FINAL PROJECT -
DATA MINING AND APPLICATIONS

TOPIC: FORECASTING CINEMA BOX OFFICE HITS: A DATA-DRIVEN APPROACH

LECTURER: VO LE NGUYEN DUY

PHAN HUY HOANG : 21520242

PHAM THI TRAM ANH: 21520146

DUONG VAN NHAT LONG: 20521561

TRUONG QUANG THIEN: 20520310

LE DINH DUC: 19521372

TABLE OF CONTENTS

INTRODUCTION

Business Understanding

CONTEXT

- The film industry is booming, with millions of new releases.
- High profit potential exists, but choosing the right films is crucial.

SOLUTION

- A decision-support tool that uncovers hidden patterns and insights to inform strategic film selection.

NEED

- Cinema owners face the daunting task of selecting which movies to invest in and showcase.

VALUE

- Enable accurate and quick decisions based on ambigious information to identify successful movies.

METHOD

METHODOLOGY

GET INSIGHTS — Linear Regression

CLASSIFY — Decision Tree

ULTIMATE — GridSearchCV

METHODOLOGY

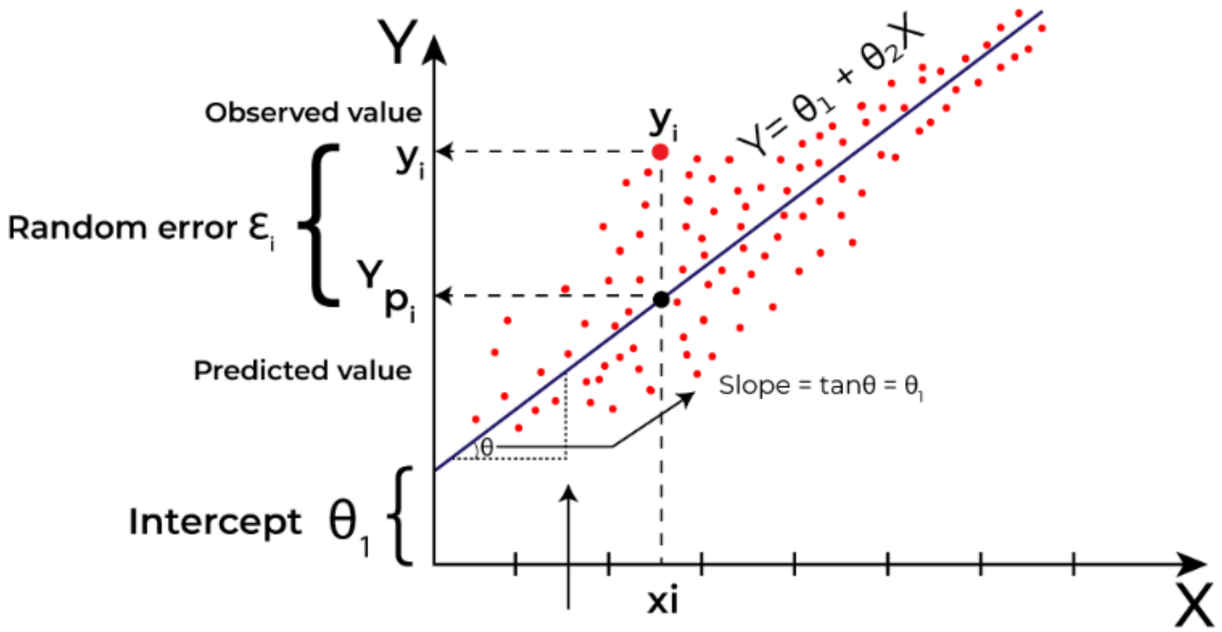
1

- **Understanding Relationships:** Identify and quantify the influence of different factors on movie-related outcomes.

2

- **Feature Selection:** Determine which independent variables are the most important predictors, helping to simplify models and focus on the most relevant factors.

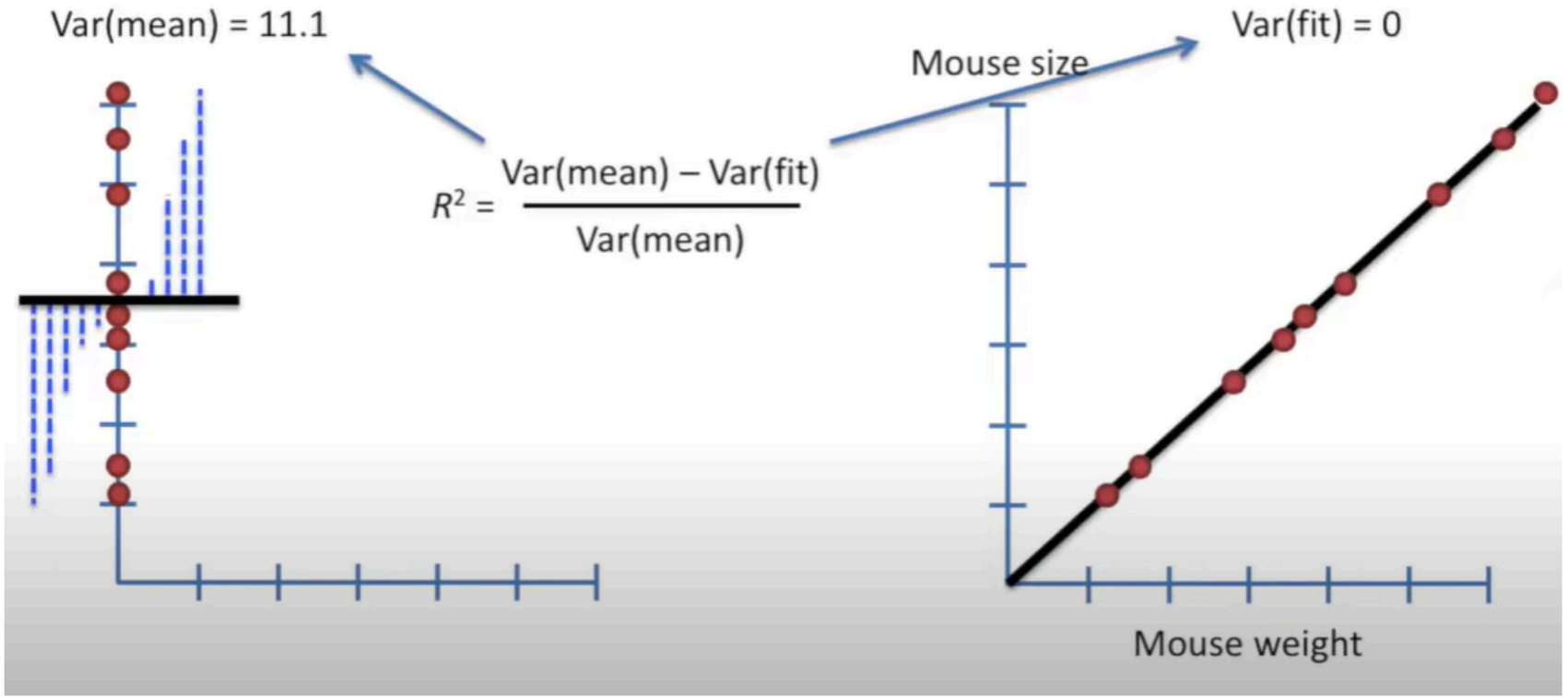
METHODOLOGY



Linear Regression

- **Statistical model to examine relationship between dependent variable (outcome) and independent variable(s) (influencers).**
- **Uses ordinary least squares (OLS) to find the best-fitting line/plane.**

