UNIVERSITY OF INFORMATION TECHNOLOGY - VNUHCM

FACULTY OF COMPUTER SCIENCE

-----□□□□□-----

REPORT PROJECT DATA MINING AND APPLICATION

Topic: Forecasting Cinema Box Office Hits:

A Data-Driven Approach

Lecturer: PhD. Vo Nguyen Le Duy

Class: CS313.O21.KHCL

Students:

Phan Huy Hoang – 21520242

Pham Thi Tram Anh – 21520146

Duong Van Nhat Long - 20521561

Truong Quang Thien - 20520310

Le Dinh Duc - 19521372

Report date: 16-May-2024

Table of Content

I. Introduction

In the fiercely competitive film industry, cinema owners face the daunting task of selecting which movies to invest in and showcase. Predicting audience preferences and market trends is crucial for maximizing revenue, but the sheer volume of films released each year makes this decision-making process inherently challenging.

Data mining offers a powerful solution. By analyzing historical movie data, we can uncover hidden patterns and insights that inform strategic film selection. This project aims to leverage data mining techniques to build a predictive model that empowers cinema owners to make informed choices, ultimately boosting their bottom line.

The data mining process encompasses five key steps:

- Business Understanding: Clearly define the project goals and the specific challenges cinema owners face.
- Data Understanding: Gather and explore relevant movie data to identify potential patterns and relationships.
- Modeling: Develop a predictive model using statistical techniques like linear regression.
- Evaluation: Rigorously assess the model's accuracy and reliability.
- Deployment: Translate model insights into actionable recommendations for cinema owners.

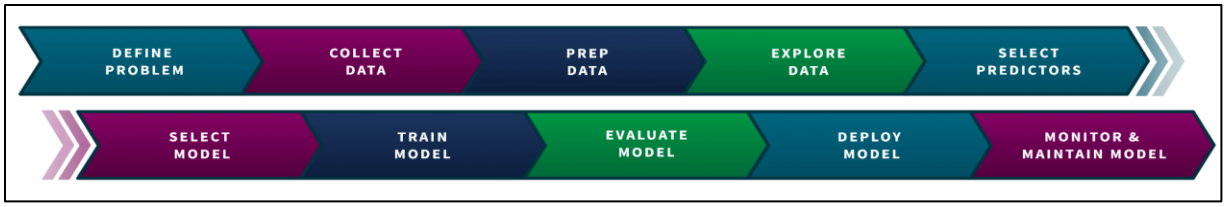

Figure 1: Data Mining Project Process

II. Problem Setup

This project addresses the specific challenge of film selection faced by cinema owners. The goal is to develop a predictive model that can

accurately assess the potential success of a movie before its release. By leveraging various data mining techniques, we aim to uncover hidden patterns and relationships within a large movie dataset, ultimately providing the cinema owner with a decision-support tool that can streamline their selection process.

III. Method

This section describes the methods which were used to analyze, mining the data, get insight and build a model to give users a suggestion about a movie.

1. Linear Regression

Definition: Linear regression is a statistical modeling technique used to investigate the relationship between a dependent variable (the outcome we want to predict or understand) and one or more independent variables (factors that might influence the outcome).

In the context of movie data, the dependent variable could be a movie's rating, box office gross, or audience score. Independent variables could include budget, genre, runtime, or the presence of specific actors or directors.

Algorithm: Linear regression uses a mathematical approach known as ordinary least squares (OLS). OLS determines the best-fitting line (or hyperplane in cases with multiple independent variables) that minimizes the sum of the squared differences between the predicted values and the actual values of the dependent variable. The resulting equation represents the linear relationship, and its coefficients quantify the strength and direction of each variable's effect.

Simple Linear Regression equation: $\boldsymbol{y = \theta_0 + \theta_1 X}$

Where:

- *θ0: intercept*
- *θ1: coefficient of x*
- *X are the independent variables*
- *Y is the dependent variable*

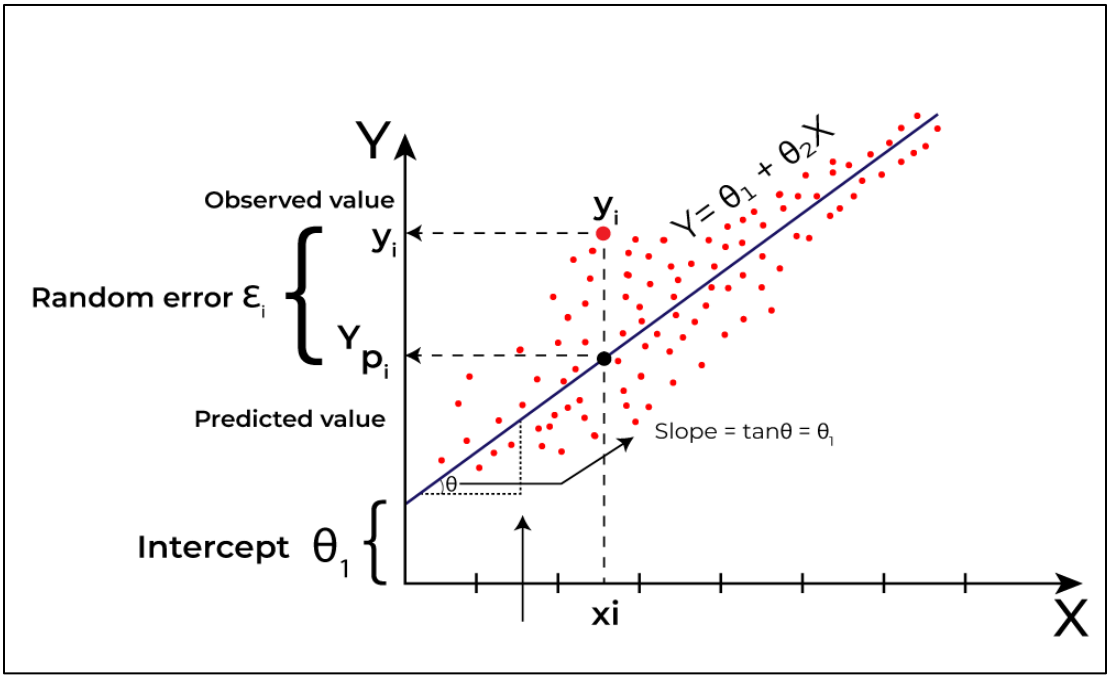

Figure 2: Linear regression

Cost function: $\boldsymbol{MSE} = \frac{1}{N}\sum_{i=1}^{n}(\boldsymbol{y_i} - (\boldsymbol{\Theta_0} + \boldsymbol{\Theta_1 X}))^2$

Usage:

- **Understanding Relationships:** Identify and quantify the influence of different factors on movie-related outcomes. For example:
  - Determine if higher budgets consistently lead to higher box office returns.
  - Analyze if certain genres tend to attract higher audience ratings.

