Group - D5: UFC Fight Prediction

GitHub repository: https://github.com/Henrik895/FightPrediction

Team members
Henrik Lepson
Tanel Orumaa
Helar Jaadla

# Homework 10

# 1. Business understanding

## 1.1 Identifying business goals

### 1.1.1 Background

People have always wanted to predict the future and sport events offer plenty of opportunities for doing that. Professional sports usually have a clear set of rules that change relatively little from year to year and well defined criteria for declaring the winner, which makes them attractive for making predictions and thanks to the mass media and the internet the audience is larger than ever before. One such sport is mixed martial arts (MMA), which has gained in popularity during the last couple of decades and has thus transformed from a relatively unknown sport to a globally well established sport that has followers all over the world. The biggest MMA promoter in the world is Ultimate Fighting Championship (UFC), which organizes MMA fights almost every week year-round and their spot as a number one promoter is unrivalled as they have the best fighters. Because of its popularity UFC attracts a lot of people who bet smaller and larger sums of money on the fight results. These people are obviously interested in improving their predictions, even though there exist people who make bets without giving it a lot of thought (maybe their favorite fighter is fighting).

Knowing the rules and watching the previous fights of the fighters is important to making correct predictions, but the predictions can be still affected by our own biases (do we like the fighter or not, what country are they from, what background do they have and so on) and we can also miss a lot of factors that have an effect on the

outcome of the fight or instead give bigger importance to things that actually do not matter. Using available data and statistics as a basis for making decisions and predictions in sports like baseball, basketball, football etc has proven to be a successful strategy and for this reason it makes it reasonable to take the same approach in MMA as well and stop relying on our own subjective evaluations.

### 1.1.2 Business goals

The goal of the project is to use data analysis and statistical methods to create a better understanding of what influences the outcome of an MMA fight and use the gained knowledge to predict UFC fights with better accuracy.

### 1.1.3 Business success criteria

The project can be considered successful when we can create a model that can be reliably used for predicting UFC fight results.

## 1.2 Assessing your situation

### 1.2.1 Inventory of resources

- Three persons who are currently taking the Introduction to Data Science course (LTAT.02.002).
- The project is using the Ultimate UFC Dataset that is available on Kaggle, which includes all UFC fight results and participating fighter statistics between the years 2010 and 2020. At the time of writing this report the dataset was last updated on 20th November 2020.
- Each project group member has their own computer.
- Python and Jupyter Notebooks will be used for analysing the data and creating graphs for presenting the results.

### 1.2.2 Requirements, assumptions and constraints

The project has strict time constraints as the results of the project have to be presented on December 17th 2020. The results must be presented in the form of a poster. The project has no security or legal constraints as the data used for this project is publicly available for anyone.

### 1.2.3 Risks and contingencies

The main risk is the inexperience of all team members when it comes to data science projects, which might cause problems when more complex issues arise and thus make it difficult to achieve all project goals before the deadline.

### 1.2.4 Terminology

Business terminology:

- bout - traditionally means a boxing match, but is used to refer to MMA fights as well
- stance - whether the fighter is right handed (orthodox) or left handed (southpaw)
- reach - “wingspan” of the fighter, which is measured from the fingertips of one hand to the other hand when arms are raised to the sides (straight from shoulders)
- submission - when fighter yields to the opposing fighter (in MMA fights submission victory usually occurs as a result of choke hold, joint lock etc, but extremely rarely as a result of strikes as well)
- round - a regular MMA round lasts 5 minutes (non-title fights consist of 3 and title fights of 5 rounds)

Data-mining terms

- classification - prediction of the class of the data object
- label - the class that is being predicted in the classification task (in this project the label is the fight result)
- clustering - grouping data objects in such way that more similar objects are closer and different objects further away from each other (created groups are called clusters)

### 1.2.5 Costs and benefits

The project has no financial costs. The only cost is time, which is roughly 30 hours per group member (up to 90 hours combined for the project). In the case of success the project can help decrease the amount of bad fight predictions and thus limit betting losses, but the exact benefits depend on the person and their betting habits.

## 1.3 Defining data-mining goals

### 1.3.1 Data-mining goals

The first goal is to identify the attributes (such as fighter age, wingspan, height etc) that affect the fight outcome the most and visualize the findings using graphs and plots.

The second goal is to train a classification model that uses the most important attributes (results of the first goal) to predict fight outcomes. The predictions will most likely be probabilities as they offer us more information than just labels 0 and 1, especially when the fight should be close.

The third goal is to do cluster analysis on fighters so that we can stylistically separate them and present it as a scatter plot. There is a saying in fighting that “styles make fights” so there is a possibility that this information could be useful for predicting fight results as well and it can be used for improving the model.

### 1.3.2 Data-mining success criteria

The data mining goals can be considered successful when the model that is created can predict fight results with an accuracy of at least 70%. Every fight can end with one punch and sometimes two fighters are very evenly matched, which makes it impossible to create a perfect prediction model, so 70% accuracy can be considered sufficient.