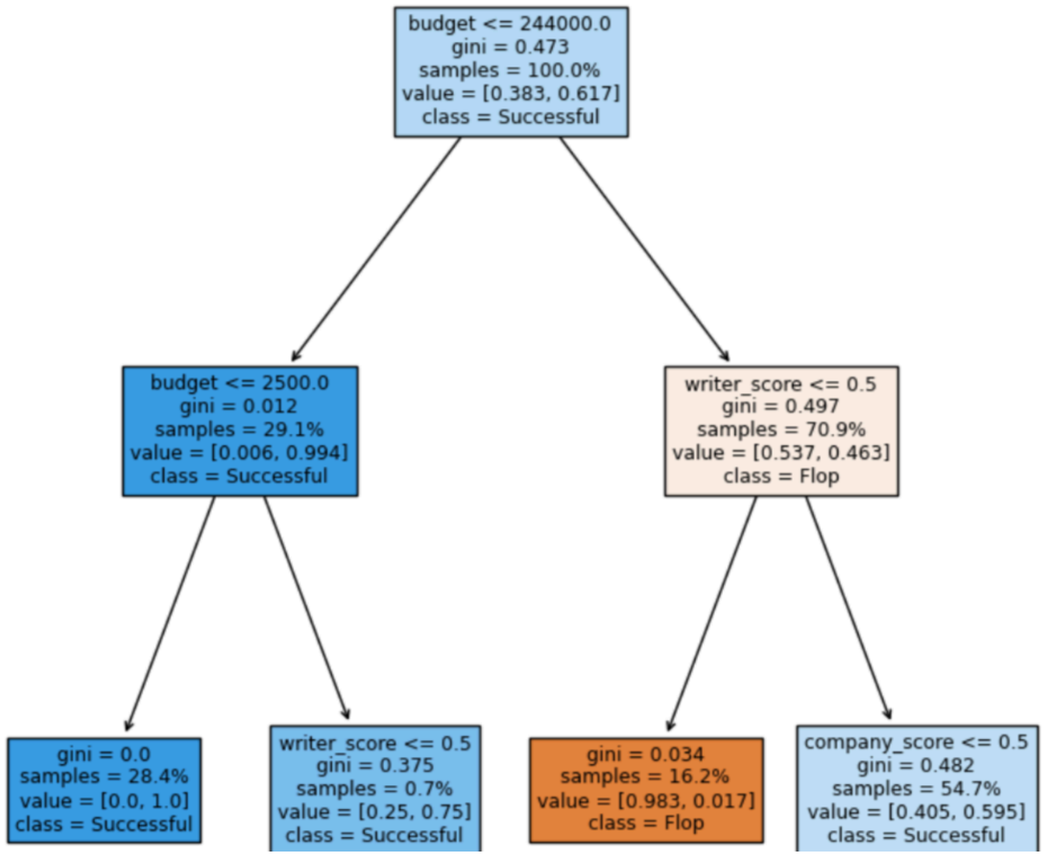
METHODOLOGY



R Square value

Use R Square to see how well the model fit and conclude the relationship

METHODOLOGY



Decision Tree

- **Supervised learning algorithm used for both classification and regression problems**

- **A tree structure**
  - Nodes: decision points.
  - Leaves: final outcomes.
  - Selecting attributes that return the highest information gain (IG).

- **Techniques:**
  - Pre-pruning
  - Post-pruning

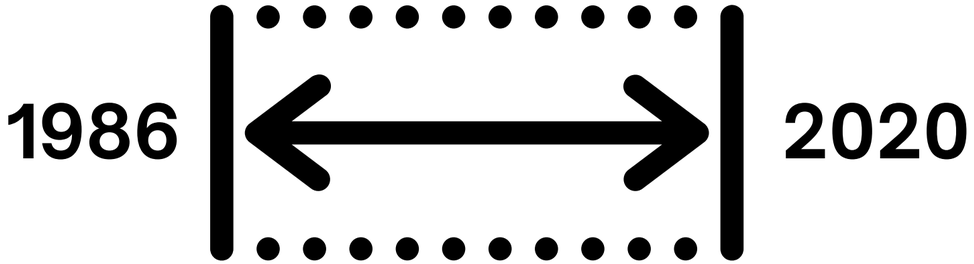
METHODOLOGY

- Hyperparameter optimization method
- Working through multiple combinations of parameter tunes, cross-validating as it goes to determine which gives the best performance.
- Define a grid of parameters, such as maximum depth, minimum samples split, and minimum samples leaf

