FILM
INDUSTRY
ANALYSIS
PROPOSAL

MOVIE SUCCESS
ANALYSIS

HTTP://EC2-3-87-168-77.COMPUTE-1.AMAZONAWS.COM:8080/

CONTENT

CONTENT

CONTENT

INTRODUCTION

The project aims to utilize historical film data to analyze and predict the factors that impact a film's success, with the goal of providing valuable insights to the film industry and improving predictions of audience reception and box office performance.

PROBLEM STATEMENT

The project aims to address the film industry's challenge of predicting a film's success by analyzing historical data, identifying influential attributes, and using them to predict performance indicators such as audience reception and box office earnings.

RELATED WORK

Box Office Pro

Box Office Pro is a platform or service that provides information, analysis, and predictions related to box office performance in the film industry.

Netflix

“Streaming Platform” (Netflix) forecasts that help highlight trends in the entertainment industry.

Academic Paper

Academic articles such as “Forecasting Box Office Revenues” (Ruus and Sharma, 2019) and “Predicting Movie Success” (Darapaneni et al.., 2020) provide valuable insights on predicting movie success using metadata, reviews, and machine learning algorithms.

DATASET

This dataset provides metadata for 45,000 movies released before July 2017, from TMDB (Movie Database) and GroupLens. It includes details like cast, crew, plot keywords, budget, revenue, release date, language, production information, and user ratings.

IMDb Dataset provides information about movie features including IMDb Movie ID, Release Year, Certificate, Runtime performance, genre, ratings, description, director, stars, votes, and box office gross. Data. The inclusion of additional details like director and star IMDb IDs enhances the granularity and precision of the dataset

INITIAL PREPROCESSING

Merging of datasets

As the movie dataset of TMDB is scattered and fragmented into genres: each genre was in a separate CSV file. We first merged all TMDb dataset then merged them with IMDb which had the final shape of 1.3 Million rows and 74 columns after dropping a few unwanted columns

Merging of columns

Once we merged the dataset we had a huge number of nan columns that needed to be addressed. We preprocessed genres by merging them then converted them to a list, We also removed few more columns.

Feature Extrapolation

Due to huge number of columns we decided to go for feature extrapolation which was performed with the columns belongs_to_collection, lanaguages_spoken, genres, production_companies, and revenues(we keept for this we used combine_first

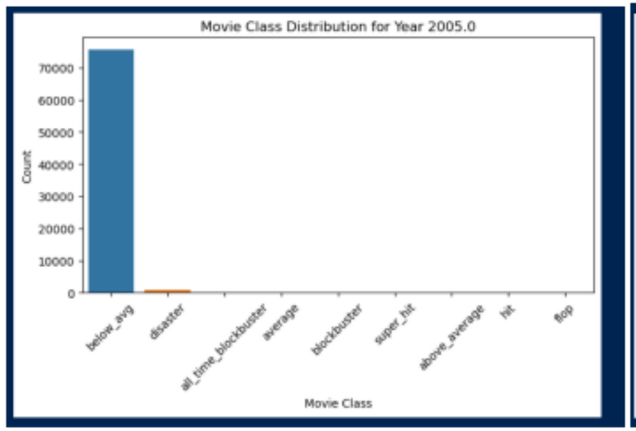
MOVIE CLASSIFICATION

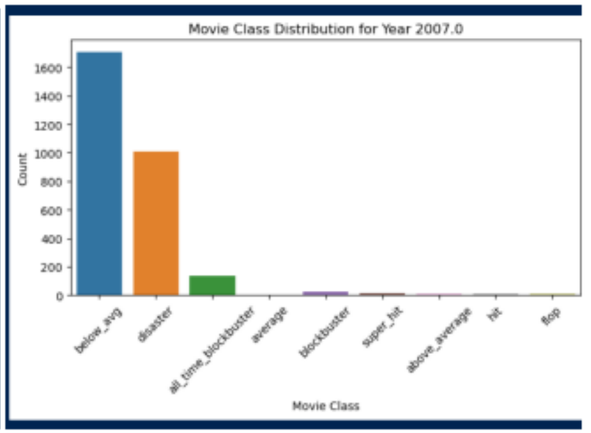
We have 9 movie classes: They are listed below with the gross profit:

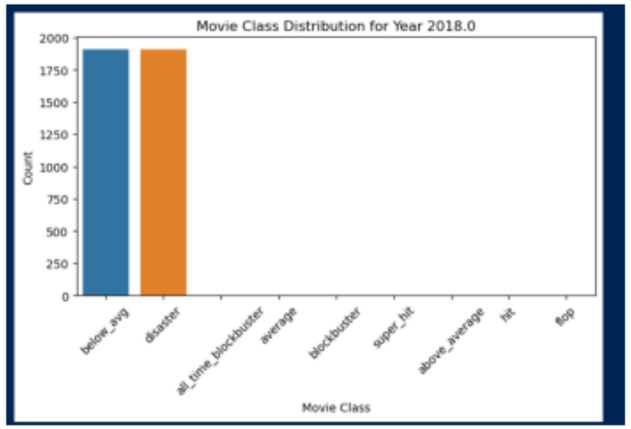
- Profit>125% (All-time blockbuster)
- Profit>75%(Blockbuster)
- Profit>40% upto 75% (Superhit)
- Profit>25% upto 40% (Hit)
- Profit>10% upto 25% (Above Average)
- Profit 0% - 10% (Average)
- Loss of Less than 15& (Below Average)
- Loss of more than 15% but less than 40% (Flop)
- Loss of more than 40% (Disaster)

EXPLORATORY DATA ANALYSIS

What is the movie success distribution for each year?







