



**Early Movie Rating Prediction Through Natural Language  
Processing and Machine Learning**

**Annie Jose**

**Word Count: 7178**

**Research Report submitted in part fulfilment of the degree of  
Master of Science in Business Analytics**

**2022 - 2023**

**Queen's Management School**

## Declaration

This is to certify that:

- i. The portfolio comprises only my original work;
- ii. AI technologies (e.g. chat GTP) have not been used in the writing of the portfolio dissertation.
- iii. No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.



---

[ Candidate's Signature ]

08 August 2023

---

Date

Annie Jose

---

Printed Name

## **Abstract**

The movie industry continues to remain dynamic, with high stakes and uncertainty. Historically, performance forecasting was made based on instances and past trends. However, the emergence of data analytics shows promising opportunities to utilize big data and make timely operational decisions. This study aims to predict the performance of a movie before its release through Machine Learning models. In addition to using metadata, the extent of influence made by unstructured textual data is evaluated through Natural Language Processing (NLP). This brings transparency to the models and provides a better understanding of the impact made by the different attributes. The capabilities of feature selection and hyperparameter tuning for improving model performance are also evaluated.

The results showed that ensemble models such as Random Forest, Extreme Gradient Boosting and Support Vector Regression, were more successful in capturing the complex interactive effects. Including text vectorizations along with base variables consistently improved the model performances. It was also found that runtime, actors, and genre were the dominant attributes that impacted movie ratings. Even though there is room for improvement, this study was successful in creating a simple and efficient method to incorporate numeric, categorical, and textual data. It further sets the foundation for further exploration into the utilisation of unstructured data for predictive analytics.

## Table of Contents

Declaration .....	ii
Abstract .....	iii
Table of Contents .....	iv
List of Figures .....	vi
List of Tables .....	vii
1. Introduction.....	8
1.2. Motivation.....	8
1.2. Research Objectives .....	9
1.3. Report Structure .....	10
2. Literature Review.....	11
2.1. Overview of Research Area .....	11
2.2. Role of various attributes in predicant movie ratings .....	12
2.3. Application of Machine Learning .....	14
2.4. Summary .....	16
3. Methodology .....	17
3.1. Method of Study .....	17
4. Findings.....	24
4.1. Understanding the relationship between the Variables.....	24

4.2. Machine Learning Models .....	30
4.3. Discussion of Theoretical Impact .....	34
4.4. Implications for Business.....	35
5. Conclusion .....	36
References.....	37
Appendix 1: Additional Information .....	44
Appendix 2: Dissertation Checklist Sheet .....	46

## List of Figures

Figure 1. The CRISP-DM Process.....	18
Figure 2. Distribution Curve of Average IMDB Rating .....	25
Figure 3. Year Wise Popularity for Movies.....	25
Figure 4. Month and Day Wise Split of Movie Releases .....	26
Figure 5. Bivariate Relation Between Rating and Duration .....	27
Figure 6. Interaction Rates Between Genres.....	27
Figure 7. World Cloud Based on Movie Title .....	29

## **List of Tables**

Table 1. Description of variables in the IMDB Dataset.....	20
Table 2. Description of Features extracted through TMDB API.....	20
Table 3. Performance Comparison of Linear and Ensemble Models .....	32
Table 4. Feature Importance Scores.....	33

# **1. Introduction**

## **1.2. Motivation**

In 2022, the global film industry's revenue weighed \$26 billion with more than half the contributions from North America and Asia (Frater, 2023); even though these were well below the pre-pandemic levels, the current statistics show promising growth for the industry. Often theatre releases fail to perform as the producers expect, in both revenue and ratings. For instance, movies like 'The Secrets of Dumbledore' and 'Lightyear', which cost more than \$200 million to produce did not generate box office collections as expected (Clark, 2022).

Considering the high volatility of the movie industry, accurate and timely predictions about movie performance and user acceptance can help the respective stakeholders strategise their operations for the future (Cizmeci and Oguducu, 2018). Especially for investors and production houses, early insight into a movie's profitability can help them decide its feasibility, budget, cast, crew, and so on (Chamani et al., 2021; Abidi et al., 2020; Ghiassi et al., 2015). Additionally, it would help them to optimally allocate resources between competing projects (Ghiassi et al., 2015), and plan marketing and distribution. Furthermore, this would aid data-based decision-making for theatre owners and streaming services, since majority of the licencing deals are finalized before a movie's theatrical debut (Ghiassi et al., 2015). Such data-oriented observations would also be beneficial to other parties such as actors, directors, and other crew members, for identifying current trends and choosing their projects (Chamani et al., 2021).

The accelerated growth of big data and analytics attracted various industry experts and academics to explore its applicability to predict the popularity of movies. Numerous studies have been made in this domain using customer reviews and comments from social media platforms, and its metadata. These studies unanimously agree on the suitability of machine learning methodologies



to predict a movie's success in terms of popularity, rating, and profitability (Won and Kim, 2022; Mhowwala et al., 2020; Ning et al., 2020).

One of the prominent methods of analysing user responses to new releases is sentiment analysis. Utilizing the data from social media platforms to analyse user responses to movies, provides a clear understanding of how the audience ranks the movie (Sahu and Ahuja, 2016). However, this classification approach fails to study critical features, such as genre, release date, actors, crews, genre et cetera, and their interrelationships. Moreover, there is a significant challenge in translating these insights into business actions. The study conducted by Ahmad et al. (2023), discovered a strong correlation between the rating for an upcoming movie and factors such as genre, crew, language and so on. Therefore, an alternative is the application of regression models that can utilize both user responses and metadata (Chamani et al., 2021).

## **1.2. Research Objectives**

Previously, researchers were primarily focused on the development and fine-tuning of Machine Learning algorithms. Even though their works contributed immensely towards accurate predictions of movie success, further investigation is required regarding the multivariate relationships and optimization of these features, to enhance customer outreach (Abidi et al., 2020). Additionally, the application of natural language processing on textual data, such as movie title, overview, and script, for identifying patterns and customer preferences were unexplored.

Inspired by these preceding works, this study aims to utilize and extract relevant textual as well as metadata for forecasting box-office success pre-release through Machine Learning. Specifically, this study aims to answer the following research questions:

1. What are the important factors in predicting a movie's rating?
2. How can attributes be used to recognize customer interests and preferences?
3. How do the performance and interpretability of linear, ensemble and non-parametric models differ in predicting a movie's rating?

To understand the influence of various attributes on movie ratings, the IMDb dataset has been used. Through adequate feature extraction, feature selection, and tuning, machine learning model performance was enhanced. Furthermore, by combining the models with appropriate statistical exploration methods, different patterns between the variables were derived and their extent of influence over the models were quantified.

### **1.3. Report Structure**

To methodically present the bivariate and multivariate relationships and evaluate the capabilities of machine learning as well as natural language processing for predicting the ratings for movies based on pre-release meta-features, this report is subdivided into four sections. The upcoming section contains a critical evaluation of related works and an outline of various approaches traditionally used for predicting movie ratings from sentiment analysis, regression, and classification models, to advanced neural networks. Section three presents the research process used for this study, which is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM), and the rationale behind it. The findings of the various quantitative analysis, graphical representations, modelling, and model assessments are presented in section four. This chapter further compares the results to the current theories and discusses their implications for businesses. The final chapter summarises the conclusion with respect to the contributions towards Machine Learning and Natural Language Processing (NLP), while reflecting on the work as a whole and presents the prospects for future work.

## **2. Literature Review**

This chapter gives a brief overview of the analytical research in the movie industry. Initially, an outline of the usage of customer reviews and ratings for sentiment analysis is provided, followed by the significance of various attributes related to the movies. Subsequently, the suitability and performance of linear, ensemble and deep learning models for predicting the performance of a movie is discussed. Finally, the challenges related to this field of study are evaluated, along with the future scope.

### **2.1. Overview of Research Area**

By combining free-flowing user-generated data and analytical techniques such as sensitivity analysis, data mining, machine learning and natural language processing, the movie industry has made significant progress in the last decade towards analytics-based decision-making. One of the finest illustrations of the effectiveness of data mining and analytics in the movie industry for market capture is Netflix's algorithm that accurately makes suggestions to customers based on their interests (Van Es, 2023). Utilizing customer data to recognize patterns in their interests and preferences makes revenue generation possible in the movie industry (Wang and Zhang, 2018).