DATA UNDERSTANDING AND MINING

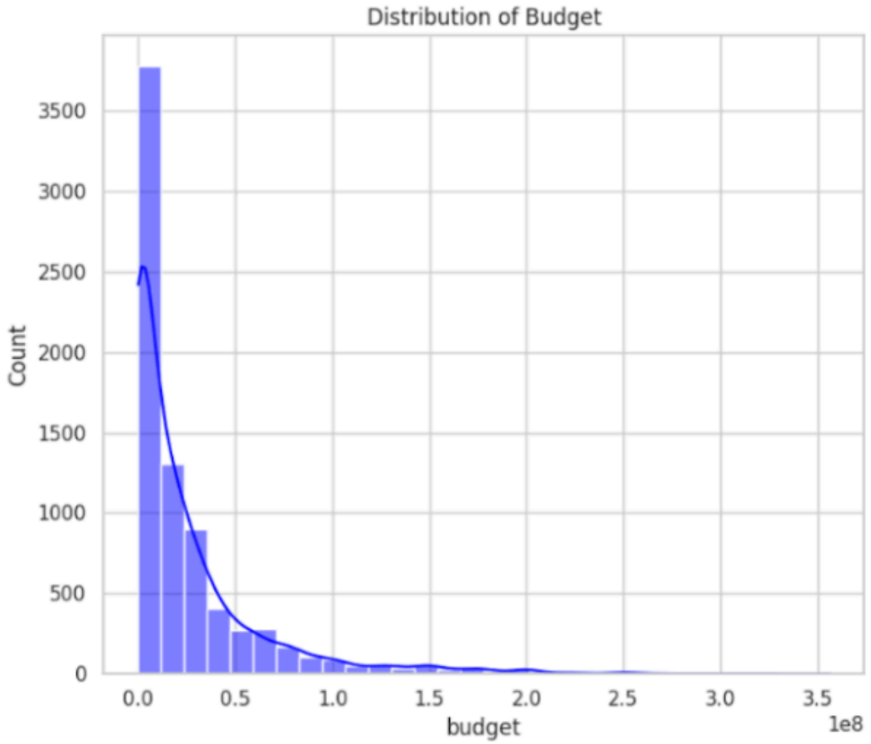
Dataset

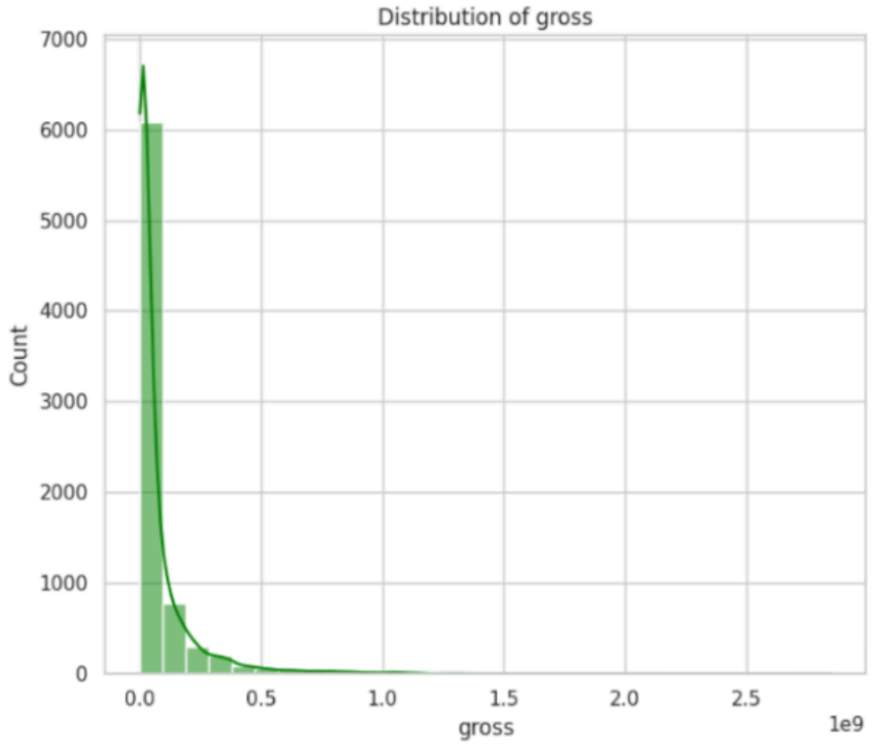
7,500 movies



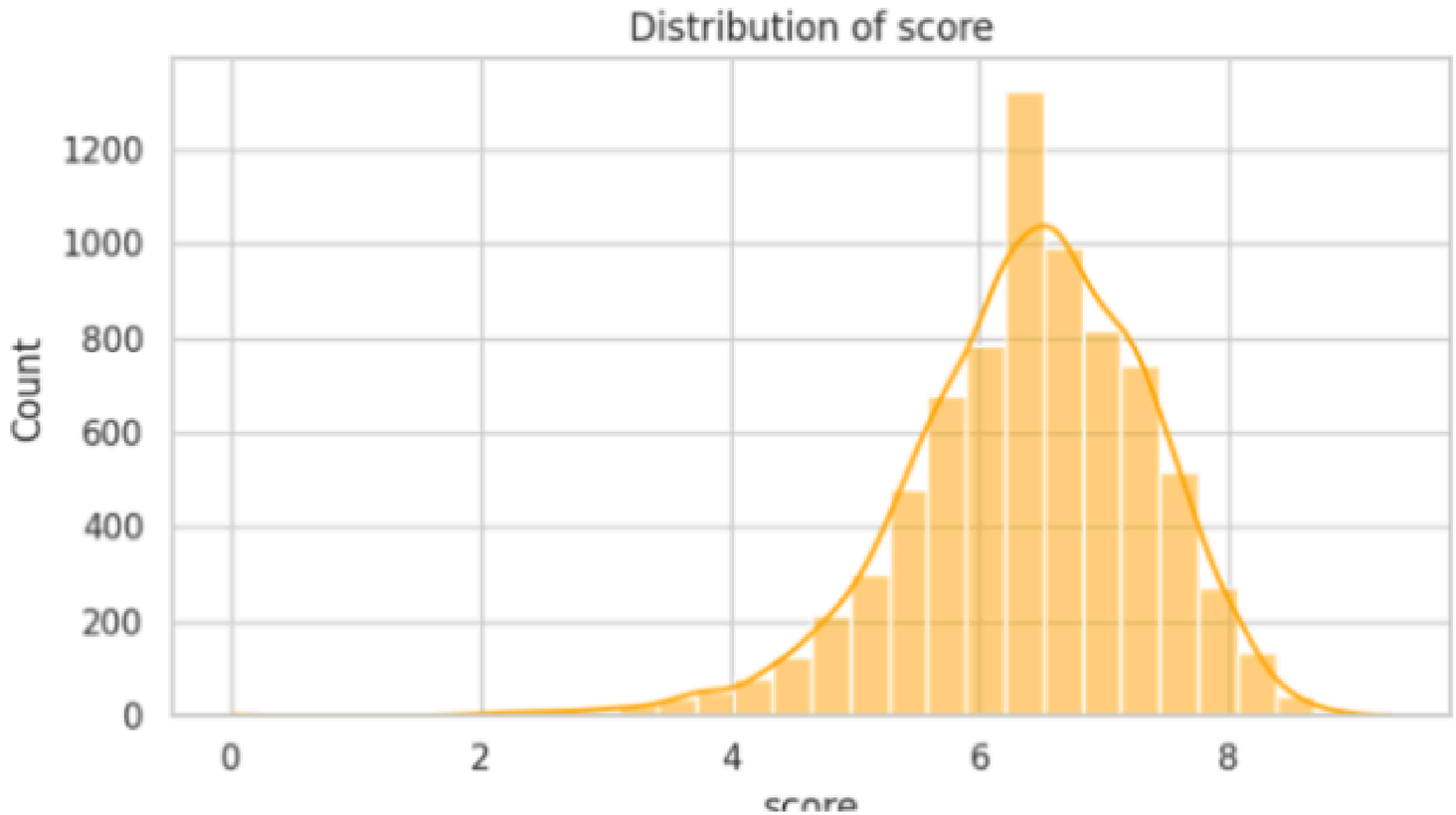
15 columns

Data Understanding and Mining

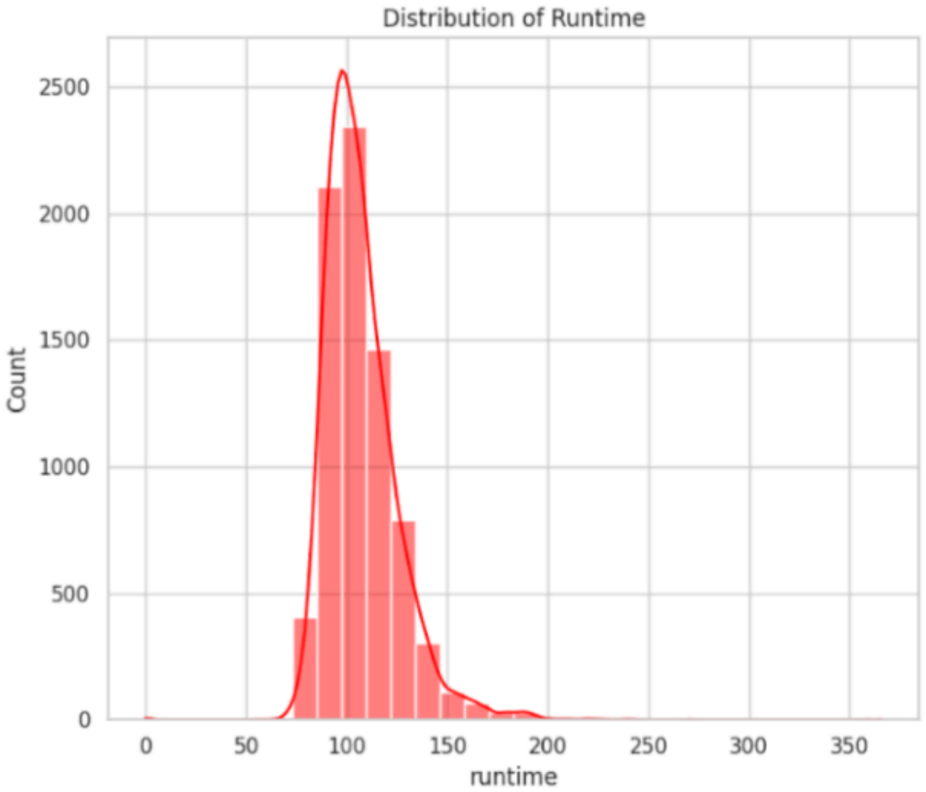


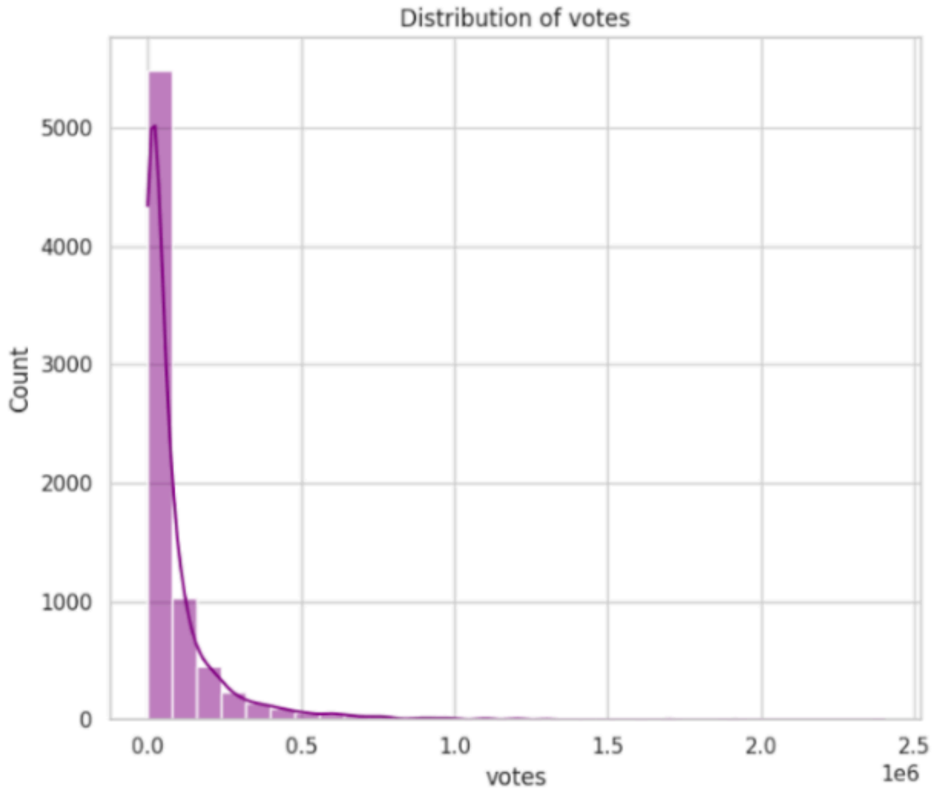


Data Understanding and Mining

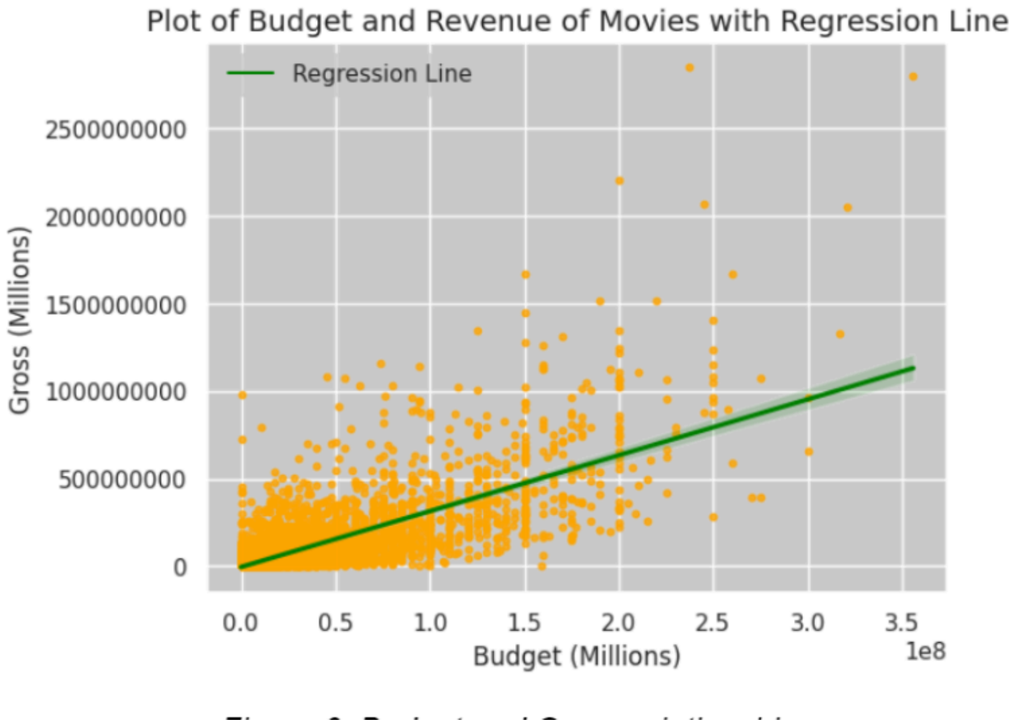


Data Understanding and Mining





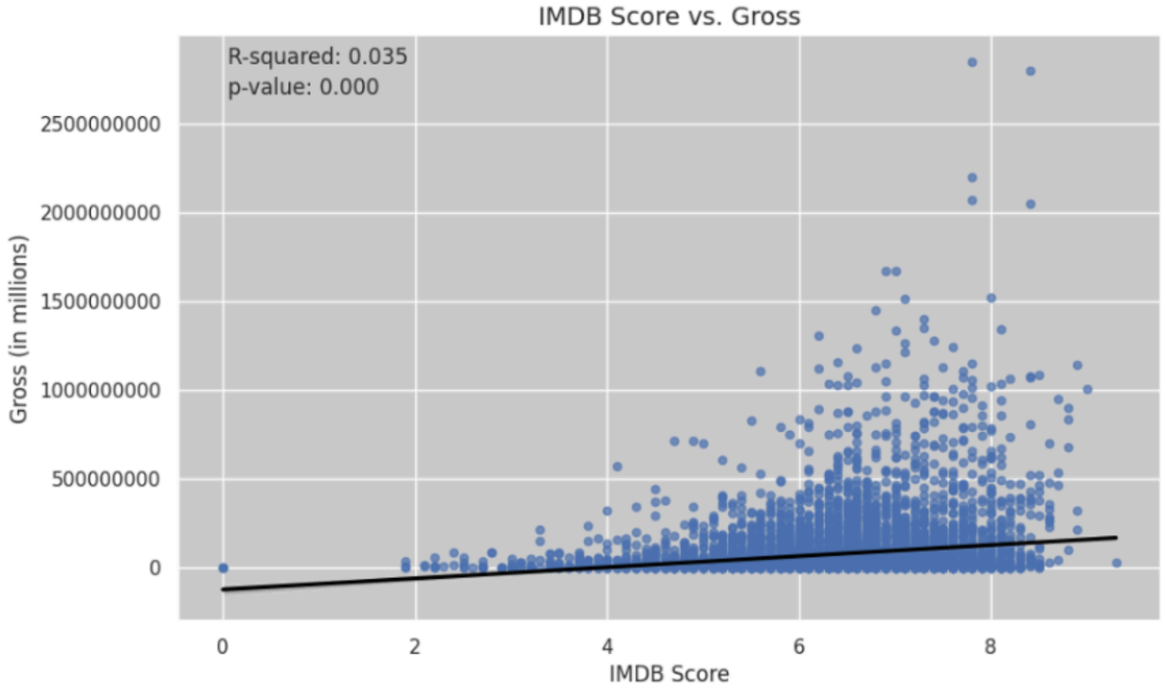
Data Understanding and Mining

