4 million rows.
- For the project we have two HP elitebook 840 G5 laptops.
- We are working with Windows 10 and Linux Mint operating system.
- For coding we use Python 3 and to make code more understandable we use jupyter notebook.
- For project management we use git and github.
- For communication we use Discord and Messenger.

Requirements, assumptions, and constraints

- Project should be ready by Thursday, December 17, 2020
- By the project ending deadline we should have data prepared, trained the model, tested the model, evaluated results and deployed.

Risks and contingencies

- Possible risk may be a broken computer, because then it is hard to work when you can not access the project.
- Lack of time. Others courses may need more time or other unexpected events.
- Data may have too less good attributes, then it is hard to train accurate model

Terminology

- Name: Name of the competitor.
- Sex: Sex of the competitor.
- Event: The lifts competed in, S refers to squat; B refers to bench press; D refers to deadlift. SBD means the competitors performed all three powerlifting lifts, B means only bench press was performed in this meet.
- Equipment:
    - “Raw” powerlifting means athletes are (generally, might vary by federation) only allowed to use approved lifting belts, an approved singlet, approved wrist wraps, approved knee sleeves, and chalk.
    - Single-ply and Multi-ply suits help the athlete lift more in all competition lifts - therefore better results should be expected.
    - Straps help the athlete deadlift more by removing grip strength from the equation. Straps in powerlifting competitions are very rare which explains the small dataset.
- Age: Age of the competitor.
- AgeClass: Age class the athlete competes in is determined by their age.
- Division: Divides the athletes into groups based on age and sex.
- BodyweightKg: Bodyweight of athlete at the time of the competition (in kilograms)

- WeightClassKg: The weight class the athlete competes in.
- SquatXKg: The result of a squat attempt. ‘X’ denotes the attempt number (1, 2, 3, 4). Note that a - (minus) before the result denotes a failed lift - it was attempted, but the judges didn’t count it.
- Best3SquatKg: Maximum of the squat attempts that was approved by the judges.
- BenchXKg: The result of a bench press attempt. ‘X’ denotes the attempt number (1, 2, 3, 4). Note that a - (minus) before the result denotes a failed lift - it was attempted, but the judges didn’t count it.
- Best3BenchKg: Maximum of the bench press attempts that was approved by the judges.
- DeadliftXKg: The result of a deadlift attempt. ‘X’ denotes the attempt number (1, 2, 3, (4 - only applies to a few federations that allow a fourth attempt to break a (world) record)). Note that a ‘-’ (minus) before the result denotes a failed lift - it was attempted, but the judges didn’t count it.
- Best3DeadliftKg: Maximum of the deadlift attempts that was approved by the judges.
- TotalKg: The sum of Best3SquatKg, Best3BenchKg, Best3DeadliftKg
- Place: The place that the athlete got in that meet.
  Wilks: The Wilks Coefficient or Wilks Formula is a coefficient that can be used to measure the strength of a powerlifter against other powerlifters despite the different weights of the lifters.
- McCulloch: Alternative to Wilks Coefficient that also uses age, made for master division athletes.
- Glossbrenner: Alternative to Wilks Coefficient.
- IPFFPoints: Alternative to Wilks Coefficient.
- Tested: Whether the athlete was tested for Performance Enhancing Drugs.
- Country: The country the athlete represents.
- Federation: The powerlifting federation the athlete is a part of.
- Date: When the meet happened.
- MeetCountry: In which country the meet happened.
- MeetState: In which state the meet happened (if applicable)
- MeetName: Name of the powerlifting meet.

Costs and benefits

- 60 hours of job done by two people, which is about 2 ECP :) (one for each group member)
- At least 10 points out of 20 points

Defining data-mining goals

Data-mining goals

- Informative dashboard about dataset
- The machine-learned model, which is capable of predicting a person’s result in competition

Data-mining success criteria

- We separate test set against which we test and the accuracy should be over 80%.

Data understanding

Gathering data

Outline data requirements

We need data,

- where is described some attributes about a competitor
  - like name, weight, age, previous experiences
- which has results of a competitor
- which is easily usable with notepad jupyter and python
  - Written in some kind of text file (like .csv, .txt, .xml, .json)
- which does not require a lot of time for preparation, because time is a limiting factor in this project
- which is free to use, because then it takes less time and resources to make it usable
- which is not too large, because for this project we use laptops
- which is not too small, because we need to train the model based on this data.
- which has recordings for at least 5 years.