Among all the algorithms that are currently used in the movie industry, the most popular one is Collaborative Filtering Techniques. Other recommendation algorithms include Content-Based Filtering, Hybrid Filtering and Deep Learning Based Methods (Sunilkumar, 2020). However, Collaborative Filtering Techniques are a simple and effective way to predict a user's interest based on others' past choices, leveraging the idea that there would be a similarity between the decisions of people who share the same viewpoint (Thakker et al., 2021). (Jha et al., 2022) confirms its superiority over traditional methods, to measure similarity between users, even for users who have given limited ratings.

However, Collaborative Filtering is far from perfect. Researchers unanimously agree on its limitations such as cold start problems, data sparsity and scalability (Ifada et al., 2020; Jha et al., 2022; Sunilkumar, 2020). In a comparative study between Collaborative Filtering and Hybrid Recommender Systems, conducted by Ifada et al. (2020), it was found that there were no significant differences in the performance of both models. However, hybrid models were better at predicting with fewer variables. Alternatively, observations of the publicly deployed recommender systems show that optimal results are made when these algorithms are used in a complimenting manner rather than exclusively.

Sentiment Analysis is another important tool that is used for understanding viewer responses to movies from textual data on the internet, such as reviews, comments, and posts across social media platforms. By analyzing textual data through Natural Language Processing techniques; keywords, linguistic patterns, context, and emotions are identified which sheds light on the underlying user preferences which can help with marketing design (Cui et al., 2012). This information is key to enhancing machine learning models and recommendation systems. However as Rahmani et al. (2019) rightly point out, for revenue generation and operational decision-making, it is more beneficial to have market insights pre-release or even pre-production.

## **2.2. Role of various attributes in predicant movie ratings**

Across the various studies related to the prediction of movie ratings, primarily two types of variables were used, first one is the textual data and the second one is the meta-features (Lash and Zhao, 2016). Unarguably the performance of the prediction model is correlated with the combination of the right attributes. Multiple studies show improved performance through feature extraction (Kim et al., 2020; Jagdale and M., 2019; Antipov and Pokryshevskaya, 2017). However, feature selection should be about the context of the study and research questions, otherwise, it will

be less meaningful for managerial insights and implementation. (Chamani et al., 2021) provides an instance for this by pointing out that including features such as revenue and budget may not be useful to predict movie performance since those insights are more necessary before the movie's release. The study conducted by Lash and Zhao (2016), shows that director rank, studio rank, genre, runtime, and popularity are some of the most influential aspects that decide a movie's success. Abidi et al., (2020) also had similar results from their analysis, which showed director, studio and genre as the most influential features.

Researchers have established a clear relationship between the genre and profitability of a movie, as well as a significant feature for model training (Wang and Zhang, 2018; Ghiassi et al., 2015; Mestyán et al., 2013). It was observed that action movies were more profitable than drama and comedy but once less than fantasy, horror and mystery once other attributes such as season, year and rating have been adjusted (Eklund and Kim, 2022). The frequently used analysis technique for genre is splitting them into separate variables and assigning binary or multiple values which would indicate if a movie fell under that category or not, however, this might turn out to be a one-dimensional approach that does not completely capture genre intersection effects.

The creativity and synergy of the cast and crew are pivotal in audience engagement and box office outcomes. However, converting them into usable features is challenging due to the limitations in quantifying artistic proficiencies. A few methods previous studies have featured engineered star values through their earnings, ranks or ratings of previous works, social media following and so on (Ning et al., 2020; Verma and Verma, 2020; Lash and Zhao, 2016). The role of directors in the commercial success of movies was also investigated similarly (Wei and Yang, 2022). Like the problem with the genre, this approach fails to capture the successful collaborative works and

partnerships between actors and directors. On the other hand, the size of the datasets used for these studies was limited, raising the question regarding bias and overfitting.

Considering data about the movie's content and context, such as script, visuals of posters and trailers are also additional features that can be incorporated into the model. The study conducted by Nemzer and Neymotin, (2020), using the descriptor words of a movie, they were able to predict movie success in terms of ticket sales, user rating and critic scores. Another study conducted by (Rahmani et al., 2019) extracted features based on the YouTube movie trailers and predicted average movie ratings with RMSE of 0.15. However, considering the size of the dataset, it could end up being very computationally heavy for implementation.

The movie industry is constantly evolving due to technological advancements, audience preferences and shifts in artistic expression, therefore, development of the analytical methodologies to assimilate these dynamic changes is critical.

### **2.3. Application of Machine Learning**

The success of a movie is measured on different levels, the popularity and hype it creates, the revenue it generates, the awards and recognitions it gets and so on. By using machine learning models to predict a movie's performance, the investors and other stakeholders will be able to transition towards data-driven decision-making (Cizmeci and Oguducu, 2018). Over the last several years, researchers have explored the scope of different algorithms in this area, such as regression (Ryu, 2020; Rahmani et al., 2019; Khopkar and Nikolaev, 2017), classification (Jagdale and M., 2019; Verma and Verma, 2020), ensemble models (Antipov and Pokryshevskaya, 2017; Subramaniaswamy et al., 2017; Wang et al., 2016), neural networks (Rehman et al., 2019; Ghiassi et al., 2015; Tarvainen et al., 2014) and so on.

The study conducted by Sahu and Ahuja, (2016) used IMDB user reviews to predict user sentiment, which was considered a reflection of performance. Through linguistics methods combined with feature extraction, the Random Forest model records an accuracy of 88.95%. A major challenge when it comes to opinion mining is the noise involved with using a real-world dataset. However, through Exponential-Salp Swarm Algorithm (ESSA), Jagdale and M. (2019) were able to develop a classifier that can distinguish positive and negative comments with more than 81% accuracy, which is suitable for real-world application. Even though it is essential to understand customer responses, as Rahmani et al. (2019) point out, treating movie performance as a classification problem causes significant data loss. (Mhowwala et al., 2020) also share the same view and elaborate that such findings lack actionable insights and do not give specific information as one might expect.

Even though Regression is an alternative to this problem, studies show that the inability of linear regression models to accommodate the interdependence of the variables and their complex relationships, calls for regularized models or advanced models such as Gaussian or Bayesian (Rehman et al., 2019; Jiaxin Zhu et al., 2016; Lee and Chang, 2009).

The comparative study of Chamani et al., (2021) between the performance of linear and ensemble methods for early movie ratings shows better performance for ensemble methods. While Extreme Gradient Boosting, the best-performing model recorded an RMSE of 0.68, Ridge Regression had an RMSE of 0.69. Even though they were able to incorporate metadata from multiple sources which contributed to improved model performance, the studies do not provide insights into the relationship between variables. The study by Mhowwala et al., (2020) generated similar results as well, with an RMSE score of .42 for Extreme Gradient Boosting and 0.53 for Random Forest.

Through their work, Antipov and Pokryshevskaya, (2017) emphasise the potential of Random Forest to predict movie performance before release since it has wider tuning options than automated algorithms. Additionally, by integrating multiple unique variables such as star power, seasonality, and competition in the release week they were able to enhance the model. The Random Forest model showed a Symmetric Mean Absolute Percentage Error (SMAPE) of 57%. Through this methodology, they were able to bring transparency into the model by evaluating the significance of each variable. Support Vector Machine is another ensemble model that can capture complex relationships within the dataset. The study of (Dhir and Raj, 2018), which used metadata and extracted social media records accuracy of 0.46 for the Support Vector Machine,

## **2.4. Summary**

Earlier studies established a clear association between attributes such as cast, director, genre, and the commercial and social success of movies. However meta features such as title, plot summary and the interaction level between the features, especially teamwork between actors and directors are under-explored. Additionally, for actionable business insights, it is necessary to understand the relationship between movie performance and its defining variables. This study's intention was to contribute to this gap by investigating the Multivariate relationships and prediction of movie performance before release based on metadata including unstructured text data. The predictive results would be useful to producers and other investors for strategic and marketing planning.



### **3. Methodology**

Various customer-centric and content-based movie attributes assist in predicting a movie's performance. In addition to model performance, it is essential to have analytical insights about profitability and performance, before the release since it can help businesses with decision-making. The suitability of a supervised model for predicting movie ratings has already been established by past researchers (Wang and Zhang, 2018; Ghiassi et al., 2015). Considering these elements, this study attempts to predict the IMDB average movie rating through metadata, that are known prior to the commercial release of a movie. Additionally, through statistical methods and visualizations, bivariate and multivariate relationships were explored.

This study will use regression models to analyse the attributes and predict the movie rating akin to the researchers by Chamani et al. (2021) and Mhowwala et al. (2020) along with focus on model interpretability. In contrast to the previous studies, for this project, a bigger dataset was used and unstructured text data such as title names, cast and crew details, as well as movie descriptors were incorporated. This section discusses the methodology following in this research while explaining the data source, pre-preprocessing, exploration, and model performances.