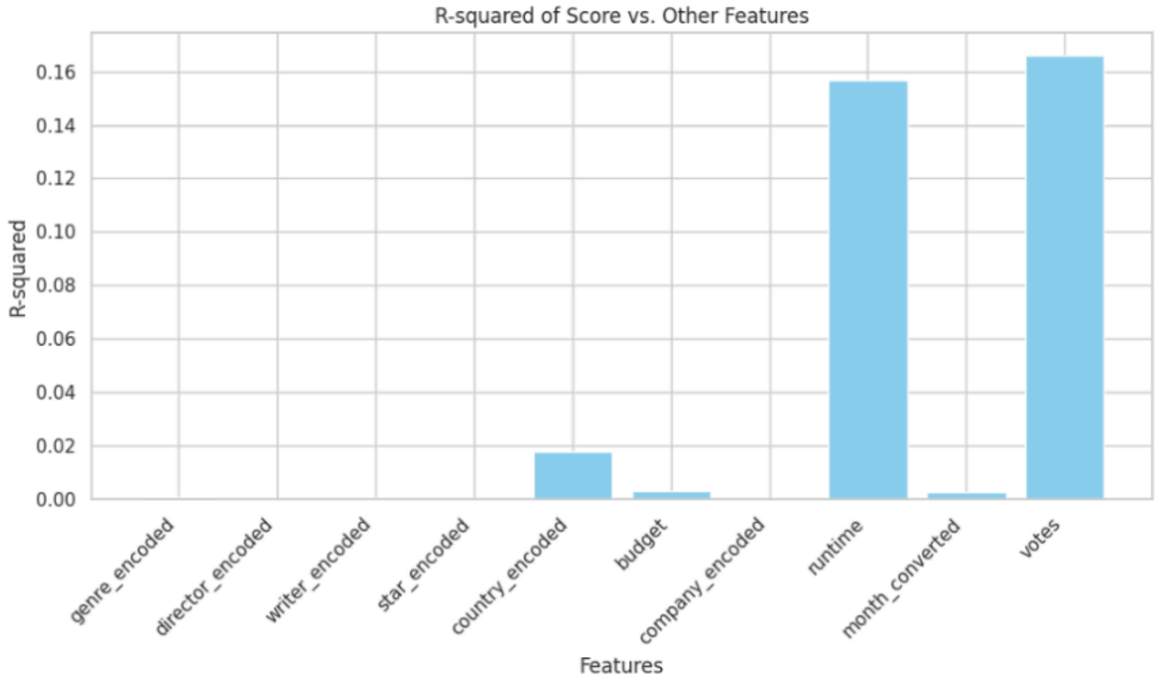


Larger budgets often mean better special effects, star-studded casts, and wider marketing campaigns, all of which can attract larger audiences.

- **A positive relationship:** movies with bigger budgets tend to make more money.

Data Understanding and Mining

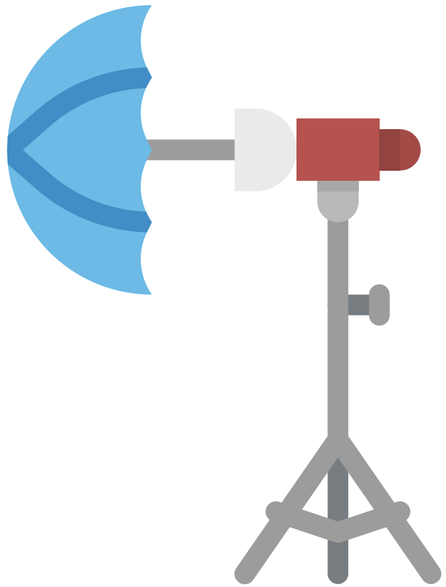




A weak positive relationship: movies with bigger budgets tend to make more money.

A strong relationship with other variables such as run time, votes, budget, country.