Verify data availability

We have confirmed that the Kaggle dataset (link here: https://www.kaggle.com/open-powerlifting/powerlifting-database), which is a public and free to use dataset, satisfies our requirements for this project. There is no need to substitute with an alternative data source or narrow the scope of our project.

Define selection criteria

Only the aforementioned kaggle dataset will be used, most tables of this dataset will be used. Most likely tables to be unused are the various coefficients (i.e Wilks, McCulloch etc).

Describing data

This dataset is a snapshot of the OpenPowerlifting database as of April 2019. OpenPowerlifting is creating a public-domain archive of powerlifting history. Powerlifting is a sport in which competitors compete to lift the most weight for their class in three separate barbell lifts: the Squat, Bench, and Deadlift.

The dataset contains over 22,000 meets and 412,574 lifters from competitions worldwide, with 1.4 million rows.
Most common powerlifting meets have athletes compete in squat, bench, deadlift with over a million participants in this kind of meet. Second most common is only participating in bench press with ~257000 participants.

Most common equipment class was single-ply followed by competing raw.

Competitors from all age ranges were represented with the oldest being as old as 97. The mean age was 31.5 and the standard deviation was 13.37.

The mean weight was 84.22kg with a standard deviation of 23.22kg. Note that this is not sorted sex.

The mean squat was 174kg with the highest squat being 575kg.
The mean bench press was 117kg with the highest bench press being 488kg.
The mean deadlift was 187kg with the highest deadlift being 585kg.
The mean total was 396kg with the highest total being 1368kg.
The mean Wilks was 288 with the highest Wilks being 779.

Note: these maximums are achieved with equipment and therefore googling deadlift world record would probably yield the raw deadlift world record (501kg) not the one here and also that there exist negative values for failed attempts that bring down the mean.

Over a million people competed in drug tested federations.

Athletes of 176 countries were represented in this dataset.

There were 222 powerlifting federations.

The first meet took place in 1964-09-05.

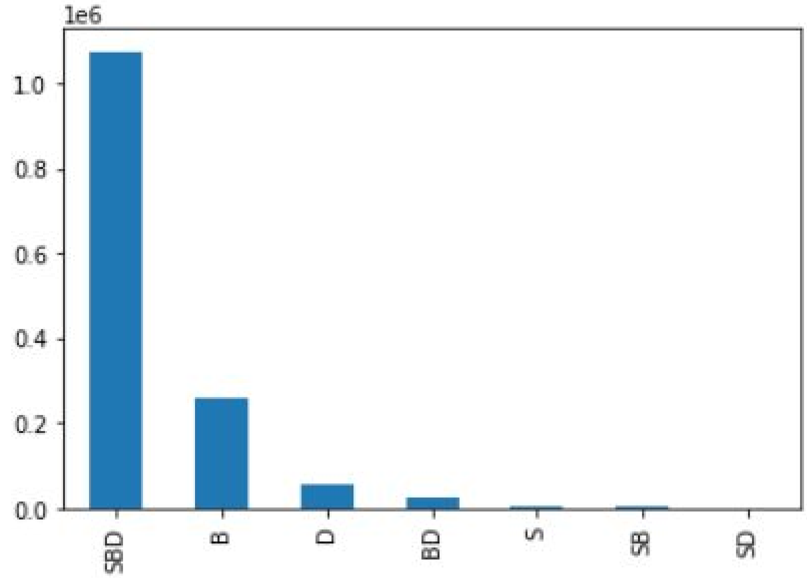
The most recent meet took place in 2019-04-20.

Most powerlifting competitions took place in the USA, with 96 countries having hosted a powerlifting competition.