#### **3.1. Method of Study**

The overall research process followed in this study is driven by Cross-Industry Standard Process for Data Mining (CRISP-DM). The cross-industry Standard Process for Data Mining (CRISP-DM) has six phases as illustrated in Figure 1. However, considering the nature of this study, the last stage, deployment, is not applicable. Instead, managerial insights and recommendations are discussed.

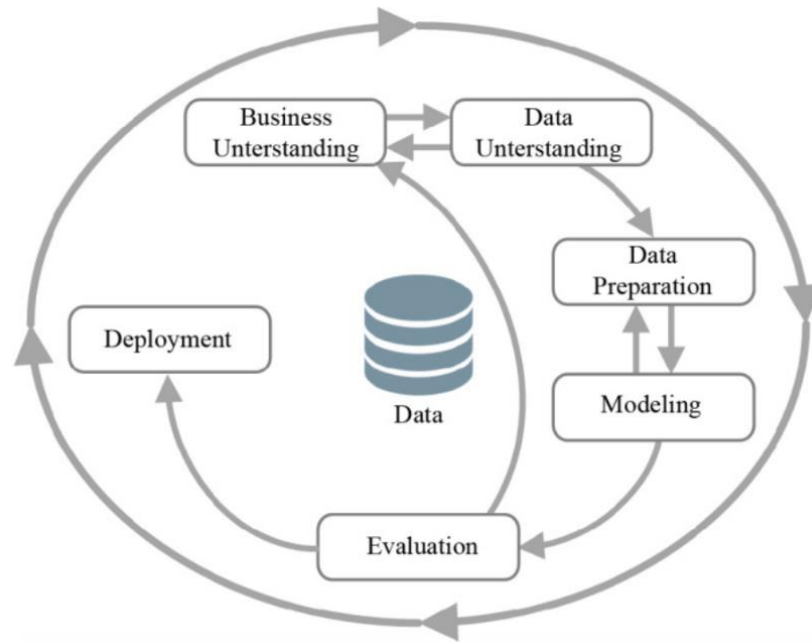


Figure 1. The CRISP-DM Process

Source: Huber et al., (2019), p. 404)

### 3.1.1. Business Understanding

The first step of by Cross-Industry Standard Process for Data Mining (CRISP-DM) framework is Business Understanding. This involves defining the objectives from the perspective of the business and understanding how they can be achieved through data analytics (Wirth and Hipp, 2000).

Considering the impact, machine learning can make on strategic planning in the movie industry, the key questions this research aims to answer were the various factors that impact a movie's rating and how it can be used to understand customer preferences along with marketing strategies design. This report also attempts to compare the performance of various models in movie rating predictions. Furthermore, the research methodology ensured that bias and misrepresentation were not happening and maintains explainability.

### **3.1.2. Data Understanding**

This phase starts with data collection and familiarisation with it through initial data explorations (Wirth and Hipp, 2000). The dataset used for this study was the IMDB non-commercial dataset (IMDb, n.d.), which had over 600,000 records across different categories such as movies, series, videos et cetera. Considering IMDB's prominent standing among movie databases, using their openly accessible dataset enabled model development and feature exploration that mirrors real world scenarios.

At the IMDB data source the attributes were split into 7 different files. Firstly, a subset was created with only details of movies. It had over 120,000 records which spans from 1870s to 2020s. Considering the transformation of the movie industry and customer interest over the year, only the movies released after 2000s were considered. The pre-release features of the movies, that were split across the datasets were merged. Table 1 presents the different variables selected from the IMDB dataset.

Considering the importance of release dates, production companies, budget, and feature extraction, through TMDB API, additional pre-release features were extracted. A summary of these variables is presented in Table 2.

To summarize, the target variable of this study is the IMDB average user rating, and the dependant variables include the metadata of the movie such as title, genre, release date, runtime, release locations, runtime, budget, director, writer, actors, actress and overview.

Table 1. Description of variables in the IMDB Dataset

(Source: IMDB, n.d)

Feature Name	Description
tconst	Unique alphanumeric ID used by IMDB
primaryTitle	Popular title of the movie
isAdult	Adult rated movie or not
startYear	Year of release
runtimeMintues	Duration of the movie
Genres	Movie Category, separated by comma
averageRating	Target Variable. Weighted average rate for a movie from all users
numVotes	Number of votes for the title
Director	Name of directors, comma separated
Actor	Name of actors, comma separated
Actress	Name of actresses, comma separated
Region_list	Countries in which the primaryTitle were used, comma separated

Table 2. Description of Features extracted through TMDB API

Feature Name	Description
Release Date	Date of release, in dd-mm-yyyy format
Budget	Cost of production
Overview	Description of the movie plot

### **3.1.3. Data Preparation**

Data preparation is a significant step because it helps to transform the raw dataset into a suitable format for modelling (Wirth and Hipp, 2000). The various tasks carried out in this phase, such as data exploration, cleaning, feature transformation and partitioning, are explained below.

#### *Exploratory Data Analysis*

Through statistical measures such as central tendencies, dispersion, and correlation; as well as visualizations using box plots, histograms, line charts and heatmaps, the underlying patterns and anomalies within the datasets were identified. Furthermore, it laid the foundation for the upcoming steps.

#### *Data Cleaning*

The various tasks that are performed in conjunction with data cleaning were the identification and rectification of outliers, missing information, categorical errors, duplicates, and wrong information.

Using boxplots and the Interquartile range (IQR) method, outliers were identified and removed from runtime. There were missing information against rating, runtime, director, actor, actress and overview. Considering the negative effect misinformation can have on the model and the large size of the dataset, records with missing information were removed.

In the IMDB dataset, a notable concern was the sampling bias created in certain movies with limited votes and elevated ratings. Therefore, to avoid skewed and misleading results, only movies with more than 100 votes were used. Based on movie ids, duplicate records were removed. Since majority of the movies are releasing worldwide and contents being available on streaming platforms, region was removed since it did not add value. After the data cleaning processes, the

dataset had 28,699 records, with three numerical variables, which were runtime, day, month, and five categorical variables, which were title, overview, genre, original movie or not, director, actor and actress.

### *Feature Transformation and Extraction*

Feature extraction, selection and transformation are crucial to enhance a model's performance and its interpretability. Using Natural Language Processing (NLP), the sentiment score and word count of the movie titles and overviews were examined. Similarly, the overview was converted into a Term Frequency - Inverse Document Frequency (TF-IDF) vector which helped to reduce the dimension while keeping the lexical significance. Genre was processed using one hot encoding. The recent gender neutrality of the term 'actor' calls for the actor and actress to be merged into a single list. For computational efficiency, only the first director (in order) was selected for modelling on the assumption that they are the primary contributors. Similarly, two actors based on their order were selected for model training. This helped to reduce the high variance while keeping the essence of the variables. Based on the impact this transformation made on model performance, binning was used over hash encoding and label encoding to transform it further for model training.

### *Data Partitioning*

After the feature engineering processes, the dataset was split into train and test sets in the ratio 80:20. The dataset was subsequently scaled using the Min-Max scaler so that the variables measured in different levels could be on a comparable scale.

#### **3.1.4. Modelling**

Based on the literature review and the nature of the attributes, various supervised machine learning algorithms were used to train the dataset. Their performance evaluation enabled the evaluation of the most suited algorithm to predict the IMDB ratings based on meta-features. The various linear and ensemble models used for this study are Linear Regression, Ridge Regression, Random Forest, Support Vector Regression and Extreme Gradient Boosting (XGBoost).

The linear regression model was used as the baseline for comparison against other models. To ensure robustness and avoid overfitting, 10-fold Cross-Validation was used for all model training. Furthermore, Backward Elimination was used for feature selection in linear models, while Recursive Feature Elimination (RFE) was used for ensemble methods. The parametric nature of the linear models made Backward Elimination a better feature selection method for them. Alternatively, Recursive Feature Elimination (RFE) enables model optimization in Random Forest, Support Vector Regression, and Extreme Gradient Boosting due to their capabilities to handle high dimensional data. Additionally, these feature selection methods helped to understand the variable's significance and reduce dimensionality. For ensemble methods, using Cross-Validated Random Search, the best combination of parameters were identified.

#### **3.1.5. Evaluation**

The predictive capacity of the trained models at best settings were examined by testing it on the previously unseen test set. Its performance was measured and compared using various measures such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) R Squared, Adjusted R Squared, and Residual Analysis. Further evaluation of model interpretability and feature significance was carried out through tree based feature importance and Accumulated Local Effect (ALE).