What is a successful movies



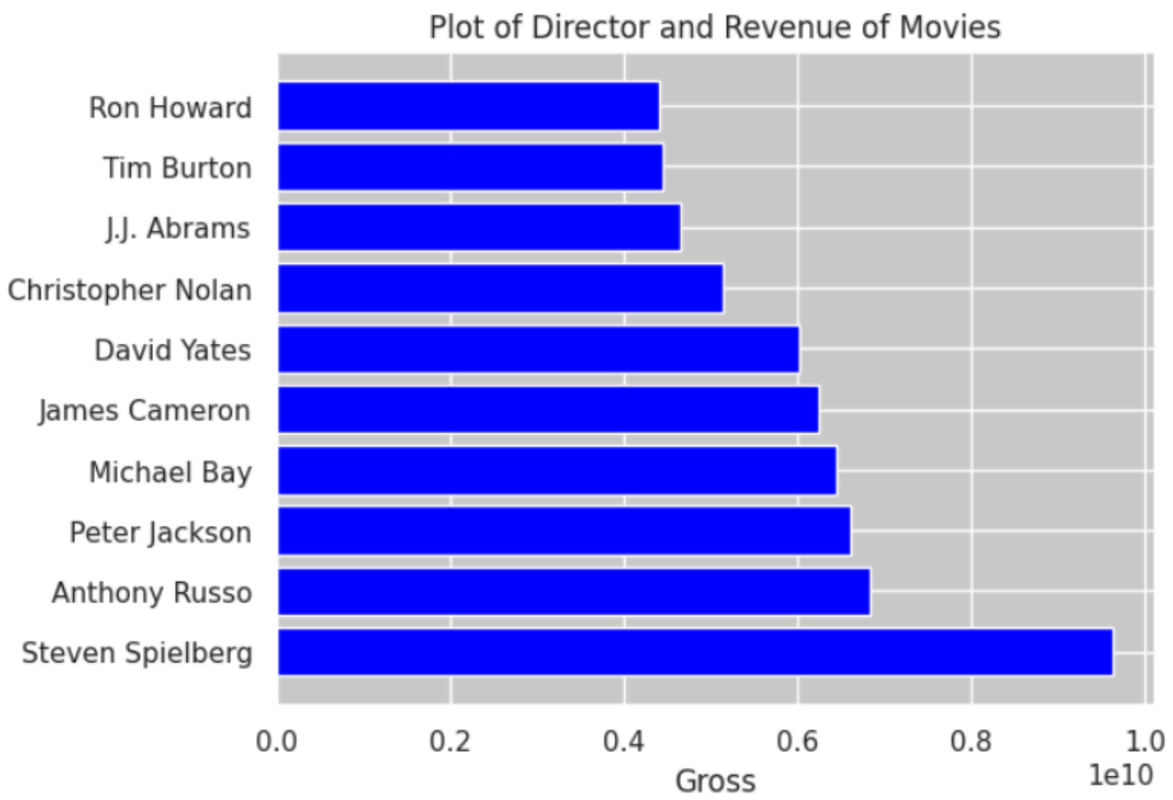
What is a successful movies

Rule of thumb

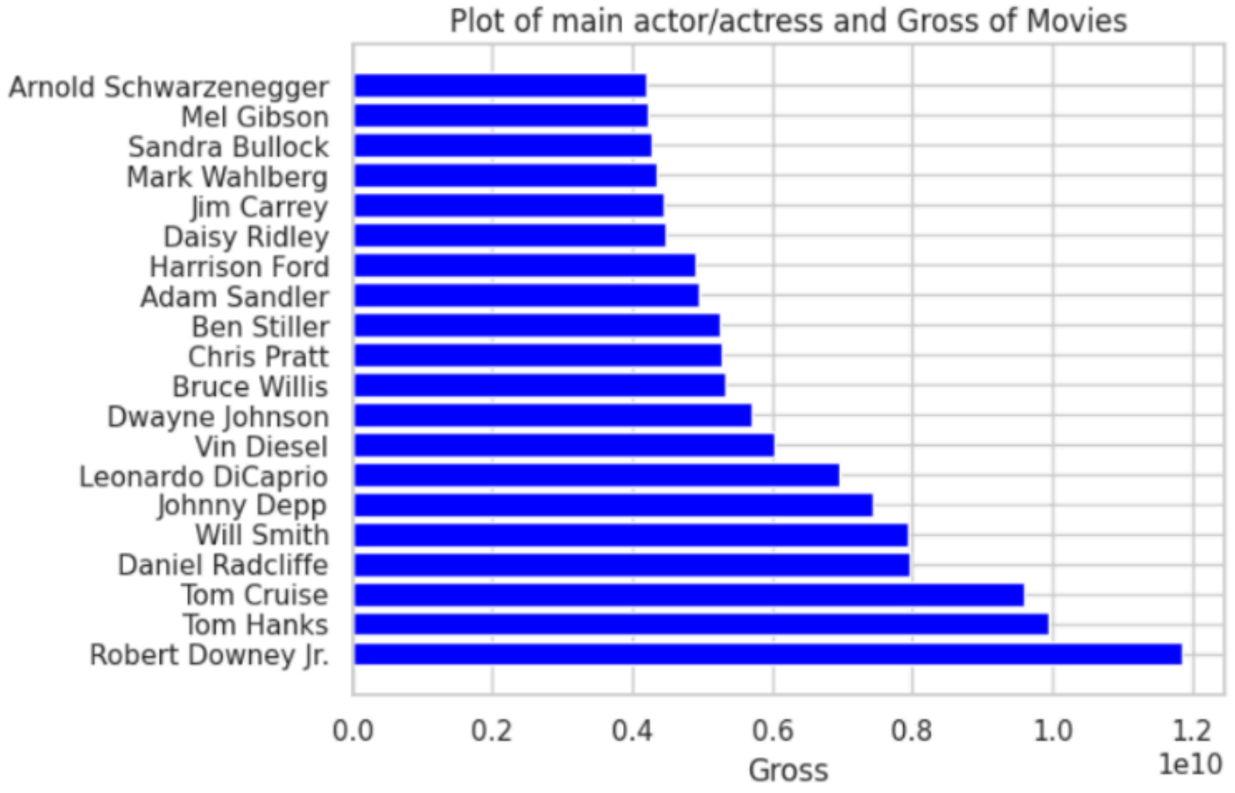
Gross Revenue: Exceeds at least twice the movie's budget.

IMDB Score: At least 7 out of 10.