    - Examine the impact of specific actors or directors on a film's commercial or critical success.
- **Feature Selection:** Determine which independent variables are the most important predictors, helping to simplify models and focus on the most relevant factors.

To assess our linear regression model and gain insights, we'll use:

- **R-squared (R^2):** This tells us how well our model fits the data. Think of it like a score from 0 to 1, where 1 means the model perfectly predicts the outcome, and 0 means it does no better than random guessing.
- **p-value:** This tells us if the relationship between a specific factor (like budget or genre) and the movie outcome (like box office gross) is likely to be real or just due to chance. If the p-value is small (usually less than 0.05), we can be more confident that the relationship is meaningful.

In this project, these metrics are combined with visual tools like charts and graphs to get a clear picture of how our model performs and which factors are most important in predicting movie success.

2. Decision Tree

Definition: The Decision Tree is a type of supervised learning algorithm used for both classification and regression problems. It is one of the most popular Machine Learning algorithms due to its simplicity and intuitive approach.

Algorithm: Decision Tree uses a tree structure as a predictive model. It breaks down a dataset into smaller subsets while at the same time an

associated decision tree is incrementally developed. It has decision points (nodes) that represent tests on features, and end points (leaf nodes) that represent the final categories or outcomes.

The training process involves selecting attributes that return the highest information gain (IG).

Pruning techniques, such as pre-pruning and post-pruning, are used to reduce the size of the tree, thereby reducing overfitting and improving the model's predictive accuracy.

- Pre-pruning is implemented by setting constraints such as maximum depth and minimum samples per split during the tree's construction.
- Post-pruning involves trimming branches post hoc to improve generalization and mitigate overfitting.

Usage: The Decision Tree model is trained on various features such as: movies, such as budget, genre, director, cast, and release date, etc. Each movie in the dataset is labeled as either “Successful” or “Unsuccessful”. The Decision Tree algorithm learns from this data, creating a feature-based classification model. This will be then used to predict the success of new, unseen movies.

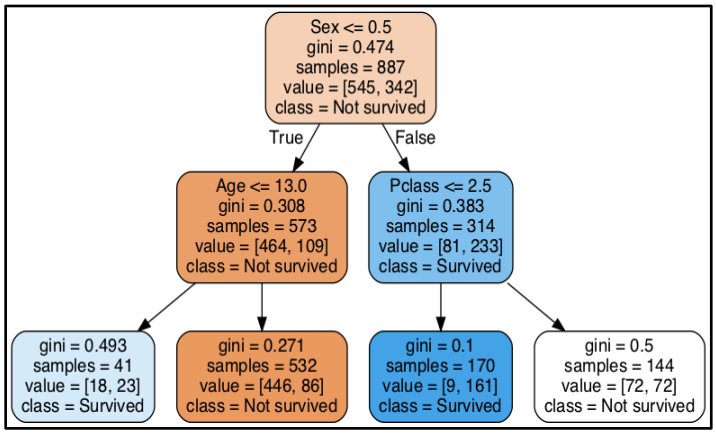

Figure 3: Decision Tree Classification - Example

3. GridSearchCV

Definition: GridSearch is a method used for hyperparameter optimization in machine learning.

Algorithm: GridSearch involves systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which gives the best performance. This technique can significantly improve the performance of a model by finding the optimal values for the hyperparameters. However, GridSearch can be computationally expensive, especially if the dataset is large or if there are many parameters to tune.

Usage: For Decision Tree, we define a grid of parameters, such as maximum depth, minimum samples split, and minimum samples leaf, among others. GridSearch then trains the Decision Tree model on all possible combinations of these parameters and uses cross-validation to evaluate the performance of each model. The optimal parameters are those that resulted in the highest cross-validation score. This approach allows us to fine-tune our Decision Tree model and achieve improved predictive accuracy.

IV. Experiment

1. Data Understanding and Mining

The dataset used in this analysis was sourced from IMDb and is available on Kaggle (https://www.kaggle.com/datasets/danielgrijalvas/movies?resource=download). This dataset comprises information on over 7,500 movies released between 1986 and 2020. It includes both quantitative variables (such as budget, gross, and IMDb score) and qualitative variables (genres, rating, country, etc.).

Key features of the dataset include:

- **Quantitative Variables:**
  - **Budget:** The financial resources allocated for film production.
  - **Gross:** The total box office revenue generated by the movie.
  - **Score:** The average user rating of the movie on IMDb.
  - **Votes:** The number of user votes contributing to the IMDb score.
  - **Runtime**: the length of the movie
- **Qualitative Variables:**
  - **Genres:** The categories or types of movies (e.g., action, comedy, drama).
  - **Rating:** The MPAA rating assigned to the movie (e.g., G, PG, R).
  - **Country:** The primary country of production for the movie.
  - **Stars:** The main actors featured in the movie.
  - **Director:** The individual(s) responsible for directing the movie.
  - **Writer:** The individual(s) responsible for writing the screenplay.
  - **Release Date:** The date on which the movie was released in theaters.
  - **Year**: The date on which the movie was released in theaters.

	name	rating	genre	year	released	score	votes	director	writer	star	country	budget	gross	company	runtime	month
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0	June
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0	Columbia Pictures	104.0	July
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	538375067.0	Lucasfilm	124.0	June
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	88.0	July
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0	39846344.0	Orion Pictures	98.0	July

Figure 4: First-five rows in dataset

Data Preprocessing

No	Status	Description	Action
1	Missing Value	name 0.000000 rating 1.004173 genre 0.000000 year 0.000000 released 0.026082 score 0.039124 votes 0.039124 director 0.000000 writer 0.039124 star 0.013041 country 0.039124 budget 28.312467 gross 2.464789 company 0.221701 runtime 0.052165	Removed missing value in Rating, Company, Runtime, gross. Replaced missing value in budget with 0
2	Duplicated Value	0%	None
3	Inconsistent Data + “Month” column has value “Unknown”	12 values	Removed
4	Change data type: +runtime (int) +Month (from string to int)		

The dataset is quite clean with low missing values; only the budget has over 20% missing. This can be easily understandable in the real world, as budgets often cannot be collected due to various reasons such as sponsorship, security, and others.