## **4. Findings**

The focus of this study was to investigate the relationship between various meta-features, especially textual attributes, and the IMDB rating to understand customer preferences. Additionally, it also aims to evaluate the predictive power of various Supervised Learning Models to forecast the movie ratings, based on these attributes.

The analytical process had three major steps, data exploration, model training and model evaluation in terms of performance and interpretability. This section presents its findings and further discusses its consistency with the current domain knowledge and offers business insights to understand and cater to customer inclinations.

### **4.1. Understanding the relationship between the Variables**

In addition to base features such as genre, runtime, release date and category, textual features such as title, overview, and details of director as well as actors.

#### **Distribution of Average Ratings**

The target variable of this analysis was the IMDB rating for each movie. For statistical evaluation and modelling, normality in its distribution is essential for regression models (James et al., 2021) Figure 2 illustrates the density plot of the average rating. Additionally, the Shapiro-Wilk Test Statistics had a score of 0.98. Both findings suggest normality.

#### **Trends in Release Dates**

A steady growth could be observed in the number of movies released since 2000 (Appendix 1: Figure 1). Even though the global pandemic caused a degrowth in the industry since 2019, it has been slowly recovering in recent years. Interestingly, some of the highest-rated movies were released prior to 2010 as suggested by Figure 3. However, for model training, rather than the year,



it would be logical to choose the month and day of release since they are more controllable. Even though the correlation between these variables and the average ratings was less than 0.5, more than half of the movies were released on Thursdays and Fridays. Additionally, a higher number of movies were released in months closer to holidays, such as August, November, and December (Figure 4). These factors suggest the importance of days and seasonality in movie releases.

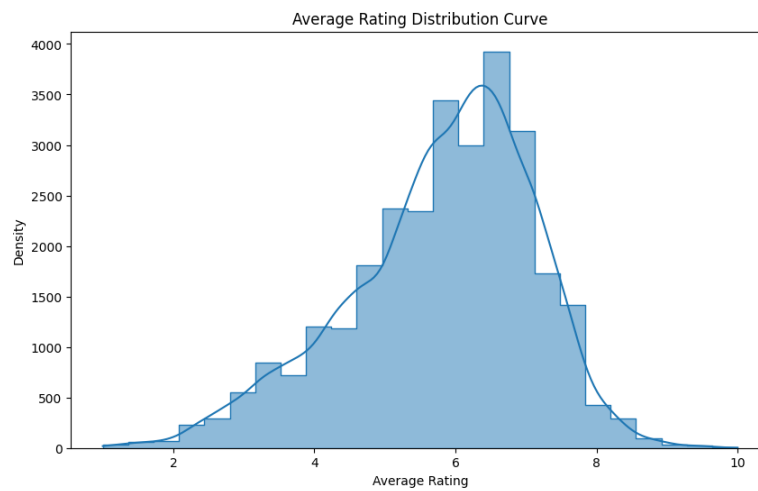


Figure 2. Distribution Curve of Average IMDB Rating

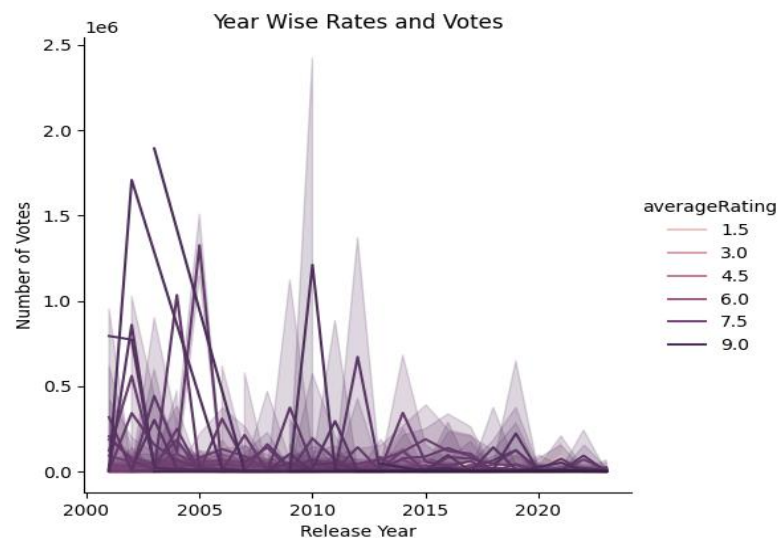


Figure 3. Year Wise Popularity for Movies

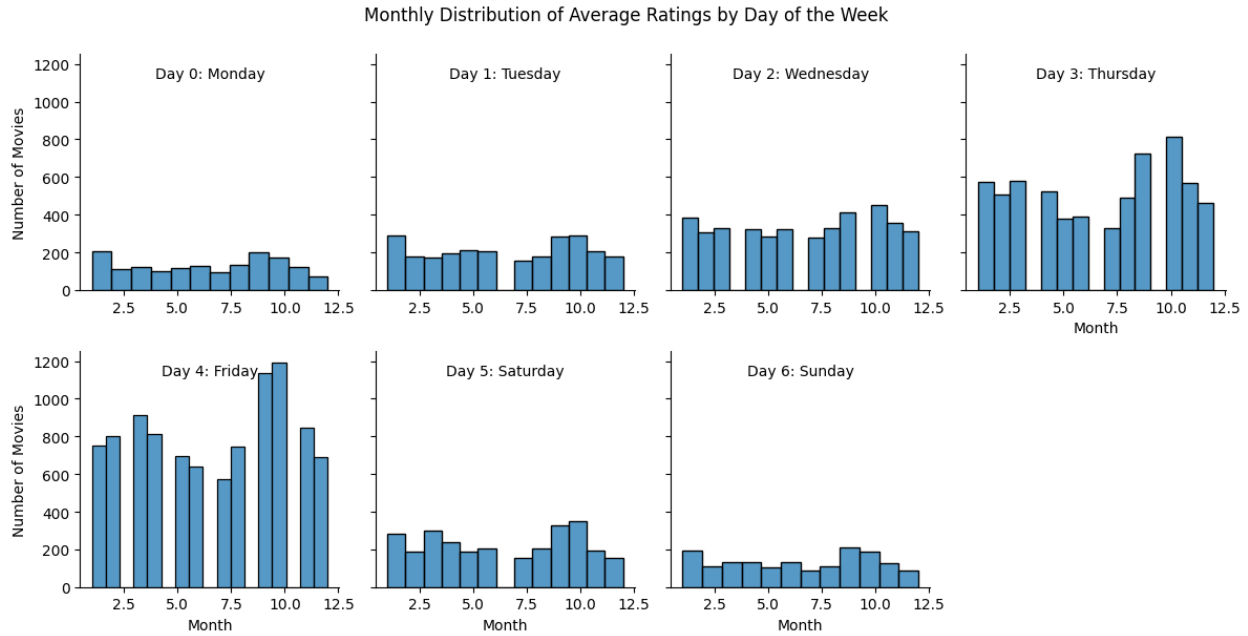


Figure 4. Month and Day Wise Split of Movie Releases

### Movie Duration

The duration of most of the movies was from 80 minutes to 100 minutes. Movies with longer runtimes were mostly from the genres of action, drama, and romance. However, a linear relationship between them could not be identified between runtime and movie ratings. Moreover, there is high variability for duration among top-rated movies (Figure 5).

### Genre: Combinations and Ratings

Approximately 50% of the movies in the dataset fall under the genre of drama. Comedy, thriller, and romance are the next most popular genres. Interestingly, most of the top-rated movies fall under these genres as well. It could be noted that many movies show multiple genre classifications suggesting the practice of genre combinations in the movie industry.