# 2. Data understanding

## 2.1 Gathering data

### 2.1.1 Outline data requirements

For meeting the data mining goals it is necessary to have UFC fight results data that includes fight outcomes and also data about fighters who participated in these fights as well. The preferred data format is csv because data that is in this format can be easily analysed using Python and Jupyter Notebooks.

### 2.1.2 Verify data availability

Required data exists on the Kaggle (Ultimate UFC dataset) and it is available for everyone at the time of writing this report. This dataset contains information about fights between the years 2010 and 2020 and fighters who participated in these fights. There is also a possibility to include more data in the research should the initial dataset turn out to be insufficient as there are UFC datasets in Kaggle that include fight results starting from the year 1993.

### 2.1.3 Define selection criteria

The file that will be used in the dataset is “ufc-master.csv” as it contains all the necessary information for this project. Everything in this file that is related to the outcome of the fight (who won, how he/she won) and to the fighters themselves (height, reach, age, rank, previous records etc) will be used. This information is essential for this project and can not be left out.

### 2.1.4 Obtaining data

There were no problems with downloading the Ultimate UFC dataset from Kaggle and the “ufc-master.csv” file can be easily read in the Jupyter Notebook using the read_csv method from Pandas library that is meant for Python. The file looks exactly like it is described on the Kaggle page and there are no issues with data types and formats. The dataset is rather small (2MB in size) and therefore there are no hardware limitations when it comes to processing and can be done with almost any computer relatively quickly.

## 2.2 Describing data

The “ufc-master.csv” file contains information about ~4500 fights. Each fight is one row in the dataset and there are a little over 130 attributes for each fight. These 130 attributes give us information about both fighters between whom the fight was made (their previous record, average previous striking and takedown ratio, height, age, reach), the fight itself (who won, how he/she won, when he/she won, how many strikes were attempted and landed by each fighter, same for takedowns and how many submission were attempted and by who) and other more miscellaneous information like the venue, bettings odds before the fight and so on. The first

impressions about the data are good and there is plenty of data to analyse. This dataset can be certainly used for the project and there should not be any problems related to the dataset itself when trying to achieve goals that were set.

## 2.3 Exploring data

The values of the columns seem to be suitable for creating a classification model for predicting fight outcomes. Most of the values are numerical so they are easily comparable and it is also quite convenient that the unit of measurement used in the dataset for fighter reach and height is centimeters so they do not have to be converted anymore. Categorical features like gender, type of finish (KO/TKO, submission, decision etc), stance and so on can be one-hot encoded without any major difficulties.

Most columns in the dataset seem to be useful as of now, but there are some redundancies as well. For example there is one very puzzling column that has constant value 1 for every row, which gives us no information at all, and some columns like the reach difference between the two fighters can be easily calculated as the dataset contains the reach of both fighters separately as well. It is possible to drop these columns to make the dataset smaller without losing any information in the process, even though having reach and height differences as a separate column might be more convenient in the end than calculating them over and over again all the time.

The initial impressions after exploring this dataset are positive. For example, when looking at the reach difference between two fighters we can already see that this is an important attribute when predicting fight results and we can include it in the prediction model. Quick look into the data shows that fighters with a reach advantage of 10cm or more win about 56% of their fights and fighters with a reach advantage of 16cm or more have a win rate of over 61%. This is a promising early sign.

## 2.4 Verifying data quality

There are no signs of serious quality problems with the dataset and nothing of importance seems to be missing either. Should there be any problems with missing values or outliers during the project they can be fixed or verified using the official UFC website, which has information about fighters stats and fights as well.

# 3. Planning the project

## 3.1 Tasks

1. Preparing the dataset (this includes, but is not limited to getting rid of redundant features in the dataset, one-hot encoding categorical features and where necessary creating new features by combining others). As the dataset contains 137 features, each team member is expected to spend at least 3-5 hours on this task.
2. Analysing different features (how much do they affect fight outcome) and choosing the most important ones for creating the model. Each team member is expected to spend at least 10 hours on this task.
3. Testing different classification models and hyperparameter tuning in order to find the best model for predicting fight outcomes. At least 5 hours per team member.
4. Cluster analysis on fighters (fight styles). At least 3 hours per team member.
5. Creating visualizations and graphs for presenting results of tasks 2 and 4. At least 7 hours per team member.
6. Designing and creating a poster slide for the final presentation. Each team member should contribute at least 2-3 hours to make sure that the presentation is flawless.

## 3.2 Methods and tools

For creating the prediction model the initial plan is to try different classification algorithms like random forest trees and decision tree. KNN is also one of the possibilities.

The tools used for this project are Jupyter Notebooks and Pandas library for Python. For creating graphs, charts and other types of visualizations matplotlib and seaborn are going to be used.