Distribution of numerical variables:

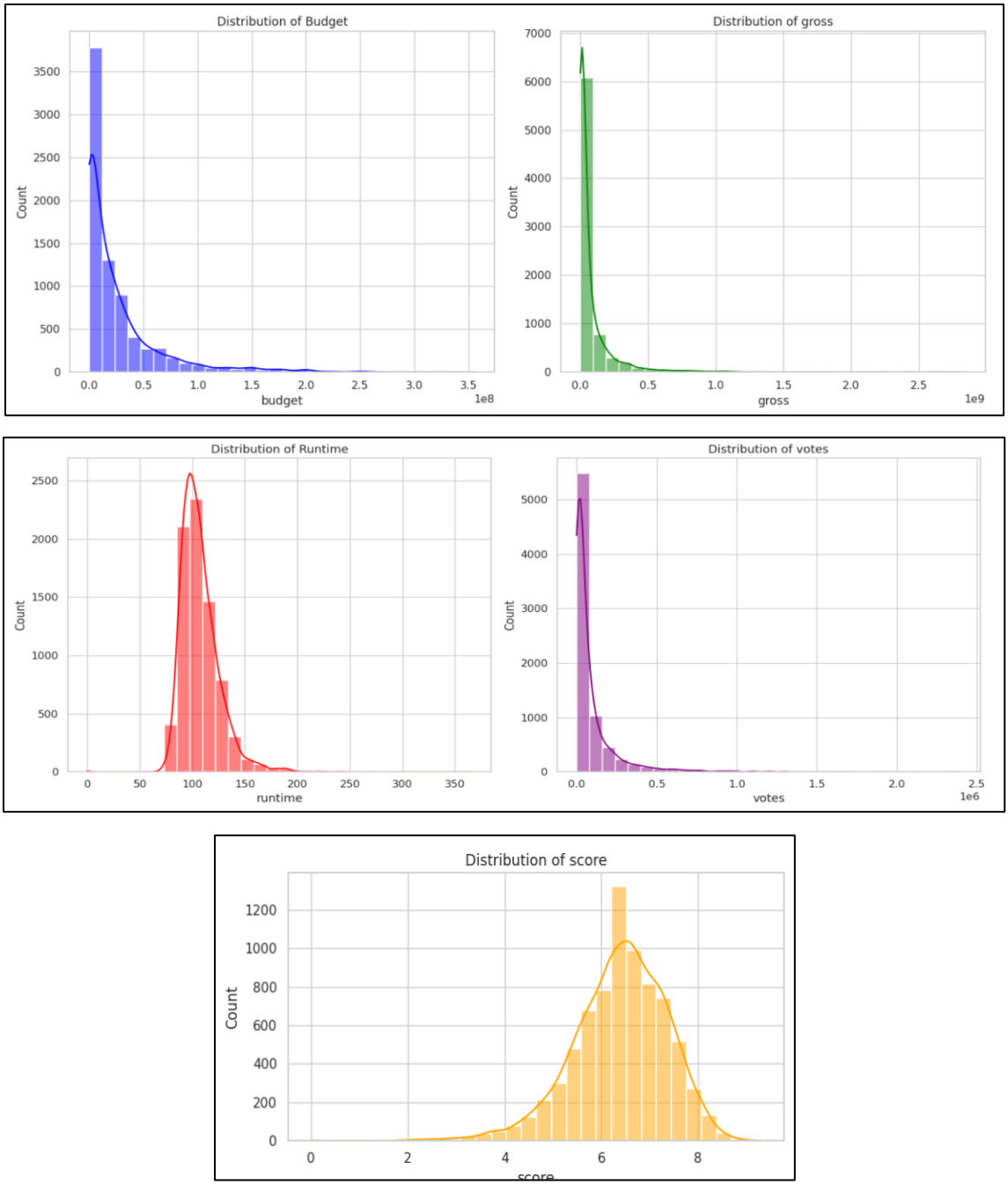

Figure 5: Distribution of numerical variables

Distributions of Key Movie Metrics (Budget, Gross, Runtime, Votes, Score):

- **Budget and Gross:** Both are heavily right-skewed, with most movies having moderate budgets and revenues, but a few outliers generating disproportionately large amounts. This suggests a potential positive relationship between budget and gross. The long tails highlight the industry's high-risk, high-reward nature.

- **Runtime:** Unimodal and slightly right-skewed, with a peak around 90-100 minutes. This suggests a typical film length that audiences generally prefer. The variation indicates some films cater to different preferences with longer runtimes.

- **Votes:** Extremely right skewed with a long tail. Most movies receive few votes, while a tiny fraction is incredibly popular.

- **Score:** Roughly bell-shaped (normal), slightly skewed to the left. Most movies receive moderate ratings clustered around the mean, indicating a wide range of critical and audience reception.

The relationship between Budget and Gross

To better understand the impact of budget on gross revenue, we will apply linear regression analysis to the data.

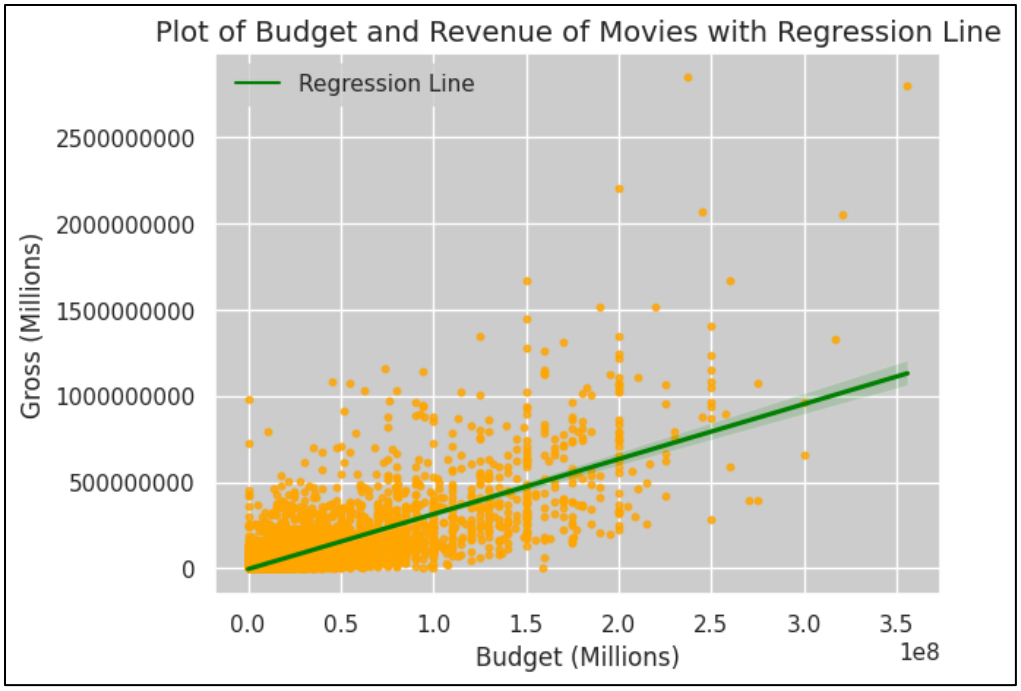

Figure 6: Budget and Gross relationship

The scatter plot of movie budgets and gross revenues clearly shows a positive relationship: movies with bigger budgets tend to make more money. This makes sense, as larger budgets often mean better special effects, star-studded casts, and wider marketing campaigns, all of which can attract larger audiences.

For instance, blockbuster films like *Avengers: Endgame*, *Pirates of the Caribbean: On Stranger Tides*, and *Avengers: Age of Ultron* all had massive budgets (around $350-400 million) and went on to achieve incredible box office success, earning well over a billion dollars each.