Analyze categorical variables

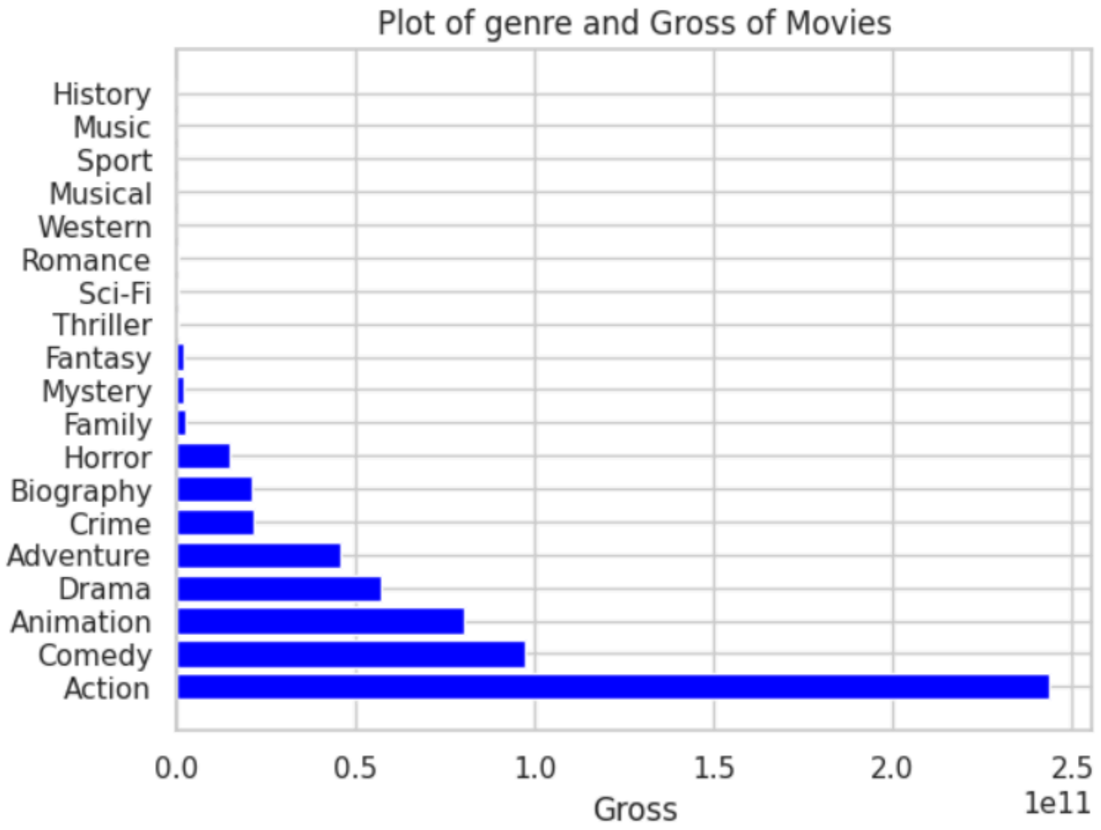


Director power based on the difference average gross

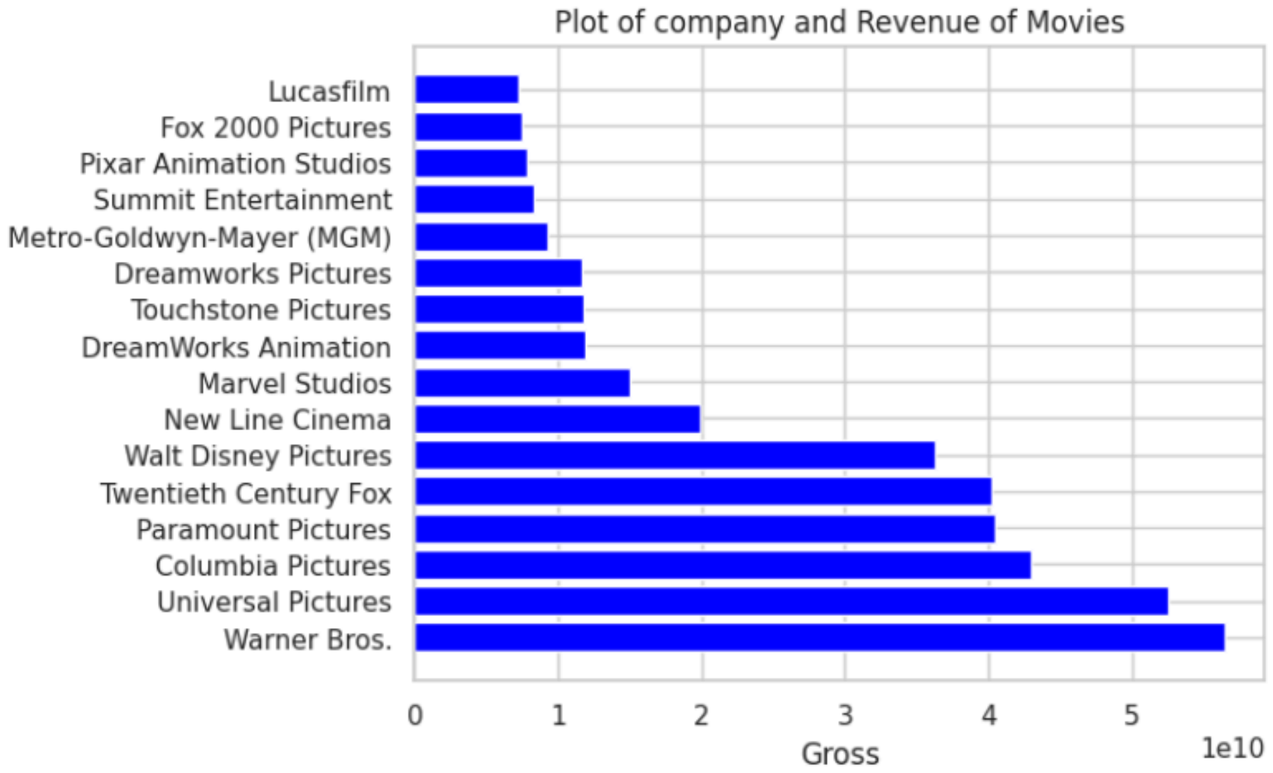


Star power based on the difference average gross

Analyze categorical variables



Top-Grossing Genres



Dominance of Major Studios

Assess the impact of directors, writers, and genres on movie profitability

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).

MODELING AND EVALUATION

Modelling & Evaluation

Key features:

- Budget
- Runtime
- Rating_converted
- Director_score
- Company_score
- Genre_score
- Star_score
- Writer_score
- Month_converted

TRAINING PROCESS

1

DICISION TREE WITHOUT DATA INSIGHTS

2

DICISION TREE WITH DATA INSIGHTS

3

DECISION TREE WITH POST-PRUNNING

4

PLOT TREE TO UNDERSTAND REASONS

DECISION TREE WITHOUT DATA INSIGHTS

Input:

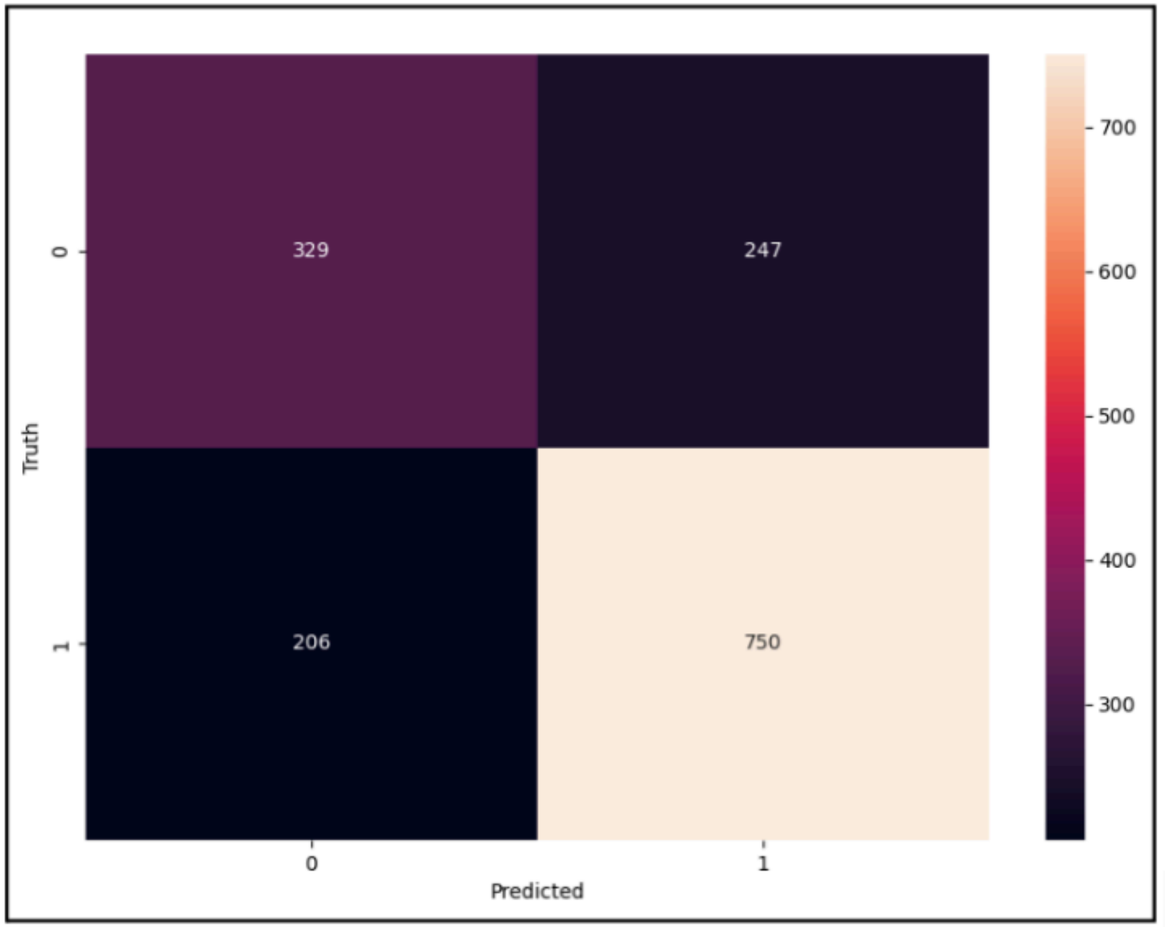
- 'rating', 'genre', 'year', 'released', 'writer', 'star', 'country', 'budget', 'company', 'runtime'

Output:

- ‘Successful’ or ‘Not successful’

Result:

- Accuracy: 70.43%.



