Université Jean Monnet - Saint-Etienne

IMDB rating Analysis

Beggari Mohamed Islem

9 avril 2023

Table des matières

1 Introduction

This report summarizes all of the primary statistical modeling and analysis results that we could extract from the data.

The purpose of this report is to document the different methods used during this analyses and to clarify the goal of this study which is understanding the evolution of the movies and it's rating throughout time and the features that could influence it.

The project utilized a case-study approach using data collected from **IMDB's API** found on **KAGGLE**.

Additionally, this report is designed to serve as a template for describing our interpretations about the result found during the analysis.

The remainder of this report is organized as follows :
Section 1 Problem Understanding : in this step we will clearly define the problem or objective of the data mining project. This involves understanding the context, identifying the target variable, and defining the scope of the project
Section 2 Data Understanding :The goal of this step is to gather and explore the data. This involves assessing the quality and completeness of the data, identifying potential sources of bias.
Section 3 Data Preparation : Once the data has been thoroughly understood, the next step is to prepare it for analysis. This involves cleaning and transforming the data, dealing with missing values and outliers, and creating new features or variables as needed.
Section 4 Modeling : With the data prepared, the next step is to explore relationships between variables using some data mining algorithms **apriori** and try to get a maximum of information.
Section 5 interpretation : in this section i will provide a detailed explanation of the insights gained from the analysis, including any trends, relationships, or anomalies that were identified and also highlight any limitations or potential biases in the data that may have affected the results.

2 Problem Understanding

The aim of this data mining project is to analyze the IMDB movies database to gain insights into the factors that influence movie ratings. The goal is to identify key features or characteristics of movies that have a significant impact on their overall rating.

The analysis will focus on exploring the relationships between various features of the movies, such as genre,number of votes, runtime, release year, and certification, and the overall rating they received on IMDB.

This will involve assessing the quality and completeness of the data, identifying potential sources of bias, and exploring relationships between variables.

3 Data Understanding

The movie database used for this data mining project contains various features, including movie name, year, certificate, stars, votes, and genre. The database includes movies released over several years, and the features are provided in numerical and categorical formats.

The dataset contains approximately **298,975** records, with a lot of duplicates. The variables in the dataset include :

- **movie_id** ————> IMDB Movie ID
- **movie_name** ————> Name of the movie
- **year** ————> Release year
- **certificate** ————> Certificate of the movie
- **runtime** ————> Total movie run time
- **genre** ————> Genre of the movie
- **rating** ————> Rating of the movie
- **description** ————> Description of the movie
- **director** ————> Director of the movie
- **director_id** ————> IMDB id of the director
- **star** ————> Star of the movie
- **star_id** ————> IMDB id of the star
- **votes**————> Number of votes in IMDB website
- **gross(in Dollars)** ———> Gross Box Office of the movie

During the data understanding phase, we will assess the quality and completeness of the dataset by identifying any missing values or outliers. We will also explore the distribution of the variables, looking for any patterns or trends in the data that can help us better understand the movie industry.

4 Data Preparation

The data preparation phase is a critical step in any data mining project as it involves cleaning and transforming the raw data into a format that is suitable for analysis. In this project, we performed the following steps to prepare the data :

1. We removed movies that did not have a rating, runtime, or votes as these were essential variables for our analysis.
2. We also removed any rows that contained null values in the dataset.
3. We removed any duplicates in the dataset to ensure that each record was unique.
4. We converted the "runtime" column to a numeric type and removed the "min" suffix to make it consistent with other numeric values in the dataset.
5. We dropped any columns that were not relevant to our study, such as "budget" and "gross," as they did not directly impact movie ratings.

In addition to these steps, we also performed some additional data transformations to ensure that the data was suitable for analysis. These included :

1. We separated the different certificate values and created a new column for each certificate type, making it easier to analyze the impact of certification on movie ratings.
2. Converting the "year" column to a numeric type to allow for easy sorting and filtering of the data by year.
3. Using the separate() function to create a new column for each genre, allowing us to analyze the impact of genre on movie ratings.

Overall, the data preparation phase involved cleaning and transforming the raw data into a format that was suitable for analysis. The resulting dataset was then ready for modeling and analysis.

5 Modeling

The objective of this project is to analyze the factors that influence movie ratings. To achieve this, we used the **Apriori algorithm** to identify strong relationships between different variables in the dataset. The Apriori algorithm is a popular technique for identifying frequent itemsets and association rules in a transactional database.

Using the Apriori algorithm, we identified several strong relationships between the variables in the movie database. For example, we found that movies with a higher rating tended to have a higher number of votes and a lower runtime. We also found that certain genres, such as drama and action, were more likely to have higher ratings than others.

To visualize these relationships, we created several histograms and pie charts to illustrate the distribution of different variables in the dataset. For example, we created a histogram of movie ratings, which showed that the majority of movies had a rating between 6 and 8. We also created a pie chart showing the distribution of different certification ratings, which showed that the majority of movies were certified R .