However, the relationship isn't always straightforward. The scatter plot also reveals that not all big-budget movies are guaranteed hits, and some smaller-budget films can still achieve significant success. This suggests that factors other than budget also play a crucial role in a movie's performance.

Will the high revenue movie have a high IMDB score?

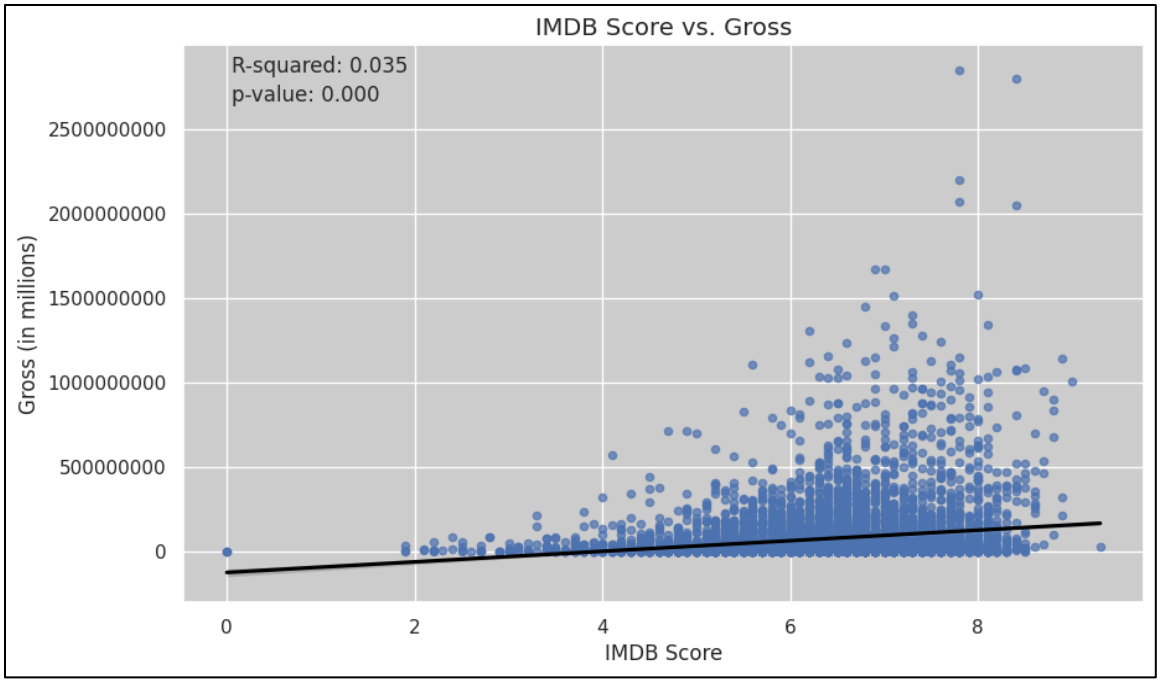

Figure 7: Score and Gross relationship

The scatter plot reveals a weak, yet statistically significant, positive relationship between a movie's IMDB score and its gross revenue, although the low R-squared value (0.035) indicates this relationship is not strong. However, the score has a strong relationship with other variables such as run time, votes, budget, country.

This suggests that an IMDB score can be a factor in determining a movie's success.

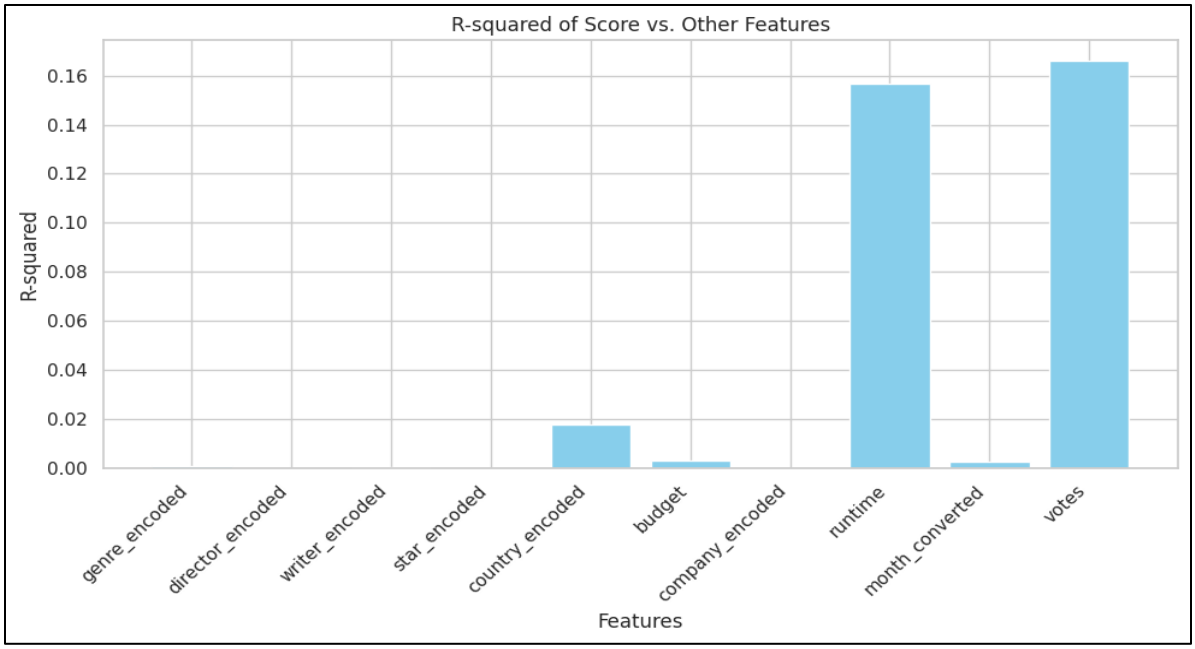

Figure 8: Score and Gross relationship

What defines a successful movie?

There is no single, definitive criterion for a successful movie. However, financial performance is a crucial factor, as revenue is needed to cover expenses like staff salaries, equipment, marketing, and other costs. A common rule of thumb is that a movie must earn at least twice its budget to be considered financially successful.

Given our previous observations that IMDB score alone has a weak relationship with revenue and is influenced by various other variables, we propose a more comprehensive definition of success. For this analysis, we'll define a successful movie as one that meets the following criteria:

- **Gross Revenue:** Exceeds at least twice the movie's budget.
- **IMDB Score:** At least 7 out of 10.

We have introduced a new column named "Successful," which classifies movies as either "1" (Successful) or "0" (Flop).

Analyze categorical variables

The reputation and star power of directors and actors are significant factors in attracting audiences and boosting a movie's box office potential. Their involvement can create anticipation, trust, and excitement, leading to higher ticket sales. Therefore, it's crucial to consider their influence when analyzing a film's financial performance.