DECISION TREE WITH DATA INSIGHTS

Input:

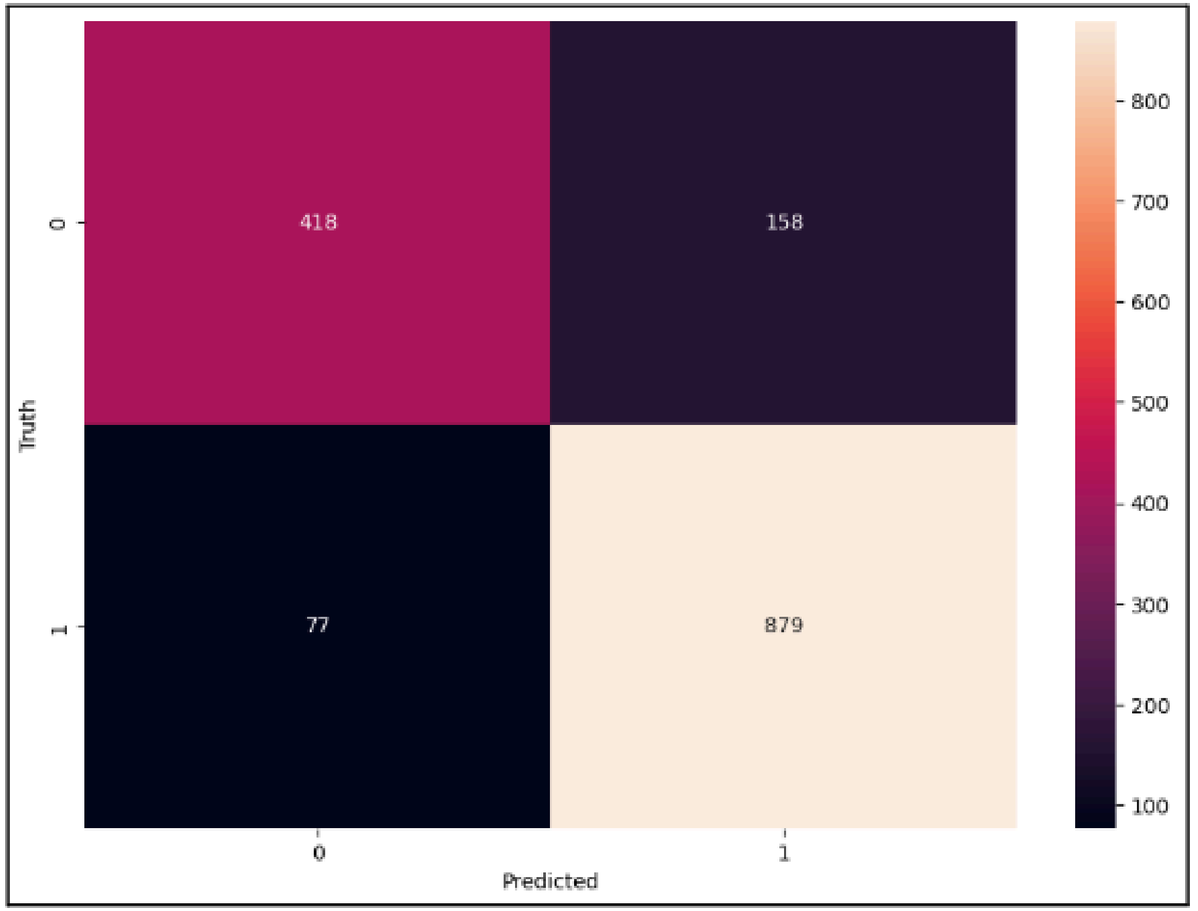
- 'budget', 'runtime','rating_converted', 'director_score', 'company_score','genre_score', 'star_score', 'writer_score', 'month_converted'

Output:

- ‘Successful’ or ‘Not successful’

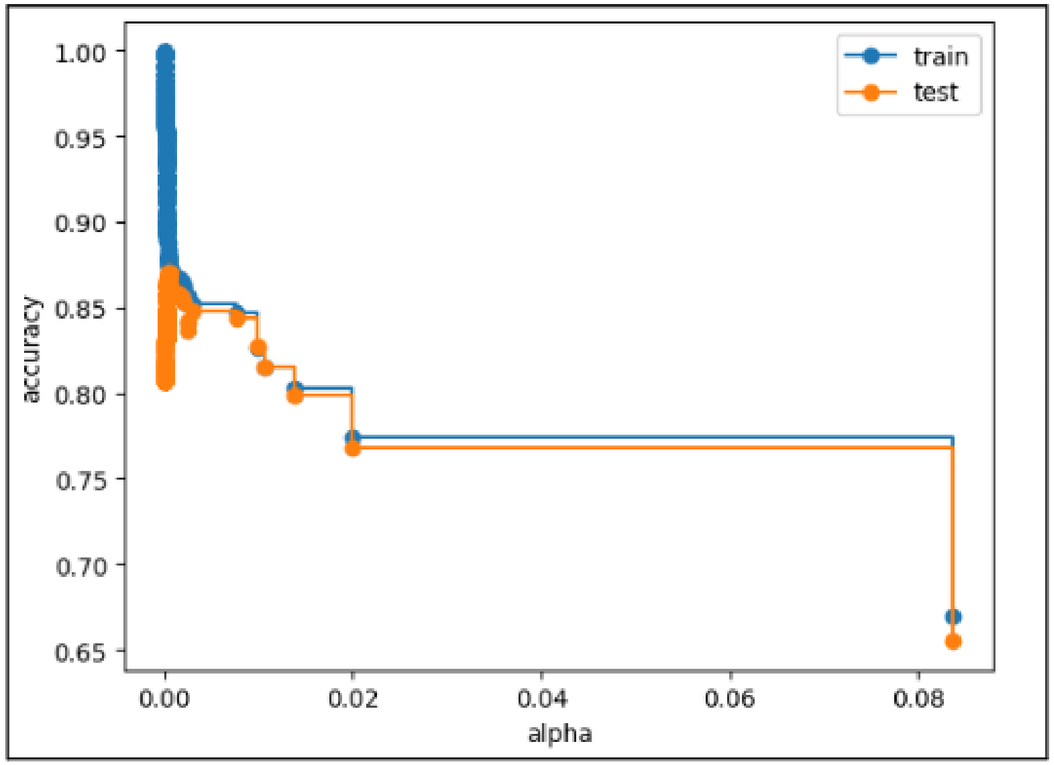
Result:

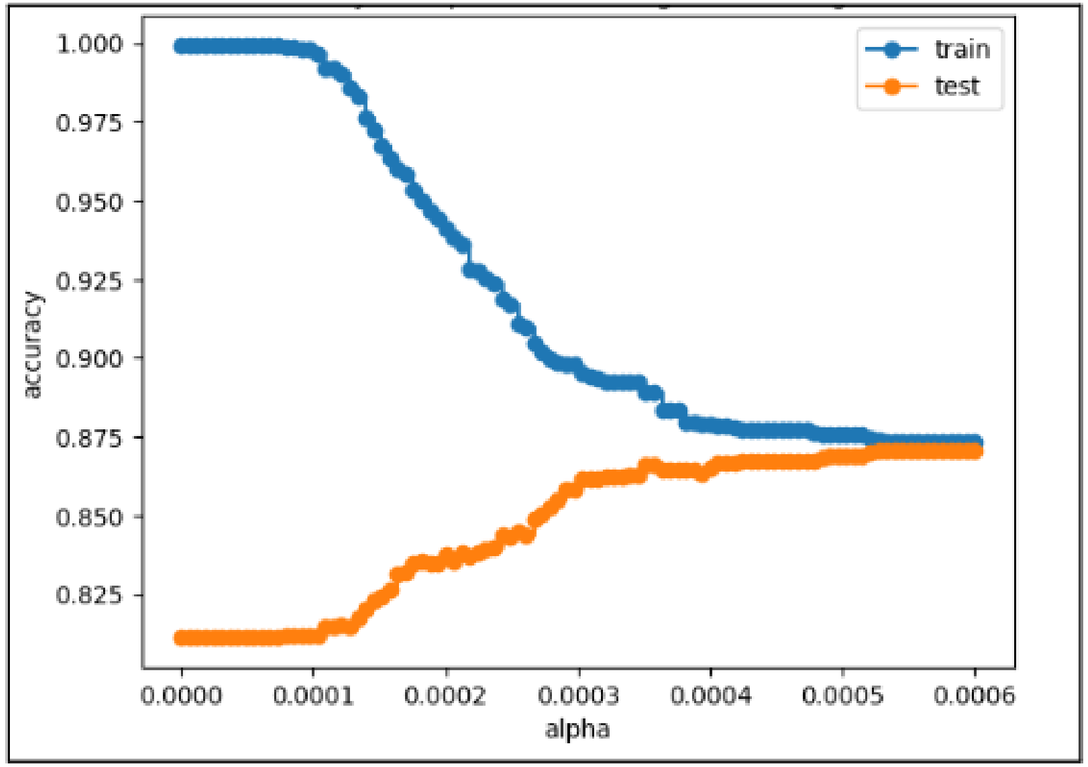
- Accuracy: 84.66%. (~14%)



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.