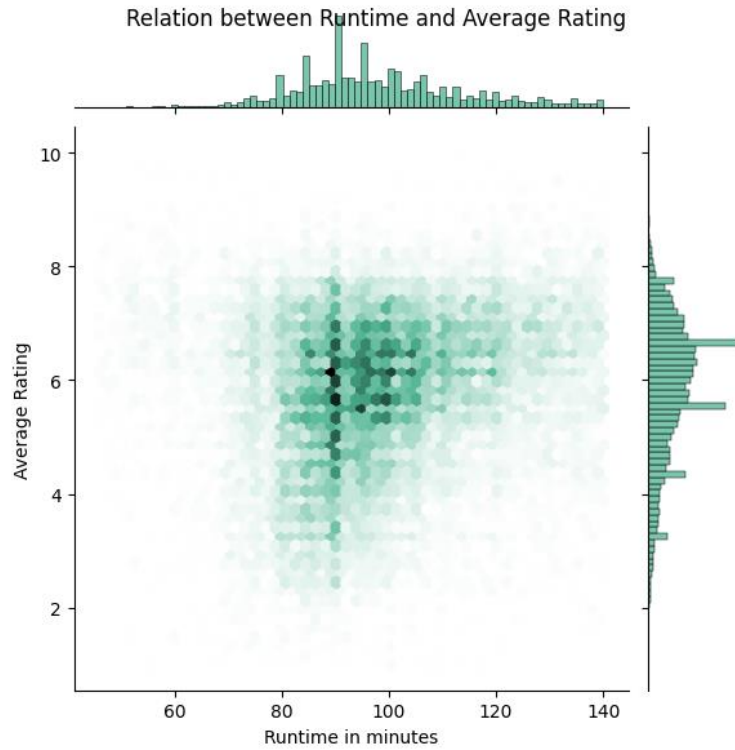


Figure 5. Bivariate Relation Between Rating and Duration

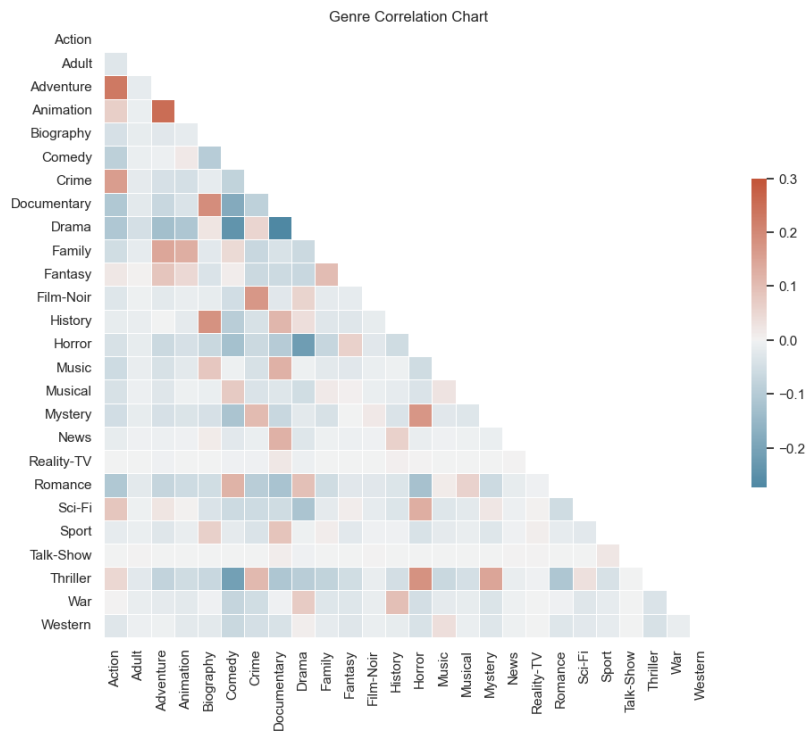


Figure 6. Interaction Rates Between Genres

Figure 6 shows the correlation between different genres. Some of the prevalent movie combinations were adventure and animation, mystery, science fiction and thriller, along with animation, family, and fantasy. However, their correlation coefficient with the average movie rating was well below 0.5, suggesting a weak linear relationship between them. But more than lack of relation, it implies a complex nature, which might not be captured in linear models.

### **Director and Cast**

Even though the director and cast are integral to the success of a movie. The main challenge pointed out by previous researchers in incorporating these features is its high variability (Wei and Yang, 2022; Lash and Zhao, 2016). Binning them based on the number of movies each person worked with helped to avoid bias and overfitting, which in turn improved model performance.

For the majority of the directors and actors, their number of movies was less than five. However, it was discovered that certain actors and directors were consistently associated with high-rated movies, which highlights their ability to attract the audience and deliver results. This emphasises the need to consider them for predictive modelling. However, there were multiple instances where a novice or moderately experienced director and actor were able to create high-rated movies, making it difficult to establish a clear-cut connection between the two.

### **Text Data: Title and Overview**

A typical movie title had two to three words with the most used word being love, day, life, man and so on. Figure 7 represents the most frequently used words for movie titles. The sentiment analysis shows that, majority of the movies are neutral. This highlights the significance of precision and brevity when it comes to movie titles. The overview vector of the movies showed that some of the highest rated words are ‘back’, ‘city’, ‘daughter’, ‘family’, ‘friends’, ‘man’, ‘together’ and so on. These give a quick glimpse into the common themes that occur in movies,

which were also a reflection of customer interests. These words further imply an interest in the customers for drama, mystery and action movies focused on a man or women, which depicts their relations with friends or family. It would not be wrong to infer that customers consistently show signs of interest towards specific themes.

Figure 7. World Cloud Based on Movie Title

Statistical measures and data visualization helped to discover the patterns within the attributes. Even though the diagrams and measures such as the correlation coefficient (Appendix 1: Figure 2) failed to show a significant linear relationship between these attributes and movie ratings, their importance in a movie's success were established, making it essential to explore non-linear and ensemble models.

## **4.2. Machine Learning Models**

### **Model 1: Multiple Linear Regression**

As a statistical technique, Multiple Linear Regression enables the evaluation of the collected effect of numerous variables on a dependent variable (James et al., 2021). This helps to generate insights about multi-variate relationships. This can be mathematically expressed as:

Incorporating Term Frequency - Inverse Document Frequency (TF-IDF) vectors derived from the movie overview and additional features extracted from the title, such as word count and sensitivity, in addition to attributes such as genre, duration, and release date, yielded better results over only using base variables. This model had an RMSE of 1.13 and R squared of 0.31. After that, through backward feature elimination, the model was trained with 11 selected features which showed a slightly better performance with an RMSE of 1.12 and R squared of 0.32. In both cases, the performance was improved after 10-fold cross-validation.

The validity of the linear regression model was evaluated through assumption checking. It was seen that the data followed a normal distribution and the residuals suggested linearity among the variables. There was no sign of autocorrelation. However, there seems to be heteroscedasticity in the dataset and multicollinearity between day and month.

### **Model 2: Ridge Regression**

Ridge Regression is a regularization method that addresses the inconsistencies of Linear Regression by adding a penalty term to the regression equation, which in turn constrains the coefficients (Hastie et al., 2009). It can reduce the impact of problems such as multicollinearity, heteroscedasticity, et cetera, and therefore increase the model stability.

The model results showed slightly better performance as compared to linear regression with an RMSE of 1.1 and R squared of 0.3. The backward elimination feature selection reduced the dataset dimension from 165 variables to 56 along with improving the model performance. The R Squared post feature selection was 0.33 and RMSE dropped by another 0.1 units. However, cross-validation did not make significant improvements to the models.

### **Model 3: Random Forest**

Random Forest is an ensemble machine learning algorithm that combines the predictive results of numerous decision trees (James et al., 2021). By training on numerous trees with random subsets of data and aggregating the results, this algorithm produces accurate and robust results (Hastie et al., 2009). The high-dimensional and complex nature of this dataset made Random Forest a suitable technique.

Leveraging the unstructured text data through NLP techniques in addition to base variables proved to be fruitful with Random Forest model generating significantly better results with an RMSE of 0.85 and R Squared 0.37. Even though Cross-validation methods were used, it did not make much improvement implying lack of overfitting. Random Search Hyperparameter tuning with over 200 iterations showed that the optimal parameters had 650 estimators with a maximum depth of 10. This further improved the RMSE score to 0.84

### **Model 4: Extreme Gradient Boosting**

Similar to Random Forest, Extreme Gradient Boosting sequentially trains an ensemble of decision trees, with each tree capable of correcting its previous error (James et al., 2021). The importance of interpretability and predictive performance, along with the high dimensional dataset in this study makes Extreme Gradient Boosting an appropriate selection. It was able to generate satisfying

results with RMSE 0.86 and R Squared of 0.37. However, feature selection and hyperparameter tuning did not significantly improve its performance indicating the possibility to explore advanced text analytics like semantic text analysis, entity recognition and so on.

### Model 5: Support Vector Regression

Support Vector Regression operates by locating a hyperplane that best fits the data and reducing the prediction errors within a specified margin (James et al., 2021). Studies show it is suitable for instances where is a complex, non-linear relationship between the variables (Subramaniaswamy et al., 2017). Initial models which only used base variables recorded an RMSE of 0.88. Incorporation of textual vector improved the performance to 0.87. However, feature selection only made minimal improvements to the model performances.