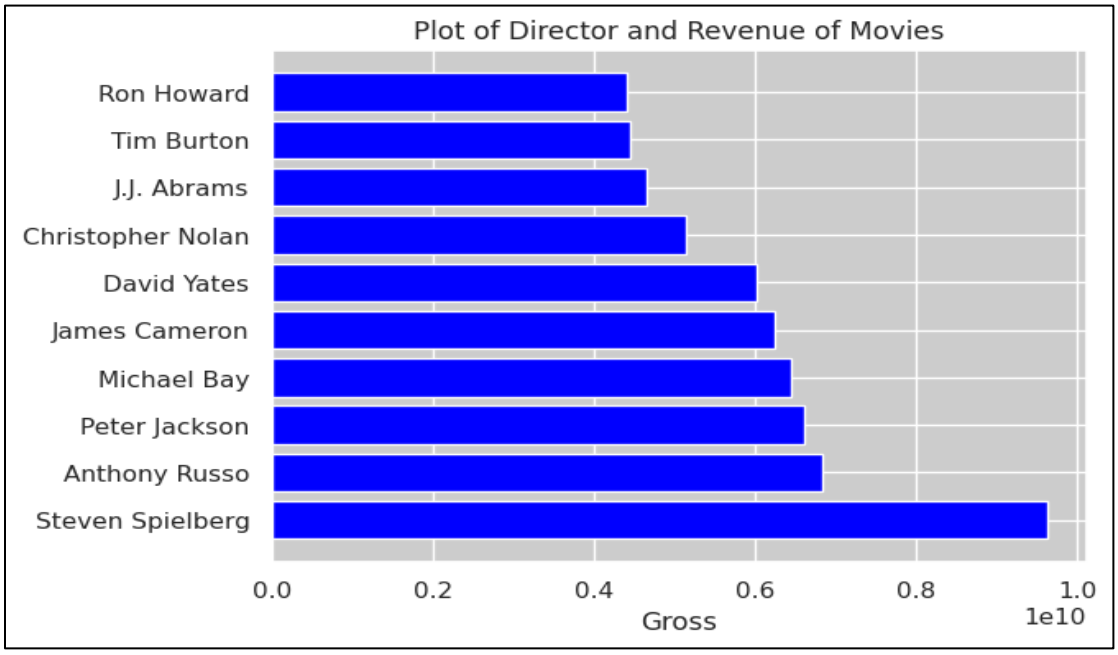

Figure 9: Director and Gross relationship

The substantial differences in average gross revenue among directors highlight their significant impact on a film's financial success. This influence can be attributed to factors such as the director's reputation, style, ability to attract top talent, and experience.

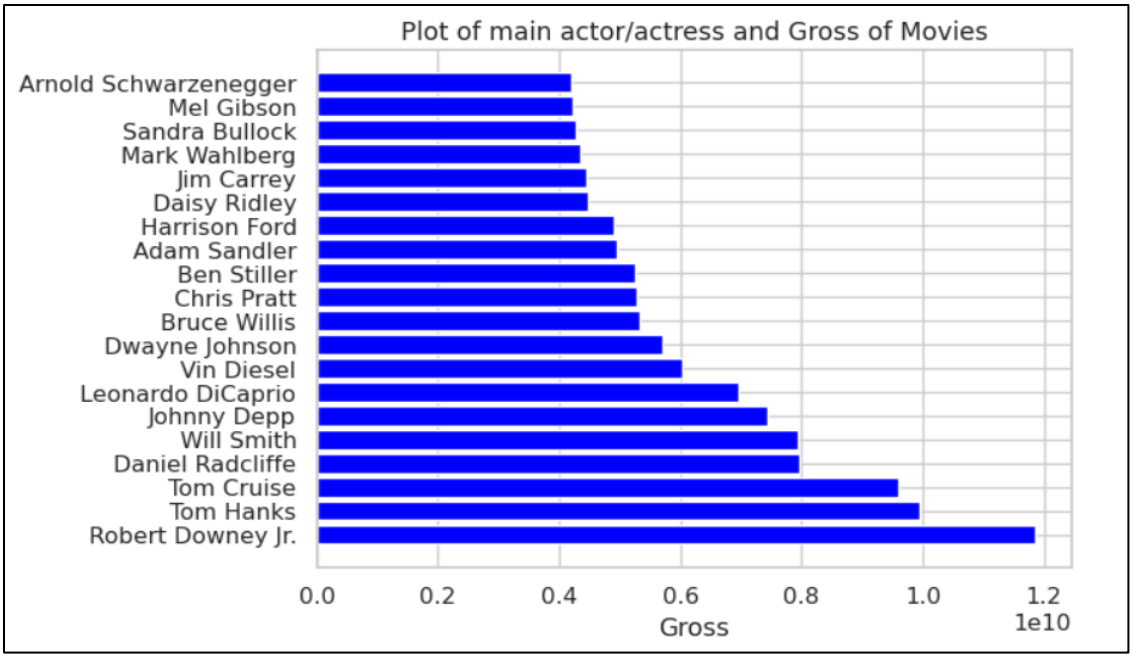

Figure 10: Star and Gross relationship

Star Power: The variation in gross revenue among actors underscores the significant impact an actor's presence can have on a film's financial performance. This could be attributed to factors like the actor's popularity, critical acclaim, association with successful franchises, and overall appeal to different demographics.

