Project A08 - Movie rating genius
Franklin Chukwuemeka (group 1), Lauri Lind (group 1), Oskar Saarepera (group 6)

HW10 report

Task 1 - Setting up

Github repository:
https://github.com/franklinchuks/Machine_Learning_Movie_Recommendation_and_Rating/blob/main/movie_ratings.ipynb

Task 2 - Business understanding

1. Identifying Business Goals:

Background:
The problem this project is tackling is the frequent and often frustrating question many of us seem to run into every once in a while "What movie should I watch?".

Business goals:
To help those in need of good recommendations this project aims to learn the user's preferences and then utilize a 5000-movie database based learning algorithm to pick out the most similar movies to the ones the user already likes.

Business success criteria:
The project will be deemed successful if the model predicts the ratings of movies with an accuracy of more than 90%. The second goal is for the model to suggest movies that the user is likely to watch based on the users' previous movie preferences.

2. Assessing the Situation:

Inventory of resources:
The project has 90 man-hours at its disposal, split between three data science students. Workload will be split optimally between the three participants, utilizing their previous experience and knowledge in cooperation with the data science skills obtained by the students.
The project will utilize the TMDB 5000 Movie Dataset. The set includes a wide variety of movies with information about the user rating, genre, keywords, budget, cast to name a few. All of this metainfo about the movies will be the input for the learning models.

Requirements, assumptions and constraints:
The completed project will be composed of a refined code and a visually appealing and informative poster. The deadline for the code is on 14th of december with the poster having to be posted on 11th.
The data necessary for this project has to include an abundant amount of movies with rating information in addition to other characteristics. The TMDB 5000 movie dataset fits this criteria and is publicly accessible.

Risks and contingencies:

The main risks related to the completion of the project include the dataset being too difficult to process or lack of computational power as there is no access to a high-end computer. Time is also a scarce resource as the project members all have lots of different responsibilities to address.
To mitigate these risks there are different movie datasets also available that might prove to be lighter to process if a problem of this sort arises. Also a comprehensive project plan is introduced to split the project into smaller tasks and help give all project members clear goals to work towards.

Terminology:
There are no specific terms that need to be described.

Costs and benefits:
As the goal of the project is not financial gain a cost-to-benefits analysis is not applicable. However, the main goal for this project is educational (gaining a deeper understanding of data processing and machine learning) and is quantitative in a sense the participants all expect to pass a university course as a direct outcome of this project. The cost is at least 30 man-hours per person and benefits include further knowledge and a completed course.

3. Defining data-mining goals

Data-mining goals:

- Explore correlations between different features and movie ratings.
- Identify key factors that influence movie ratings.
- Find the movies that are most similar to each other.

Data mining success criteria:
The model will be evaluated with metrics such as mean squared error for rating predictions and precision, recall and F1-score for the recommendation system. Project will be deemed successful if the predicted ratings have an error of less than 0.5, 90% of the time and the movie recommendations vaguely represent the reality of similar movie

Task 3. Data understanding

1. Gathering data:

Outline data requirements:
To achieve the goals of the project, the data used to train the necessary machine learning model needs to have enough attributes to predict the movie’s rating and find similarities between different movies. The time range of used data is not critical since the project covers movies of all timeframes.

Verify data availability:
Data for the project is available on the Kaggle.com environment and is free to use for everybody. In case this specific dataset should become unavailable, other similar datasets are also available on the same website.

Define selection criteria:
For this project, two datasets are going to be used. First dataset, tmdb_5000_movies, contains technical and descriptive information about 5000 movies and the second dataset,

tmdb_5000_credits contains data about people associated with those movies. The two datasets are linked with a movie id.

2. Describing data:

tmdb_5000_movies.csv:
The dataset includes various aspects of movies, such as financial details (budget, revenue), categorical data (genres, languages), and subjective metrics (popularity, vote average/count).
The data is a mix of numerical, textual, and nested JSON structures (for genres, keywords, production companies, etc.).
Variables: Budget, genres, homepage, original language, title, popularity, production companies, release date, revenue, runtime, spoken languages, status, tagline, vote average, vote count.

tmdb_5000_credits.csv:
Contains detailed cast and crew information for each movie.
The cast and crew data are in a nested structure, likely containing details like names, roles/positions, and possibly demographic information.
Variables: Movie ID, title, cast, crew.

The data seems suitable for analysis with respect to the project goals. It includes essential fields for building a recommendation system and analyzing movie ratings.

3. Exploring data:

tmdb_5000_movies.csv variables analysis:
Budget and revenue: Budget ranges from 0 to 380 million USD; Revenue ranges from 0 to 2.78 billion USD. Both fields show a right-skewed distribution with a significant number of movies having 0 as their value, indicating potential missing or placeholder values.
Runtime: from 0 to 338 minutes. Most movies have a runtime between 90 and 120 minutes. The presence of 0 minutes suggests data quality issues or missing data.
Vote average and vote count: Vote average ranges from 0 to 10, vote count ranges widely, indicating varying levels of viewer engagement and popularity. Vote average shows a normal distribution around the mean, while vote count is heavily right-skewed, indicating a few movies with exceptionally high vote counts.
Popularity: Varies significantly, showing the varying degrees of movie popularity.
Genres, Keywords, production companies and spoken languages: nested json structures of different strings which can be used to find similarities and correlations between movies and ratings respectively.

tmdb_5000_credits.csv analysis:
The credits database only has two unique variables: cast and crew which are both nested json structures. Further parsing needs to be done in order to analyze the diversity and composition of the variables.

4. Verifying data quality:

Quality Assessment:
movies database: Some fields have missing values (e.g., homepage, tagline). 0 values in budget, revenue, and runtime could indicate missing data.
credits database: No missing values detected, but nested JSON structures require careful handling.
Major Issues:
No severe issues detected that would prevent analysis, but attention needed for handling 0 values and missing data.
Possible Remedies:
Handling missing values (e.g., imputation, exclusion) based on analysis needs.
Parsing JSON fields for detailed analysis of cast and crew.

Task 4. Project plan

The project will be split into smaller subtasks and all of these will be allocated to different project members as follows. The estimated time for each task is also listed.
Starting the project:

- Choosing a project idea, meeting with team members, presenting the idea (Franklin, Lauri, Oskar, 2h)
- Creating a repository, preparing files. (Franklin, 1h)
- Creating business assessment, compiling a project plan, allocating work-hours (Oskar, 3h)
- Examining the given data, describing, exploring, verifying it. (Lauri, 3h)

Data Preparation:

- Data importing and cleaning. (Franklin, 3h)
- Feature selection and encoding. (Franklin, 3h)
- Data normalization. (Franklin, 2h)

Recommendation System:

- Data splitting and implementing KNN algorithm. (Franklin, 3h)
- Iterating the data structure, trying different approaches to get to optimal results (Franklin, 3h)
- Evaluation and visualization (Franklin, Lauri, Oskar, 3h)

Rating Prediction:

- Data splitting and model development: Implement various machine learning algorithms (Random Forest, KNN, SVM, etc) for rating prediction. (Franklin, Lauri, Oskar, 10h)
- Hyperparameter Tuning: Optimize the parameters of the models for improved performance. (Franklin, Lauri, Oskar, 1h)
- Evaluation and visualization (Franklin, Lauri, Oskar, 1h)

Documentation and Reporting:

- Writing a report, drawing conclusions, formulating the project. (Lauri, 4h)
- Ongoing documentation of various tasks in the project (Lauri, 3h)

Poster creation:

- Compiling the text and graphics for poster (Oskar, 2h)
- Forming the poster (Oskar, 2h)