Accuracy vs alpha for training and testing sets

IMPROVEMENT WITH POST-PRUNING

Input:

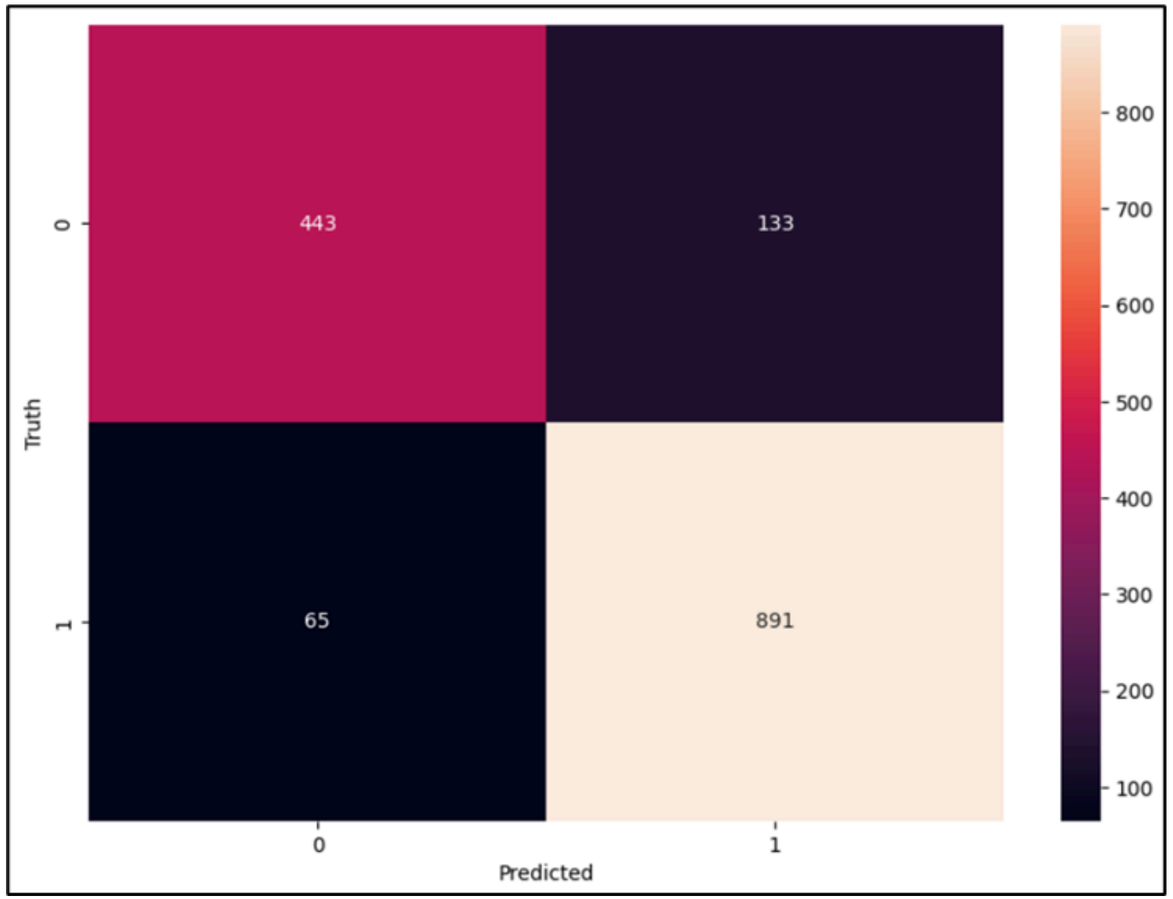
- 'budget', 'runtime','rating_converted', 'director_score', 'company_score','genre_score', 'star_score', 'writer_score', 'month_converted'

Output:

- ‘Successful’ or ‘Not successful’

Result:

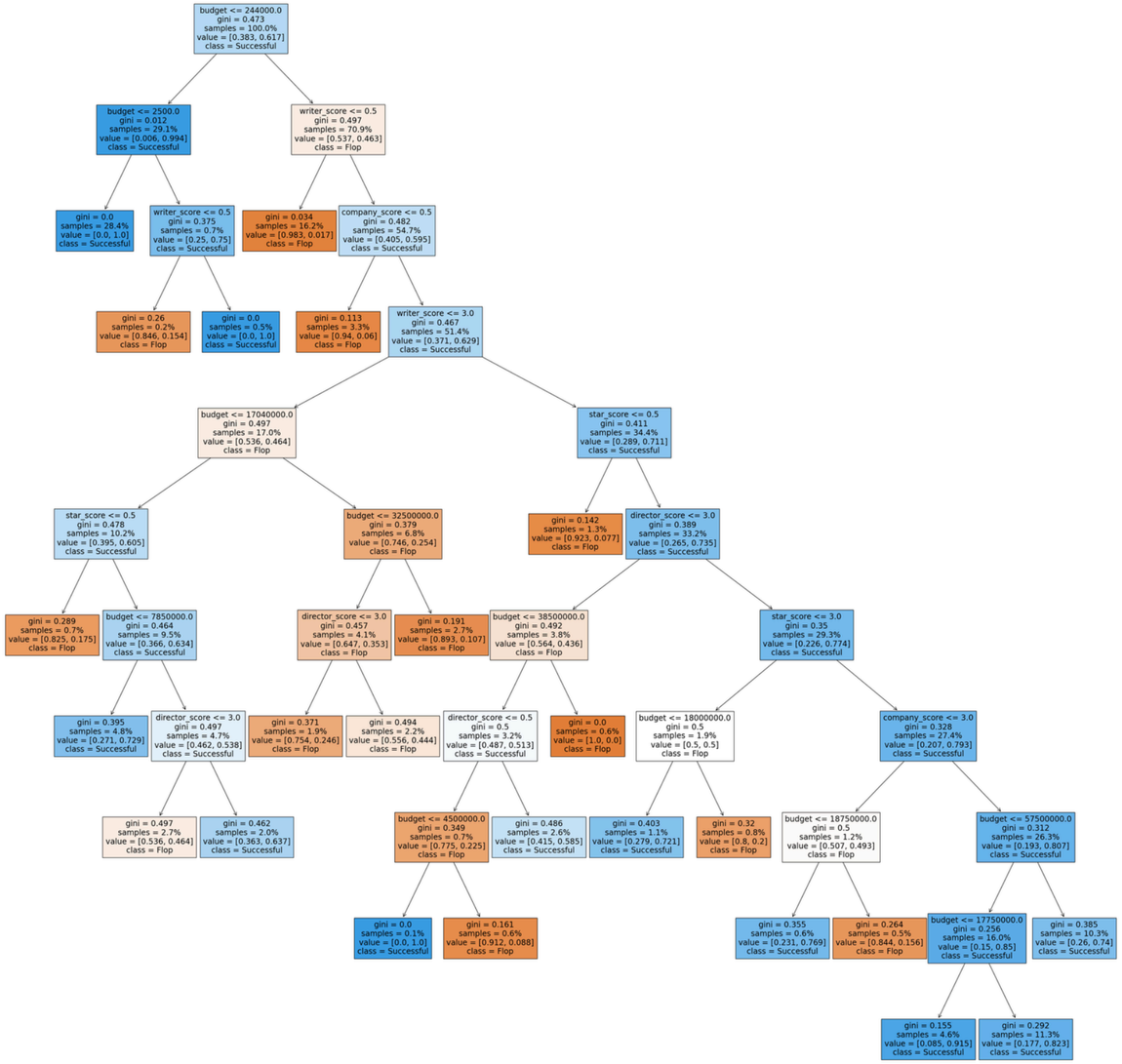
- Accuracy: 87.08%. (~2.5%)



COMPARISION

METHODS	ACCURACY	INCREASE
Decision Tree (without data insight)	70.43%	
Decision Tree (with data insight)	84.66%	~15%
Decision Tree (post-pruning)	87.08%	~2.5%

PLOT DECISION TREE



DEPLOYMENT

Deployment

Streamlit application github link:

https://github.com/IrisPham74/CS331.O21.KHCL.git

REFERENCES

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.html
- https://patshih.luddy.indiana.edu/publications/Gao-MovieSuccess-iConf19.pdf
- https://www.researchgate.net/publication/24072360_Successful_Movies_A_Preliminary_Empirical_Analysis

Assign Work Table

	Huy Hoàng	Tram Anh	Dinh Duc	Nhat Long	Truong Thien
Research and Data Mining	X	X	X	X	X
Implement code demonstration			X	X	X
Prepare Report	X	X	X	X	X
Create Slide	X	X			
Deliver the presentation	X	X			