### Comparison of Model Performances

The overall performances of the various linear and ensemble models used in this study are illustrated in Table 3. The best performance was recorded by Random Forest which included both text data and base variables such as genre and duration after Feature Selection.

Table 3. Performance Comparison of Linear and Ensemble Models

Model Name	MAE	MSE	RMSE	R2	CV
Linear Regression	0.87	1.28	1.12	0.32	1.11
Ridge Regression	0.86	1.27	0.98	0.33	0.97
<b>Random Forest</b>	<b>0.85</b>	<b>1.26</b>	<b>0.84</b>	<b>0.37</b>	<b>0.84</b>
Extreme Gradient Boosting	0.85	1.26	0.86	0.37	0.86
Support Vector Regression	0.86	1.27	0.87	0.34	0.86



## Interpretability and Feature Importance

The significance of the various features was evaluated by comparing their influence on the Linear Regression model and Random Forest. These helped to bring transparency to the models and discover important features.

In the Linear Regression Model, runtime was the most important feature. Among genres, documentary and animation were the most significant features. The model further proves the importance of actors in the success of a movie. Alternatively, the tree based feature importance scores in Random Forest shows horror as the most prominent genre. Apart from the word length of the title, actors were also found as significant features. The comparative scores of the top features based on both the models are recorded in Table 4.

Table 4. Feature Importance Scores

Feature	Linear Regression Score	Random Forest Score
Runtime	1.68081	0.21278
Sentiment of the title	0.07258	0.02795
Documentary	1.48876	0.08794
Horror	-	0.13259
Actress	0.19780	0.05578
Actor	0.10269	0.05764

### 4.3. Discussion of Theoretical Impact

This study was successful in contributing to the current domain of predictive analytics for movie performance. The key results about the different attributes and model training not only aligns with the current knowledge, but also adds to it.

One of the key findings of this study was the superiority of ensemble models over linear models to capture complex, high-dimensional, interaction effects, which aligns with the findings of Chamani et al., (2021). The primary factor that contributed to the consistent improvement of model performances was feature extraction and training the model, with a combination of metadata and textual data. Similar to the findings of Lee et al. (2018), in addition to performance enhancement, adequate feature extraction also enabled interpretability and transparency for the models. Additionally, the wider tuning capabilities of Random Forest, which was highlighted by Antipov and Pokryshevskaya (2017), helped to get the best results in this study with an RMSE of 0.84.

Overfitting is a critical challenge that could be observed across the data analytics domain. Mhowwala et al. (2020), points out that smaller datasets are more prone overfitting. The average sample size used by numerous past researches were less than 5000 since they were limited to a specific region (Xiao et al., 2021; Abidi et al., 2020; Lee and Chang, 2009). However, by considering the global audience for movies that transcend geographical borders, it was possible to create a much larger dataset. Training the model with a larger dataset helped to reduce overfitting, which is implied by the lack of improvement in model performance with Cross-validation.

Even though there were differences across studies, regarding the specific genre, that might have higher influence on the movie's performance, its overall significance is unarguable. The common method for used for their transformation is one hot encoding. However, this could create sparsity,

which in turn causes overfitting (Cizmeci and Oguducu, 2018). Therefore, it is possible that the whole relationship between multiple genres might not have been accurately captured.

While the results suggest actors are important attributes similar to the results of Lash and Zhao (2016), a significant relationship could not be established between directors and movie ratings. This contrasts with the findings of (Wei and Yang, 2022). This calls for further feature engineering and inclusion of variables related to their original number of works and other skills that might reveal further insights about them.

#### **4.4. Implications for Business**

The predictive modeling techniques developed in this study along with the insights can be used by different stakeholders in the movie industry for risk management and project planning. Considering the significance of duration of the movie, movies can be planned to be within the preferred limits of 80 to 120 minutes. While action, drama, and romantic movies are more frequent and popular, constantly stable performances were observed for movies in the genres of animation and horror. Based on the risk appetite of the production house, they can choose to have a more experienced crew and cast which in turn would increase the expenses.

In addition to predicting movie prediction, NLP are powerful techniques that can be used by the stakeholders to understand customer experience, opening a two-way communication between the parties.

## 5. Conclusion

This study investigated the pre-release movie prediction using linear and ensemble machine learning models and text analytics. By combining the pre-release meta-features and unstructured textual data, and appropriate feature selection as well as model tuning, the Random Forest model generated an RMSE score of 0.84. The study was successful in using a bigger dataset unlike previous works and additionally reduce overfitting through feature selection and dimension reduction. The models revealed that duration, actors, and genre, especially horror, animation, and documentary, were significant attributes that can influence model performance. The statistical explorations further showed movies with themes such as family, friends, love et cetera were widely accepted by the audiences. Planning the theatrical release around weekends and holidays would help investors to earn more revenue.

However, there is still room for significant improvements. Especially in model performance. Using advanced methods like neural networks might improve the model's performance due to their ability to capture complex relations. Considering the limited time and computational capacity at hand to complete this study, data loss was compromised during data cleaning. Furthermore, this study does not include any of the variables that are directly linked to revenue and profitability such as ticket collection, marketing, and broadcast deals and so on. These factors would have helped to quantify movie success in terms of profitability.

Considering the significant transformation, analytics can make of the movie industry, in future, it would be interesting to combining the text data with deep learning generative models like GANs to suggest movie titles or even tag words,

## References

- Abidi, S.M.R., Xu, Y., Ni, J., Wang, X., Zhang, W., 2020. Popularity prediction of movies: from statistical modeling to machine learning techniques. *Multimed. Tools Appl.* 79, 35583–35617. <https://doi.org/10.1007/s11042-019-08546-5>
- Ahmad, J., Duraisamy, P., Yousef, A., Buckles, B., 2023. Movie Success Prediction Using Data Mining.
- Antipov, E.A., Pokryshevskaya, E.B., 2017. Are box office revenues equally unpredictable for all movies? Evidence from a Random forest-based model. *J. Revenue Pricing Manag.* 16, 295–307. <https://doi.org/10.1057/s41272-016-0072-y>
- Chamani, H., Zadeh, Z.S.H., Bahrak, B., 2021. An Overview of Regression Methods in Early Prediction of Movie Ratings, in: 2021 11th International Conference on Computer Engineering and Knowledge (ICCCKE). Presented at the 2021 11th International Conference on Computer Engineering and Knowledge (ICCCKE), IEEE, Mashhad, Iran, Islamic Republic of, pp. 1–6. <https://doi.org/10.1109/ICCCKE54056.2021.9721453>
- Cizmecic, B., Oguducu, S.G., 2018. Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media, in: 2018 3rd International Conference on Computer Science and Engineering (UBMK). Presented at the 2018 3rd International Conference on Computer Science and Engineering (UBMK), IEEE, Sarajevo, pp. 173–178. <https://doi.org/10.1109/UBMK.2018.8566661>
- Clark, T., 2022. The biggest box-office flops of the year show how the movie business still has a long road to recovery [WWW Document]. *Bus. Insid.* URL <https://www.businessinsider.com/biggest-box-office-flops-of-the-year-2022-5> (accessed 9.7.23).

Cui, G., Lui, H.-K., Guo, X., 2012. The Effect of Online Consumer Reviews on New Product Sales.

Dhir, R., Raj, A., 2018. Movie Success Prediction using Machine Learning Algorithms and their Comparison, in: 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC). Presented at the 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), IEEE, Jalandhar, India, pp. 385–390. <https://doi.org/10.1109/ICSCCC.2018.8703320>

Eklund, J., Kim, J.-M., 2022. Examining Factors That Affect Movie Gross Using Gaussian Copula Marginal Regression. *Forecasting* 4, 685–698. <https://doi.org/10.3390/forecast4030037>

Frater, P., 2023. Global Box Office Notched 27% Gain in 2022 to Hit \$26 Billion Total, Research Shows. *Variety*. URL <https://variety.com/2023/data/news/global-box-office-in-2022-1235480594/> (accessed 9.7.23).

Ghiassi, M., Lio, D., Moon, B., 2015. Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Syst. Appl.* 42, 3176–3193. <https://doi.org/10.1016/j.eswa.2014.11.022>

Hastie, T., Tibshirani, R., Friedman, J., 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. ed, *Elements of Statistical Learning*, Springer Series in Statistics. Springer, New York.

