Project title: Predictive Cinematics: A Model for Forecasting Movie Ratings

Team members: Lauri Leppik, Jan Erik Köst, Kaur Veere

Github repository: https://github.com/laurileppik/DS-Project

Business understanding

- Identifying your business goals
    - Background

The movie industry is a multi-billion dollar industry with a rich history and a wide range of products. Understanding the factors that contribute to movie ratings and popularity can provide valuable insights for filmmakers, distributors, and marketers.

    - Business goals

Examine how different features(revenue, runtime, popularity etc) of the movie impact the ratings of the movie: We plan to analyze the impact of different features on their ratings. This will help those interested to understand what factors contribute most to a movie's success and tailor their strategies accordingly.

Examine how movie popularity has changed over the years: By analyzing trends in movie popularity over time, we can gain insights into changing audience preferences and market trends. This can guide future investment and production decisions.

Examine how different movie descriptions impact the ratings of the movie: We plan to analyze the impact of movie descriptions on their ratings. This can help in crafting more effective descriptions and marketing content.

    - Business success criteria

The success of our first goal will be measured by the accuracy of our predictive model. A higher accuracy indicates that our model is effective in predicting movie ratings. We aim to create a model capable of accurately predicting at least 80% of the movie ratings correctly.

The success of our second goal will be determined by the insights we gain from our trend analysis. These insights should be actionable and provide value to stakeholders.

The success of our third goal will be determined by the insights we gain from our analysis. These insights should be actionable and provide value to stakeholders.

- Assessing your situation
    - Inventory of resources

We have a dataset of approximately 900,000 entries from the movie industry. This dataset includes a variety of features such as genres, budget, runtime, release date, original language, and popularity. We also have access to Jupyter Notebooks from where most of this project will be created. We use Github for version control. Our team consists of 3 second year university students who are currently taking a course named Introduction to Data Science in the University of Tartu and have no prior experience in the field.

    - Requirements, assumptions, and constraints

Requirements:

Business Understanding	29.11.2023
Data Understanding	01.12.2023
Data preparation	04.12.2023
Modeling	08.12.2023
Evaluation	11.12.2023
Presentation	14.12.2023

We must abide by university rules and regulations throughout making this project.

Assumptions:

We assume that our dataset is representative of the overall movie industry and that movie ratings are influenced by the features included in our dataset.

Constraints:

- Our dataset may consist of 900,000 entries, though a huge portion of that is unusable because of a lack of votes on each entry. (only around 25,000 entries have had more than 50 people vote for the rating of the movie)

    - Risks and contingencies

There is a risk that our model may not be accurate if our assumptions are incorrect or if our data is not representative or complete. In this case, we may need to revise our assumptions.

We may also face challenges in interpreting the results of our text analysis due to the subjective nature of movie descriptions.

○ Terminology

We use standard terminology from the movie industry and data science.

○ Costs and benefits

The main costs of our project are the time and resources needed to preprocess the data, develop the model, and perform the analysis. Financial cost of this project is 0 euros. The potential benefits are the insights we can gain about the movie industry, which can guide decision-making for filmmakers, distributors, and marketers.

● Defining your data-mining goals

○ Data-mining goals

1) Predictive Modeling: Develop a predictive model that can estimate movie ratings. The deliverables will include:

   Predictive Model: A machine learning model trained using algorithms such as Linear Regression and Decision Trees. This model will predict movie ratings based on features such as genres, budget, runtime, release date, original language, and popularity.

   Model Evaluation Report: A detailed report evaluating the performance of the model using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

2) Trend Analysis: Analyze trends in movie popularity over the years. The deliverables will include:

   Trend Analysis Report: A comprehensive report detailing the trends identified in movie popularity over the years, supported by visualizations.

3) Text Analysis: Analyze the movie descriptions to identify common themes, keywords, or phrases that may impact movie ratings. The deliverables will include:

   Text Analysis Report: A comprehensive report detailing the themes, keywords, or phrases identified in movie descriptions and their potential impact on movie ratings, supported by visualizations.

   Keyword-Performance Model: A machine learning model trained using algorithms such as Naive Bayes or Support Vector Machines. This model will predict movie ratings based on the presence of certain keywords or phrases in the movie description.