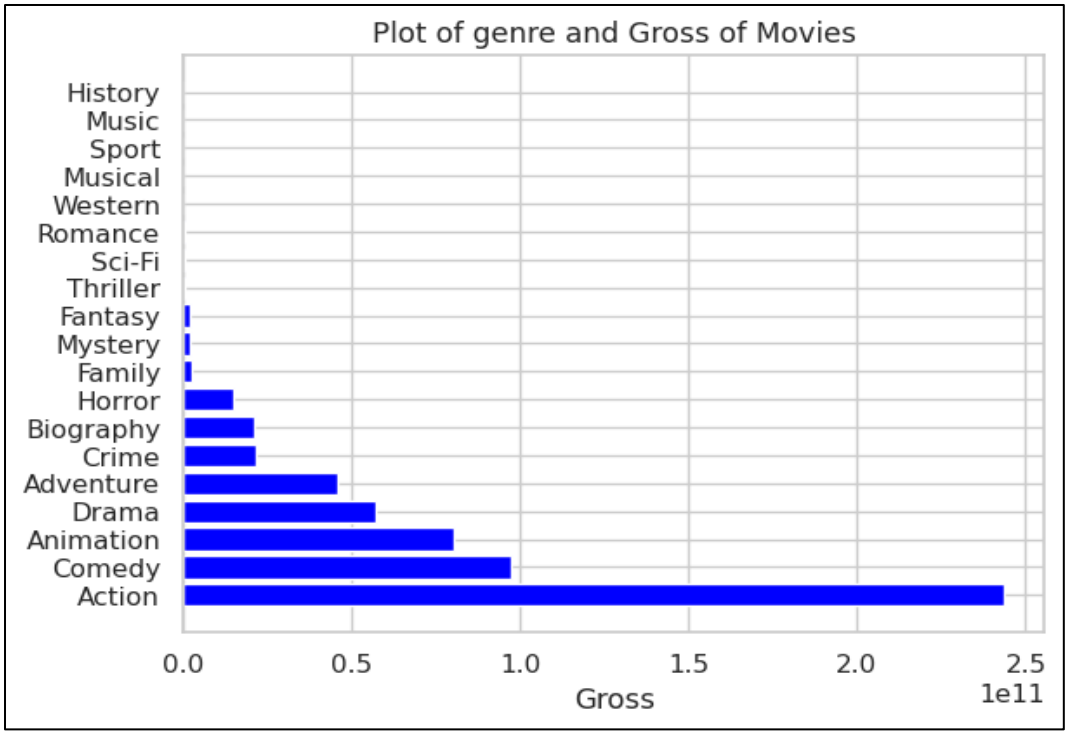

Figure 11: Genre And Gross relationship

Top-Grossing Genres: Action, Comedy, Adventure, Drama, and Animation films consistently outperform other genres at the box office. This suggests these genres have a broad appeal and the potential to attract large audiences, making them potentially safer investments for studios.

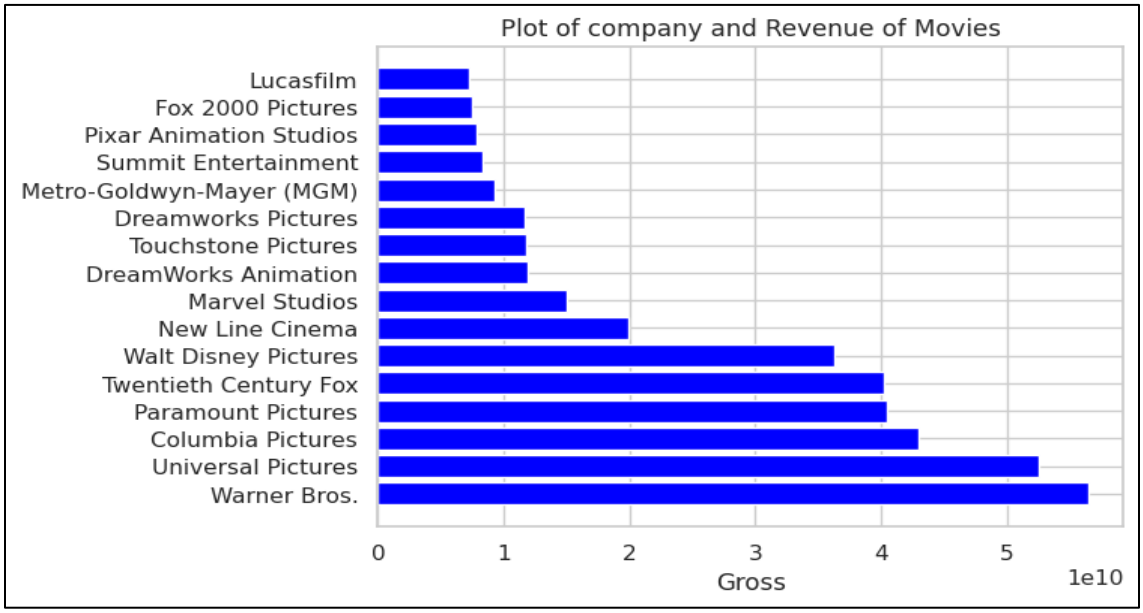

Figure 12: Company and Gross relationship

Market Share: This chart provides insights into the market share of different production companies. The top three companies seem to dominate the market, while the others compete for a smaller portion.

Dominance of Major Studios: The significantly higher gross revenue of some giant companies, like Warner Bros., Universal Pictures, and Columbia Pictures, compared to smaller ones indicates a competitive advantage in attracting viewers. This could be due to factors like established brand recognition, larger marketing budgets, and the ability to secure top talent.

To assess the impact of directors, writers, and genres on movie profitability, we'll assign them scores columns based on how their films perform financially compared to the average profit of all movies.

- **Score = 5:** If the total profit of all their movies exceeds the average profit.
- **Score = 3:** If the total profit of all their movies roughly equals the average profit.
- **Score = 1:** If the total profit of all their movies is less than the average profit.
- **Score = 0:** If the total profit of all their movies is negative (they incurred a loss).

2. Modeling and Evaluation

a. Success Movie Classification:

In this research, we utilize a Decision Tree classifier for movie success prediction. We conduct two sets of experiments with the Decision Tree classifier, one with raw features and another with processed features. The comparative results from these experiments substantiate the accuracy of our Data Insight. GridSearchCV is incorporated for hyperparameter tuning, supplemented with pre-pruning and post-pruning strategies.

After data analysis, key features are as follows: **budget, runtime, rating_converted, director_score, company_score, genre_score, star_score, writer_score, and month_converted.** They are expected to significantly contribute to the performance of our predictive model.

Pre-pruning with optimal hyperparameters

Without data insights, a Decision Tree model is trained using GridSearch, a method for hyperparameter tuning. The optimal parameters are: 'class_weight' set to None, 'criterion' set to 'gini', 'max_depth' set to 5, 'min_samples_leaf' set to 4, 'min_samples_split' set to 2, 'random_state' set to 42, and 'splitter' set to 'best'. The MAE is approximately 0.295 and the accuracy is approximately 70.43%.

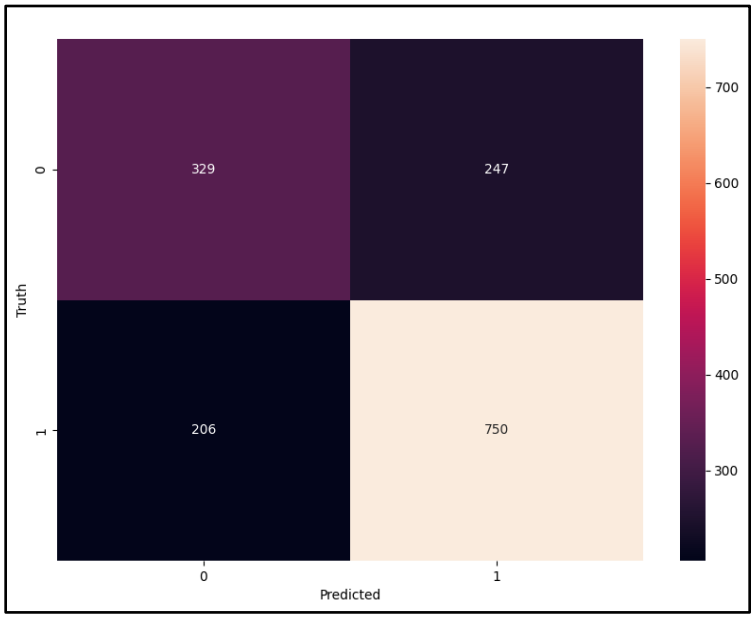

Figure 13: Confusion matrix (Decision Tree without Data insight)

With our data insight, the hyperparameters of the Decision Tree classifier are 'class_weight': None, 'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'random_state': 42, 'splitter':

'random'. The MAE is 0.1534 and the accuracy score is approximately 84.66%.