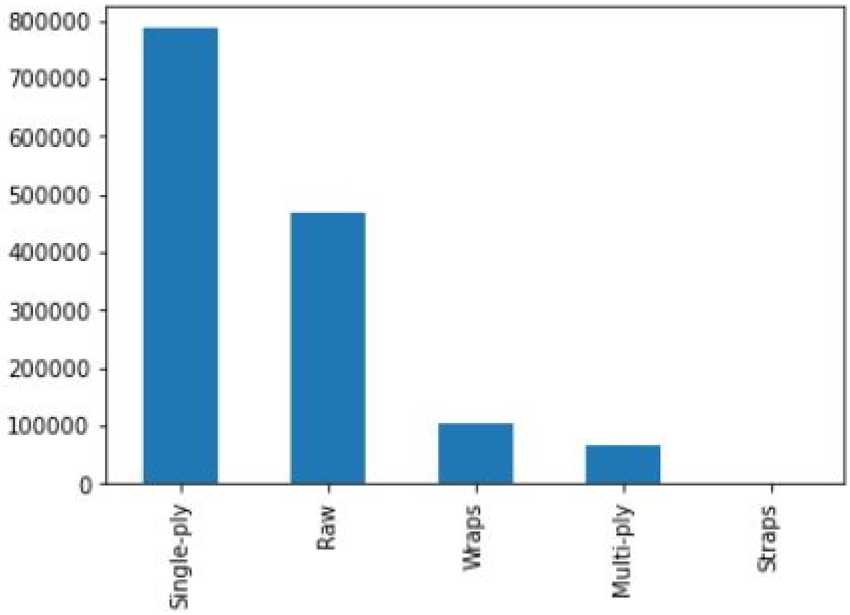
The most common meet was named “World Championships” with 32615 participants.

Exploring data

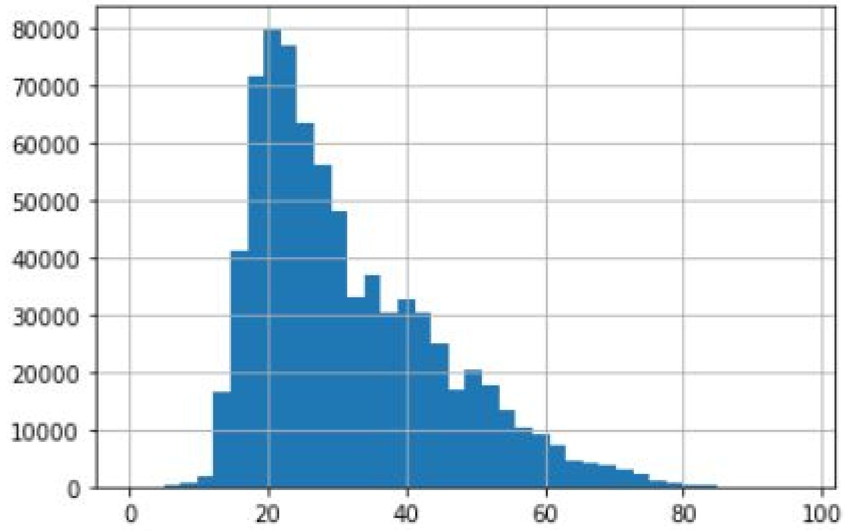
- Event distribution, note that y axis is in millions:



- Equipment distribution:

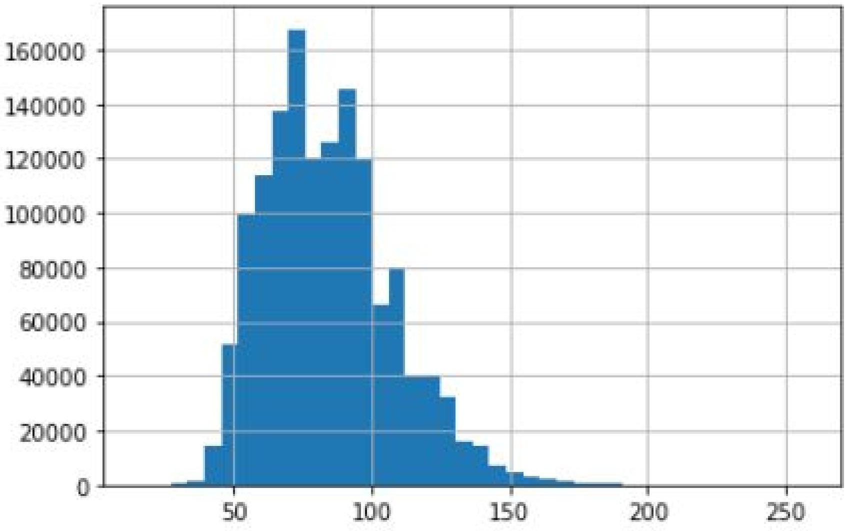


- Age distribution:



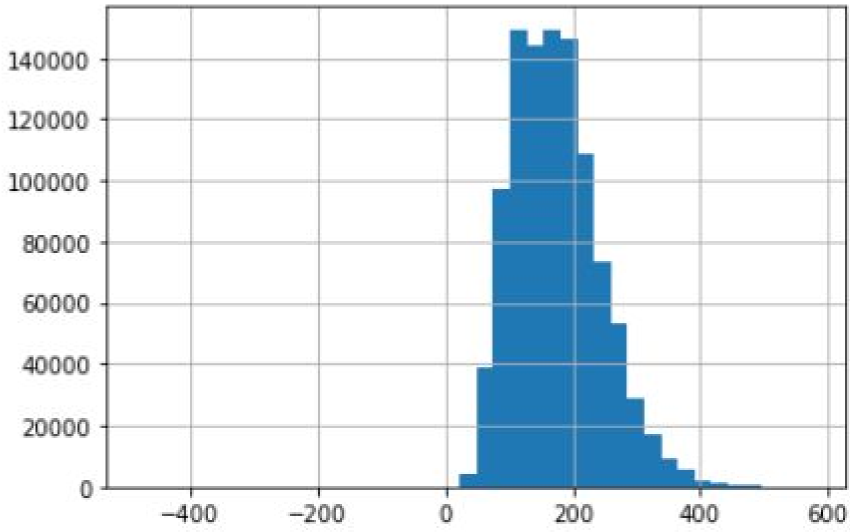
As division is not universal across federations and has way too many unique values it is impossible to make meaningful conclusions from it. Considering federations determine division from age, sex and weight, no information is lost.

- Bodyweight distribution (KG)

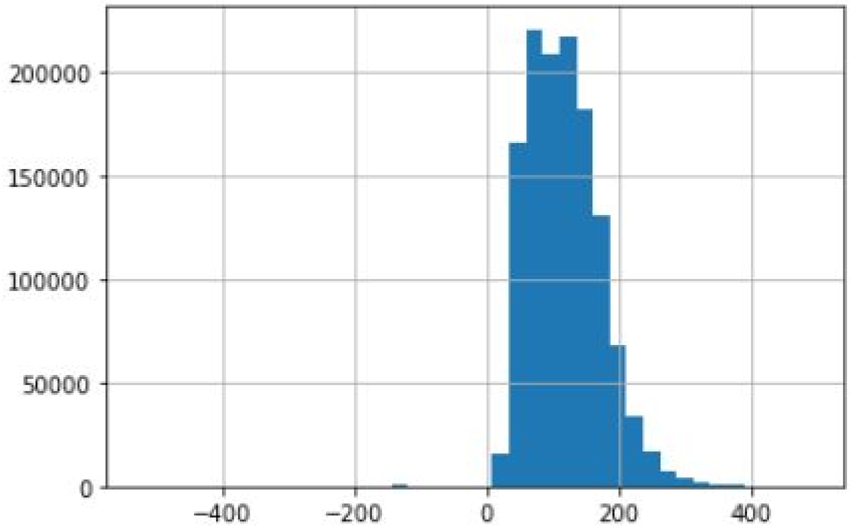


:

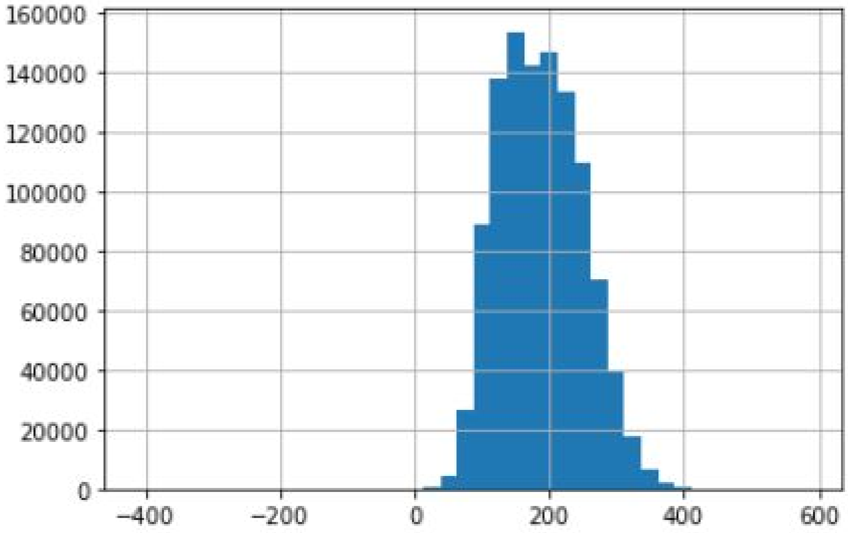
- Best squat out of 3 attempts:



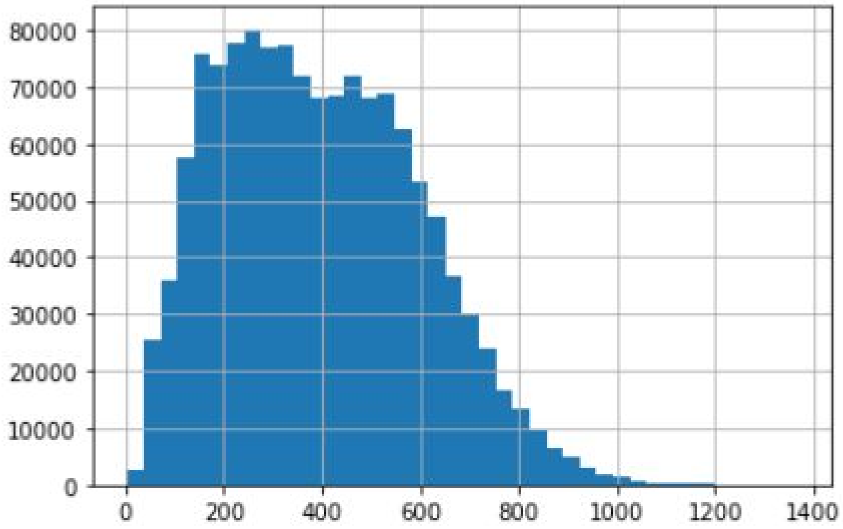
- Best bench press out of 3 attempts:



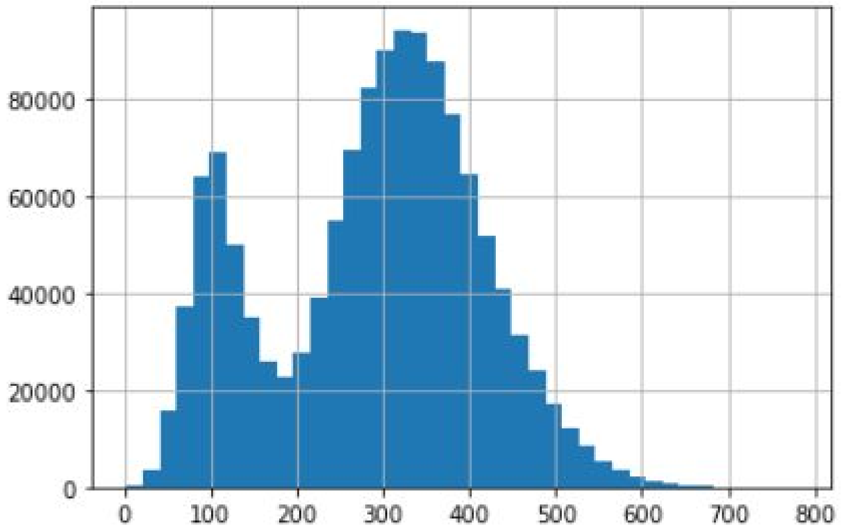
- Best deadlift out of 3 attempts:



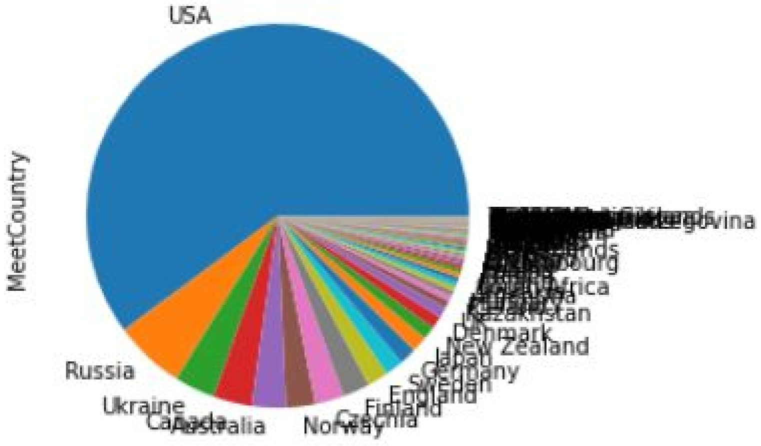
- Total distribution (sum of best lifts):



- Wilks distribution:



- Meet country pie chart:



Verifying data quality

The data quality is good enough for our purposes, but it needs some cleaning.

Project plan

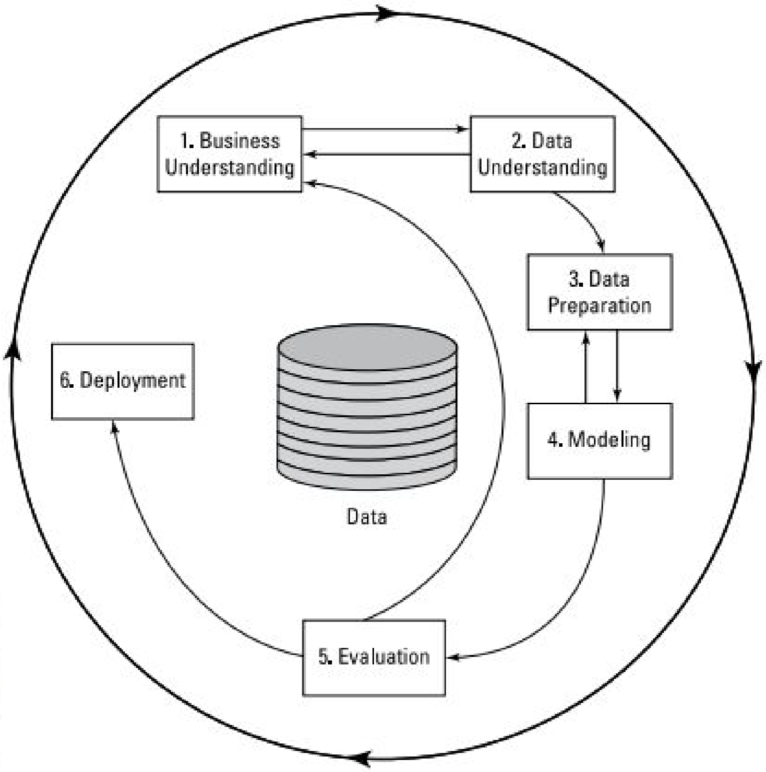


Figure 5-1: The CRISP-DM process model.

Business Understanding and Data Understanding

The details of business understanding and data understanding are described above.

Data preparation

1. Cleaning data
    a. Removing invalid values
    b. Removing NaN values
    c. Removing unnecessary columns
2. Adding new columns (like birth year, competition count)
3. Separating data into test and training set.
4. Separating training set to multiple cross validation sets

Modeling

5. Trying out different models like:
    a. KNN
    b. Decision Tree
    c. Random Forest
    d. SVM diffe
    e. K-Means
    f. etc
6. Finding the best model
7. Improve the best model

Evaluation

8. Testing model on test set
9. Comparing results to the goals set in business understanding.
10. Deciding if we have to remodel

Deployment

11. Reporting final results
    a. Making a final report
    b. Presenting the final report
12. Review project

Work distribution

We aim to split the work evenly, so in the case of a 60 hour project we should both work 30 hours. We are planning to use pair programming.

Methods and Tools

- We are working with Windows 10 and Linux Mint operating system.
- For coding we use Python 3
- To make code more understandable we use jupyter notebook.
- For project management we use git and github.
- For communication we use Discord and Messenger.
- We are planning to use virtual pair programming