○ Data-mining success criteria

Goal 1:

The predictive model should have an accuracy higher than a predetermined threshold. This can be measured using metrics such as R-squared for regression models or accuracy for classification models.

Goal 2:

Statistical Significance: The trends identified in movie popularity over the years should be statistically significant. This can be measured using statistical tests such as the Mann-Kendall trend test or Spearman's rank correlation.

Visual Confirmation: The identified trends should be visually apparent in the data. This can be confirmed by plotting the data and visually inspecting for trends.

Goal 3:

Keyword Correlation: The identified keywords or phrases in movie descriptions should have a significant correlation with movie ratings. This can be measured using correlation coefficients.

Data understanding

Gathering data

Outline data requirements:

For achieving our data mining goals we need to have data that must contain the following information:

- Movie title
- Average rating
- Number of people who voted for average rating
- Release Date
- Revenue
- Runtime
- Whether the movie is an adult movie or not
- Original language
- Popularity of the movie
- Description of the movie
- Genres
- Languages spoken in the movie

We use .csv for all of our tabular data storing.

Verify data availability:

The data is available on kaggle:

https://www.kaggle.com/datasets/asaniczka/tmdb-movies-dataset-2023-930k-movies?rvi=1

Describing data

We use 4 different .ipynb NoteBooks and 3 different .csv files for our project. Notebook 1 (Cleaning_Data.ipynb) is used for cleaning the dataset. Notebook 2 (Predict_movie_ratings.ipynb) is used for creating and validating the accuracy of several predictive models. Notebook 3 (Popularity_Through_Years.ipynb) is used for looking for and analyzing trends that have changed throughout movie history. Notebook 4 (Examine_Ratings_Description.ipynb) is used for analyzing what impact movie descriptions have on the popularity of movies. CSV 1 (original_TMDB_movie_dataset_v11.csv) (402MB) is the original dataset. CSV 2 (fil_data_movies_lessthan10k_removed.csv) (4.74MB) is the cleaned dataset in which movies with less than 10 000 dollars in budget or revenue have been removed. CSV 3 (fil_data_movies_lessthan10k_replaced.csv) (15.9MB) is the cleaned dataset in which movies with less than 10 000 dollars in budget or revenue have been replaced by the median value of all budgets and revenues.

CSV1:

23 columns (id, title, vote_average, vote_count, status, release_date, revenue, runtime, adult, backdrop_path, budget, homepage, imdb_id, original_language, original_title, overview, popularity, poster_path, tagline, genres, production_companies, production_countries, spoken_languages)

- id - id of the movie in the CSV
- title - title of the movie
- vote_average - average rating for the movie
- vote_count - count of people who voted for the rating
- status - is the movie released or not
- release_date - release date of the movie
- revenue - revenue that the movie generated
- runtime - runtime of the movie
- adult - is the movie an adult movie or not
- backdrop_path - path of the backdrop of the movie
- budget - budget of the movie
- homepage - link to the homepage of the movie
- imdb_id - imbd id for the movie
- original_language - the language in which the movie was released
- original_title - original title of the movie
- overview - description of the movie
- popularity - how popular the movie is on a scale of 0 to 3000
- poster_path - path of the poster of the movie
- tagline - tagline of the movie
- genres - genres of the movie

- production_companies - production companies of the movie
- production_countries - production countries of the movie
- spoken_languages - languages spoken in the movie

CSV2: same but without status column

CSV3: same but without status column

The data includes 900,000 entries but when we narrow it down to entries that have at least a sufficient amount of rating voters (we chose at least 50 people to be sufficient), we are left with about 28,000 entries, which is suitable for our goals.

Exploring data

NOTE!: These ranges are taken from the dataset after removing the entries with less than 50 voters removed. Some columns are described as irrelevant meaning that these columns only have unique values. Here are the ranges of which the values are from and the distribution of these ranges. For categorical data with more than 10 entries, 10 most popular are described here.