Ifada, N., Rahman, T.F., Sophan, M.K., 2020. Comparing Collaborative Filtering and Hybrid based Approaches for Movie Recommendation, in: 2020 6th Information Technology International Seminar (ITIS). Presented at the 2020 6th Information Technology International

Seminar (ITIS), IEEE, Surabaya, Indonesia, pp. 219–223.  
<https://doi.org/10.1109/ITIS50118.2020.9321014>

IMDb, n.d. IMDb Non-Commercial Datasets [WWW Document]. IMDb Dev. URL  
<https://developer.imdb.com/non-commercial-datasets/> (accessed 7.7.23).

Jagdale, J., M., E., 2019. Optimization driven actor-critic neural network for sentiment analysis in social media. *VINE J. Inf. Knowl. Manag. Syst.* 49, 457–476. <https://doi.org/10.1108/VJIKMS-12-2018-0116>

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An Introduction to Statistical Learning: With Applications in R, 2nd ed. 2021. ed, An Introduction to Statistical Learning, Springer Texts in Statistics. Springer, New York, NY.

Jha, G.K., Gaur, M., Thakur, H.K., 2022. A trust-worthy approach to recommend movies for communities. *Multimed. Tools Appl.* 81, 19655–19682. <https://doi.org/10.1007/s11042-021-11544-1>

Jiaxin Zhu, Yijun Guo, Jianjun Hao, Jianfeng Li, Duo Chen, 2016. Gaussian Mixture Model based prediction method of movie rating, in: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). Presented at the 2016 2nd IEEE International Conference on Computer and Communications (ICCC), IEEE, Chengdu, China, pp. 2114–2118.  
<https://doi.org/10.1109/CompComm.2016.7925073>

Khopkar, S.S., Nikolaev, A.G., 2017. Predicting long-term product ratings based on few early ratings and user base analysis. *Electron. Commer. Res. Appl.* 21, 38–49.  
<https://doi.org/10.1016/j.elerap.2016.12.002>

- Kim, J.-M., Xia, L., Kim, I., Lee, S., Lee, K.-H., 2020. Finding Nemo: Predicting Movie Performances by Machine Learning Methods. *J. Risk Financ. Manag.* 13, 93. <https://doi.org/10.3390/jrfm13050093>
- Lash, M.T., Zhao, K., 2016. Early Predictions of Movie Success: The Who, What, and When of Profitability. *J. Manag. Inf. Syst.* 33, 874–903. <https://doi.org/10.1080/07421222.2016.1243969>
- Lee, K., Park, J., Kim, I., Choi, Y., 2018. Predicting movie success with machine learning techniques: ways to improve accuracy. *Inf. Syst. Front.* 20, 577–588. <https://doi.org/10.1007/s10796-016-9689-z>
- Lee, K.J., Chang, W., 2009. Bayesian belief network for box-office performance: A case study on Korean movies. *Expert Syst. Appl.* 36, 280–291. <https://doi.org/10.1016/j.eswa.2007.09.042>
- Mestyán, M., Yasseri, T., Kertész, J., 2013. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLoS ONE* 8, e71226. <https://doi.org/10.1371/journal.pone.0071226>
- Mhowwala, Z., Razia, A., D., S., 2020. Movie Rating Prediction using Ensemble Learning Algorithms. *Int. J. Adv. Comput. Sci. Appl.* 11. <https://doi.org/10.14569/IJACSA.2020.0110849>
- Nemzer, L.R., Neymotin, F., 2020. How words matter: machine learning & movie success. *Appl. Econ. Lett.* 27, 1272–1276. <https://doi.org/10.1080/13504851.2019.1676868>
- Ning, X., Yac, L., Wang, X., Benatallah, B., Dong, M., Zhang, S., 2020. Rating prediction via generative convolutional neural networks based regression. *Pattern Recognit. Lett.* 132, 12–20. <https://doi.org/10.1016/j.patrec.2018.07.028>



Rahmani, H.A., Deldjoo, Y., Schedl, M., 2019. A Regression Approach to Movie Rating Prediction using Multimedia Content and Metadata.

Rehman, A.U., Malik, A.K., Raza, B., Ali, W., 2019. A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimed. Tools Appl.* 78, 26597–26613. <https://doi.org/10.1007/s11042-019-07788-7>

Ryu, S., 2020. How does film adaptation influence box office performance? An empirical analysis of science fiction films in Hollywood. *Arts Mark.* 10, 125–143. <https://doi.org/10.1108/AAM-05-2019-0018>

Sahu, T.P., Ahuja, S., 2016. Sentiment analysis of movie reviews: A study on feature selection & classification algorithms, in: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom). Presented at the 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, Durgapur, India, pp. 1–6. <https://doi.org/10.1109/MicroCom.2016.7522583>

Subramaniaswamy, V., Vaibhav, M.V., Prasad, R.V., Logesh, R., 2017. Predicting movie box office success using multiple regression and SVM, in: 2017 International Conference on Intelligent Sustainable Systems (ICISS). Presented at the 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, pp. 182–186. <https://doi.org/10.1109/ISS1.2017.8389394>

Sunilkumar, C.N., 2020. A review of movie recommendation system: Limitations, Survey and Challenges. *ELCVIA Electron. Lett. Comput. Vis. Image Anal.* 19, 18. <https://doi.org/10.5565/rev/elcvia.1232>

- Tarvainen, J., Sjöberg, M., Westman, S., Laaksonen, J., Oittinen, P., 2014. Content-Based Prediction of Movie Style, Aesthetics, and Affect: Data Set and Baseline Experiments. *IEEE Trans. Multimed.* 16, 2085–2098. <https://doi.org/10.1109/TMM.2014.2357688>
- Thakker, U., Patel, R., Shah, M., 2021. A comprehensive analysis on movie recommendation system employing collaborative filtering. *Multimed. Tools Appl.* 80, 28647–28672. <https://doi.org/10.1007/s11042-021-10965-2>
- Van Es, K., 2023. Netflix & Big Data: The Strategic Ambivalence of an Entertainment Company. *Telev. New Media* 24, 656–672. <https://doi.org/10.1177/15274764221125745>
- Verma, H., Verma, G., 2020. Prediction Model for Bollywood Movie Success: A Comparative Analysis of Performance of Supervised Machine Learning Algorithms. *Rev. Socionetwork Strateg.* 14, 1–17. <https://doi.org/10.1007/s12626-019-00040-6>
- Wang, H., Zhang, H., 2018. Movie genre preference prediction using machine learning for customer-based information, in: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). Presented at the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), IEEE, Las Vegas, NV, pp. 110–116. <https://doi.org/10.1109/CCWC.2018.8301647>
- Wang, X., Luo, F., Qian, Y., Ranzi, G., 2016. A Personalized Electronic Movie Recommendation System Based on Support Vector Machine and Improved Particle Swarm Optimization. *PLOS ONE* 11, e0165868. <https://doi.org/10.1371/journal.pone.0165868>
- Wei, L., Yang, Y., 2022. An empirical investigation of director selection in movie preproduction: A two-sided matching approach. *Int. J. Res. Mark.* 39, 888–906. <https://doi.org/10.1016/j.ijresmar.2021.11.001>

Wirth, R., Hipp, J., 2000. CRISP-DM: Towards a Standard Process Model for Data Mining.

Won, D.U., Kim, H.S., 2022. A Prediction Scheme For Movie Preference Rating Based on DeepFM Model, in: 2022 International Conference on Information Networking (ICOIN). Presented at the 2022 International Conference on Information Networking (ICOIN), IEEE, Jeju-si, Korea, Republic of, pp. 385–390. <https://doi.org/10.1109/ICOIN53446.2022.9687136>

Xiao, X., Cheng, Y., Kim, J.-M., 2021. Movie Title Keywords: A Text Mining and Exploratory Factor Analysis of Popular Movies in the United States and China. *J. Risk Financ. Manag.* 14, 68. <https://doi.org/10.3390/jrfm14020068>