Comparison: With a nearly 0.15 reduction in MAE and 15% increase in the accuracy score, our data analysis methodology appears to be well-suited for the task.

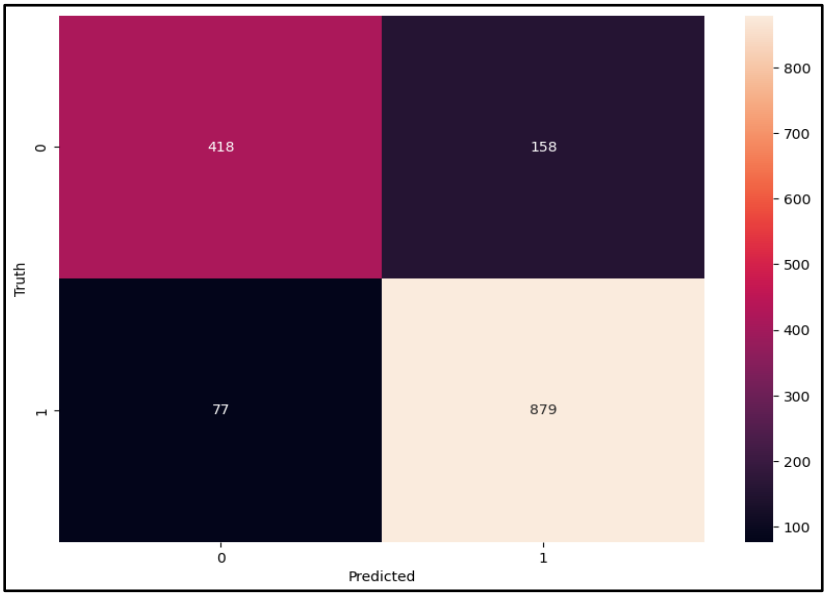

Figure 14: Confusion matrix (Decision Matrix with Data insight)

b. Improvement with post-pruning

We implement the post-pruning technique and select the alpha. Nodes are removed from the tree if the accuracy of the model does not improve after the split. MAE is about 0.129 and Accuracy score is about 87.08%.

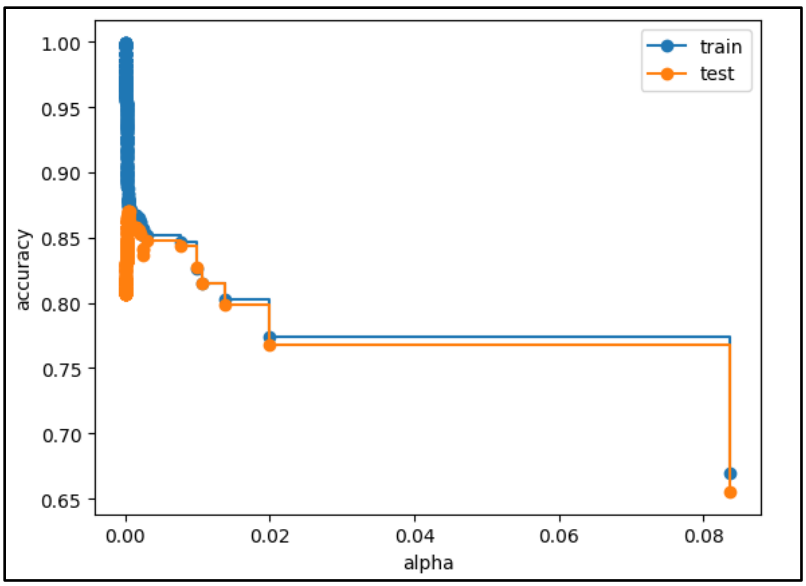

Figure 15: Accuracy vs alpha for training and testing sets

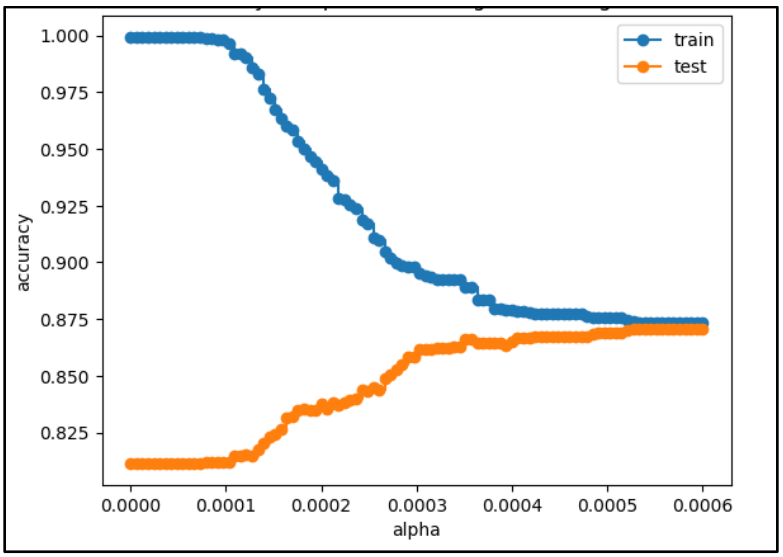

Figure 16: Accuracy vs alpha for training and testing sets (alpha from 0.0000 to 0.0006)

Comparison: With Pros-pruning, the result improves with almost 2.5% increase in accuracy and ultimate hyperparameters for model are listed:

- class_weight: None
- criterion: gini
- max_depth: None

- min_samples_leaf: 1
- min_samples_split: 2
- random_state: None
- splitter: best