- id - irrelevant
- title - irrelevant
- vote_average - 1.840 - 9.980, normal distribution
- vote_count - 50 - 34495, a lot more of the values are closer to 50
- status - all of the values are “relased”
- release_date - 1870s to 2020s, a lot less of the movies in this dataset made before 1950s
- revenue - 0 - 2923706026, since a lot of entries are 0, data is skewed towards 0
- runtime - 0-585, most movies are between 1 and 2 hours with some exception
- adult - True and False, only 13 movies are adult movies in this dataset
- backdrop_path - irrelevant
- budget - 0 - 460000000, since a lot of entries are 0, data is skewed towards 0
- homepage - irrelevant
- imdb_id - irrelevant
- original_language - en, fr, it, ja, es, de, ko, hi, zh, ru, most movies have english as the original language
- original_title - irrelevant
- overview - irrelevant
- popularity - 0.600 - 2994.357, most values are between 0 and 100
- poster_path - irrelevant
- tagline - irrelevant
- genres - Comedy (2337), Drama (2078), Drama, Romance (831), Comedy, Drama (807), Documentary (694), Horror (643), Comedy, Romance (640), Horror, Thriller (473), Comedy, Drama, Romance (401), Drama, Comedy

- production_companies - Paramount (206), Universal Pictures (184), 20th Century Fox (178), Metro-Goldwyn-Mayer (172), Warner Bros. Pictures (153), Walt Disney Productions (133), Columbia Pictures (120), RKO Radio Pictures (63), Pixar (63), The Asylum (59)
- production_countries - USA (11845), France (1629), United Kingdom (1196), Italy (1072), Japan (1001), United Kingdom, USA (638), Canada, USA (502), India (465), Canada (452), South Korea (380)
- spoken_languages - English (13808), French (1612), Italian (1120), Japanese (936), Spanish (753), English, Spanish (506), English, French (471), No Language (395), English, Italian (353), German (315)

Verifying data quality

We noticed a number of movies having the same name. We removed all of the duplicates and kept only the movie with the largest amount of votes. This resulted in losing around 1500 entries.

We removed the column status entirely, since all of the values in this column are the same.

We removed all of the movies with a runtime of 0, since its rather unlikely that a movie lasts 0 minutes. This resulted in losing around 100 entries.

Main weakness in our data (that has less than 50 voters removed) is the fact that we lack adequate entries for budget and revenue columns since a lot of these entries have a value of 0. We fixed this problem creating two different filtered datasets, one of which has all of the revenus and budgets less than 10000 replaced with median value of all of the other entries in the dataset. In the second dataset we removed all of these entries. (and are left with around 7000 entries)

Other weaknesses include a small amount of missing values in each of the columns. Which we can just remove or ignore without causing problems in our data quality.

The data we have is not ideal since we would like to have a dataset with a larger amount of people having voted for the rating (vote_count column), but its the best we could find and sufficient for our goals.

Planning your project

Task1: Create a PDF-file using CRISP-DM methods for developing a business understanding of the project. - Lauri (4 hours)

Task2: Create a PDF-file using CRISP-DM methods for developing a data understanding of the project. - Jan (4 hours)

Task 3: Make the project plan. - Kaur (1.5 hours)

Task 4: Prepare and clean the data for general use throughout all of the .ipynb files. - Kaur, Lauri (both 4 hours)

Task 5: Prepare the data for use inside separate .ipynb files (i.e remove irrelevant columns for these files specifically) - Kaur, Jan, Lauri (each 3 hours)

Task 6: a) Create a predictive model and test validate its accuracy inside Predict_movie_ratings.ipynb. - Kaur (10 hours)

b) Create a trend analysis report by analyzing the data throughout years inside Popularity_Through_Years.ipynb, - Jan (10 hours)

c) Train a Keyword-Performance Model for accurately predicting movie ratings and validate its accuracy inside Examine_Ratings_Description.ipynb - Lauri (10 hours)

Task 7: Evaluate the findings and make conclusions. - Lauri, Jan, Kaur (4 hours each)

Task 8: Make a poster for presenting the findings and conclusions. - Lauri, Jan, Kaur (6 hours each)

Task 9: Present the data. - Lauri, Jan, Kaur (3 hours each)

Tools we plan on using:

Github, Jupyter Notebook, Python and its libraries, Google Docs for writing documentation.