## Appendix 1: Additional Information

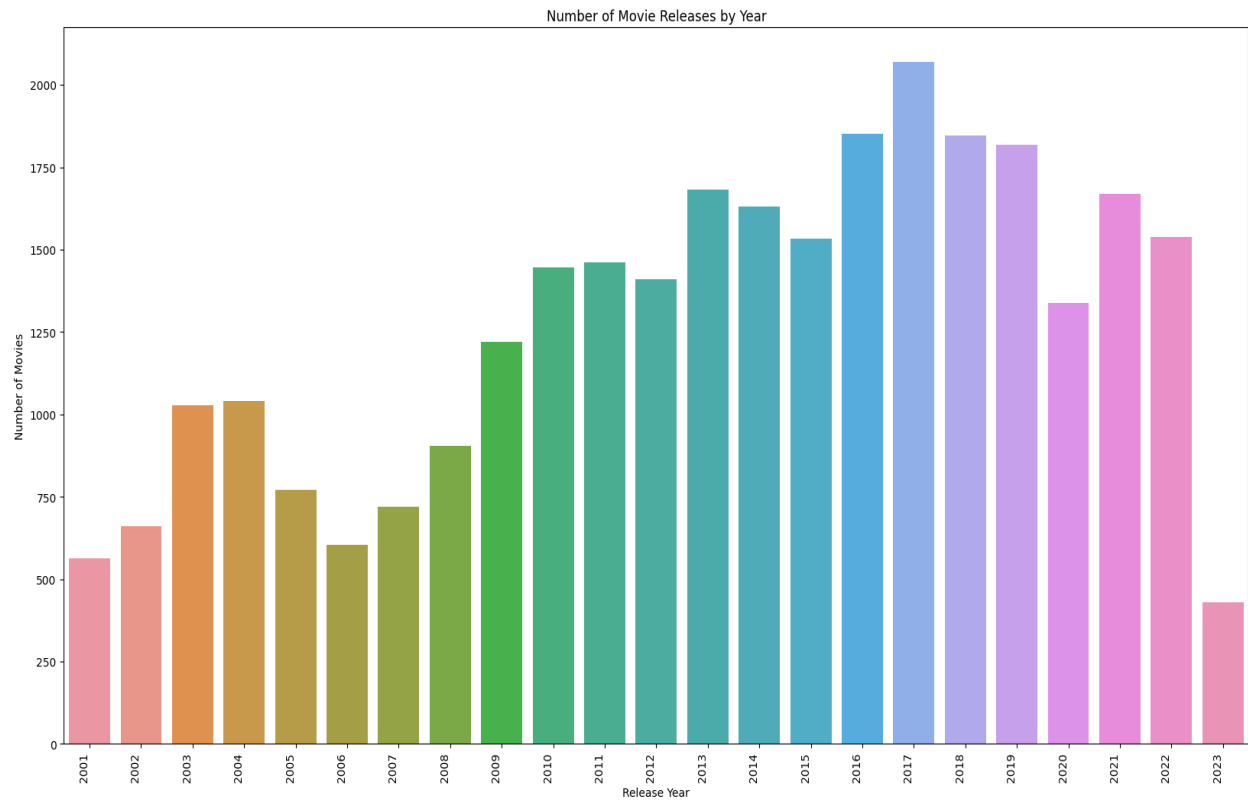


Figure 1. Number of movie releases throughout the years

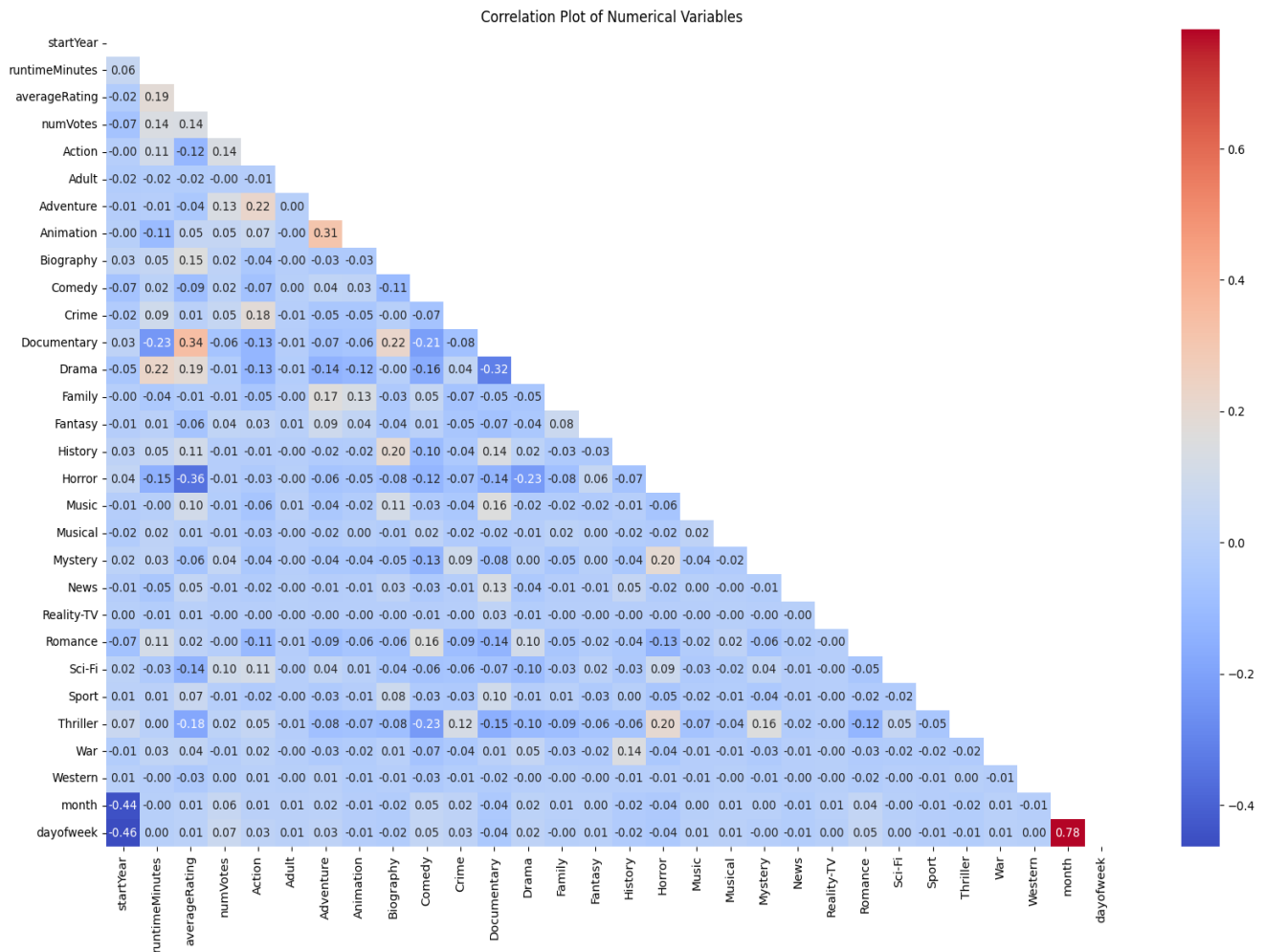



Figure 2. Correlation Plot of Variables

## Appendix 2: Dissertation Checklist Sheet

**Name:** Annie Jose \_\_\_\_\_

**Date Submitted:** 08 August 2023 \_\_\_\_\_

**Signature (Digital):**  \_\_\_\_\_

I confirm that my dissertation contains the following prescribed elements:

- ☐ My dissertation portfolio meets the style requirements set out in the MSc Business Analytics Portfolio Dissertation Handbook including a word count on the front page of each element.
- ☐ I have reviewed the Turnitin similarity report prior to submission.
- ☐ My dissertation title captures succinctly the focus of my dissertation
- ☐ My title page is formatted as prescribed in the MSc Business Analytics Portfolio Dissertation Handbook
- ☐ The abstract provides a clear and succinct overview of my study
- ☐ Each element contains a Table of Contents, and List of Figures and Tables (where appropriate)
- ☐ My dissertation contains a statement of acknowledgement (optional)
- ☐ The Introduction section of the research report, at a minimum, covers each of the following issues:
  - Background to/context of the project
  - Research question(s), aim(s) and objectives
  - Why the project is necessary/important
  - A summary of the Methodology
  - Outline of the key findings
  - Overview of chapter structure of the remainder of dissertation

- The Background section of the research report, at a minimum, covers each of the following issues:
  - Synthesises the key technical literature relating to the topic
  - Synthesises the key theoretical literature relating to the topic
  
- The methodology section of the research report:
  - Contains justification for the tools and method(s) selected
  - Details the procedures adopted (e.g. the data source/acquisition, data processing, procedures for maximising rigour and robustness, methods of data analysis etc) -
  - Contains ethical considerations and decisions
  
- The findings section of the research report reports the results in detail and provides possible explanations for the various findings
  
- The discussion section of the research report makes appropriate linkages between the findings and the literature reviewed
  
- The conclusions section of the research report includes:
  - Conclusions about each research question and/or hypothesis
  - General conclusions about the research problem
  - Implications for theory, for policy and/or management practice
  - Limitations of the research
  - Suggestions for practice and future research
  
- The technical report, log book, and reflective discussion have each been included.
  
- The reference list is in alphabetical order and follows the Harvard system
  
- I have signed and dated the Candidate Declaration