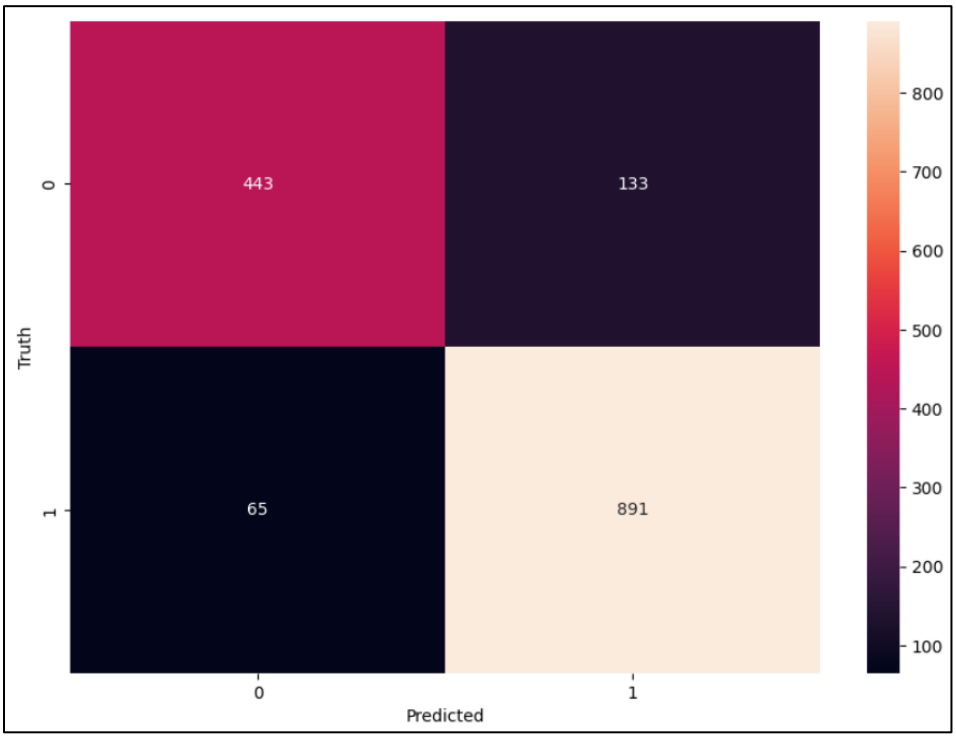

Figure 17: Accuracy vs alpha for training and testing sets (alpha from 0.0000 to 0.0006)

Evaluation

Model	Accuracy score	Increase
Decision Tree (without data insight)	70.43%	
Decision Tree (with data insight)	84.66%	~15%
Decision Tree	87.08%	~2.5%

Model	Accuracy score	Increase
Decision Tree (without data insight)	70.43%	
Decision Tree (with data insight)	84.66%	~15%
(pos pruning)		

Table 1: Comparison 3 methods in terms of Accuracy score

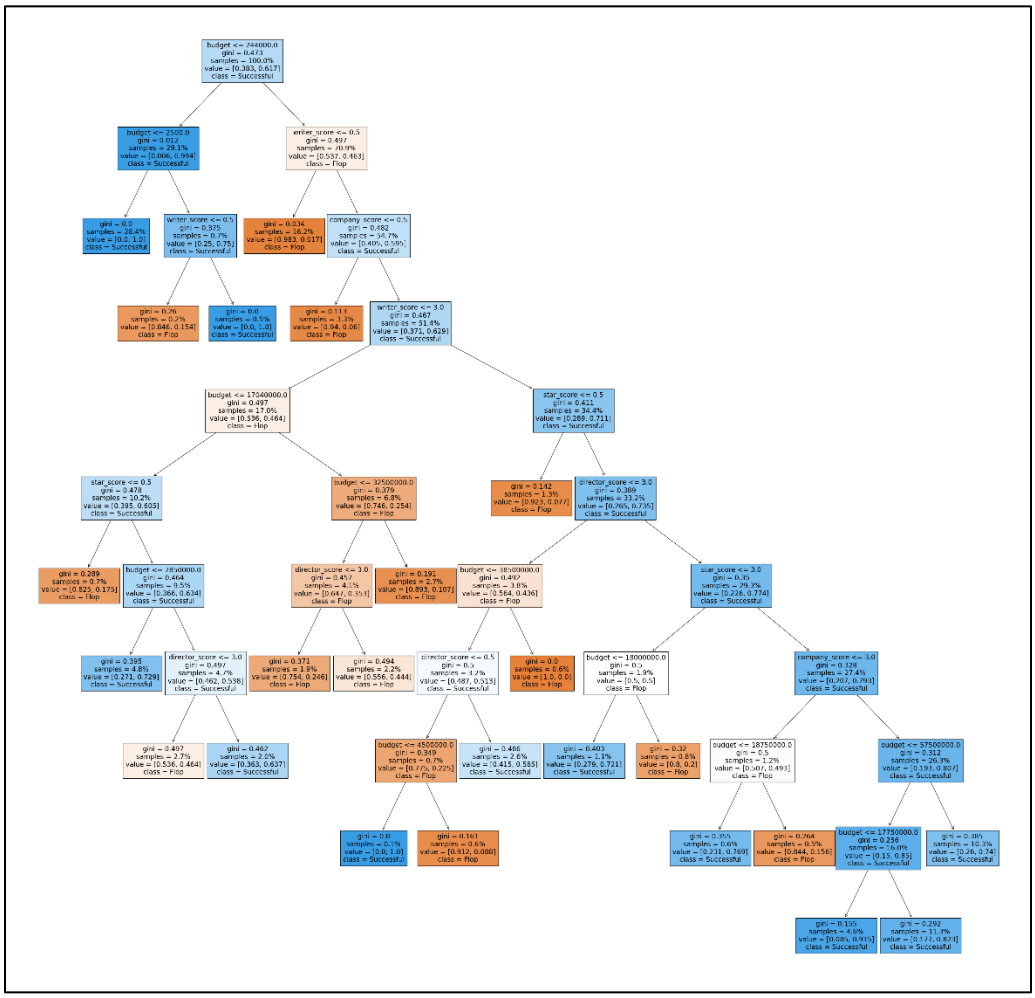

Figure 18: Final Decision Tree

V. Conclusion

Our analysis reveals that predicting movie success is a complex endeavor, influenced by a multitude of factors beyond simply critical acclaim. By understanding the nuances of audience engagement and the impact of key figures like directors and writers, studios and investors can make more informed decisions, potentially leading to increased profitability and a more vibrant film landscape.

This analysis offers valuable insights into the factors that contribute to a movie's financial success, but it's just the beginning. By incorporating additional data sources and refining our model, we can develop even more sophisticated predictive tools to empower decision-makers in the film industry. The future holds exciting possibilities for leveraging data-driven insights to shape the future of cinema.

Our source code: https://github.com/IrisPham74/CS331.O21.KHCL.git

VI. Task Assignment

	Research and Data Mining	Implement Application	Prepare Report	Create Slide	Presentation
Huy Hoang	x		x	x	x
Tram Anh	x		x	x	x
Dinh Duc	x	x	x		
Nhat Long	x	x	x		
Truong Thien	x	x	x		

VII. References

- https://www.kaggle.com/datasets/danielgrijalvas/movies
- https://scikit-learn.org/stable/modules/tree.htm
- https://aws.amazon.com/what-is/linear-regression/#:~:text=Linear%20regression%20is%20a%20data,variable%20as%20a%20linear%20equation.
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.htmll
- https://patshih.luddy.indiana.edu/publications/Gao-MovieSuccess-iConf19.pdf
- https://www.researchgate.net/publication/24072360_Successful_Movies_A_Preliminary_Empirical_Analysis