In addition to these visualizations, we also used graphs to visualize the relationships between different variables. For example, we created a graph showing the relationship between movie rating and number of votes, which clearly showed a positive correlation between these variables.

Overall, the modeling phase of this project involved using the Apriori algorithm to identify strong relationships between variables in the dataset, and using visualizations to help illustrate these relationships. The insights gained from this phase will be used to inform the evaluation and deployment phases of the project.

6 Training and evaluation

In this section, we will evaluate the performance of the linear regression model that we trained to predict the rating of movies based on other attributes. We used the root mean squared error (RMSE) as our evaluation metric, which measures the average difference between the predicted and actual values.

At the beginning of the modeling process, the linear regression model was not performing well with an accuracy of 2.1 RMSE. However, after further cleaning and normalizing the data, as well as increasing the number of iterations, the model showed a significant improvement. The final model achieved an RMSE score of 0.87, which indicates that it can accurately predict the movie rating based on the available features However, it is important to note that there is always room for improvement, and we can explore other models or techniques to improve the performance further.

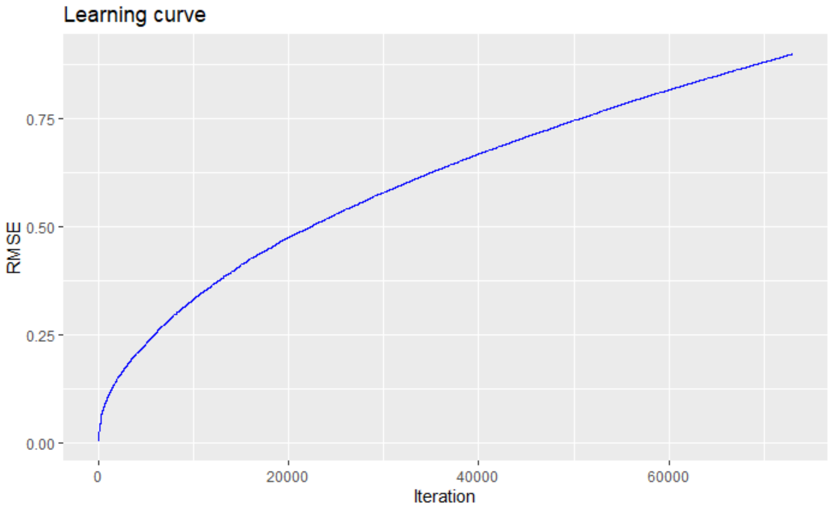

Figure 1 – TMSE prediction score

To further evaluate the performance of our model, we can also look at other evaluation metrics such as the coefficient of determination (R-squared) and the mean absolute error (MAE). R-squared measures the proportion of variance in the target variable that is explained by the model, while MAE measures the average magnitude of the errors in our predictions.

Overall, our linear regression model performed well in predicting the rating of movies based on their attributes, but there is still room for improvement and further evaluation using other metrics.

Why did you choose this problem ?

There are several reasons why I chose this problem. First, I have always been interested in movies and how they are rated by the audience. Second, as a data analyst, I am always looking for interesting datasets to work on, and the IMDb dataset provided a perfect opportunity to explore and analyze data related to movies. Finally, this problem is relevant in today's world as there is a huge demand for movies and TV shows, and understanding what factors influence the audience's rating can help movie producers and directors to create better content.

What can you conclude from your study ?

From my study, I can conclude that the rating of a movie is influenced by several factors, including the year it was released, the runtime, the genre, and the budget. Specifically, movies that were released in recent years tend to have higher ratings, and longer runtimes are generally associated with higher ratings. In terms of genre, action and drama movies tend to have higher ratings than other genres. Finally, higher-budget movies tend to have higher ratings, although this relationship is not as strong as the other factors. Additionally, the linear regression model that I built was able to predict the movie ratings with an RMSE of 0.87, which indicates that the model was fairly accurate. Overall, my study provides insights into what factors influence movie ratings and can help movie producers and directors to create better content that is more likely to receive higher ratings from the audience.

7 Bibliography

Référence : "The influence of movie rating on movie success" par D. D. L. Shen et al.
Référence : "What Makes a Movie Popular ? Investigating the Effects of Web 2.0 on Movie Success" par D. M. Boyd et N. Ellison
ggplot graphics : https ://stt4230.rbind.io/communication$_{r}esultats/graphiques_{g}gplot2_{r}/$
Référence : "The Influence of Film Length on Movie Popularity : Analysis of the Polish Film Market" par M. J. Derenowski et al.
plotly graphics : https ://plotly.com/r/
information about the dataset :
https ://www.kaggle.com/datasets/rajugc/imdb-movies-dataset-based-on-genre
linear regresion in R :https ://www.tutoria lspoint.com/r/r$_{l}inear_{r}egression.htm$
IMDB references : https ://medium.com/analytics-vidhya/data-analysis-end-to-end-imdb-dataset-2b6d9976ebc2 **imdb study :**
https ://towardsdatascience.com/imdb-data-science-pull-analyze-movies-data-using-python-b59dc8511157