EXPLORATORY DATA ANALYSIS

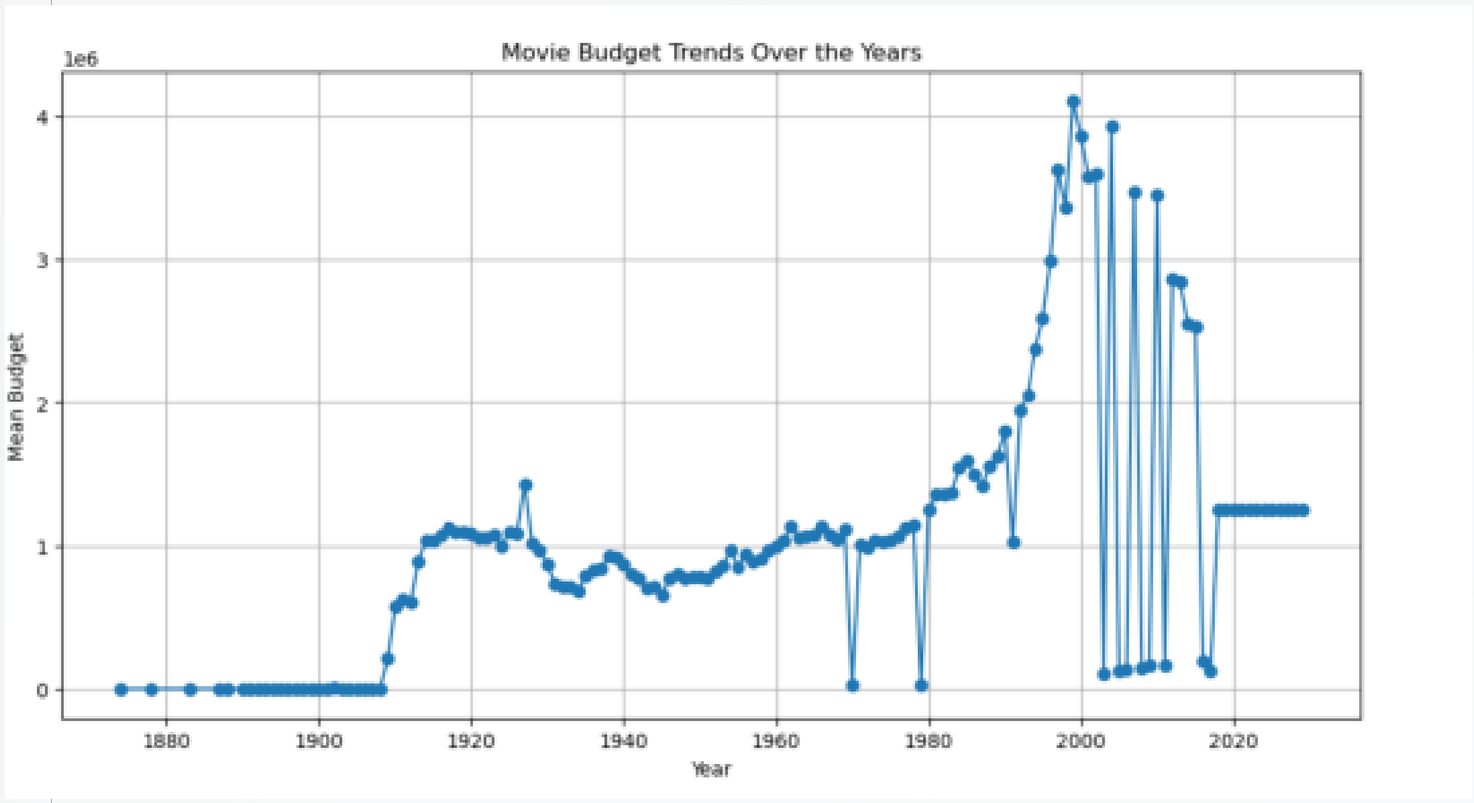
What is the yearly investment trend in movies? n for each year?

The investment across the years seems to grow significantly with the highest in the year around 2000 which decreased around 2020 which might be due to the outbreak of coronavirus. There was also a slight peak in the year around 1925 during major events such as the transition from silent films to "talkies" with synchronized sound, which was a revolutionary development in cinema and the prominent rise of Hollywood.

EXPLORATORY DATA ANALYSIS

What is the yearly investment trend in movies?

n for each year?



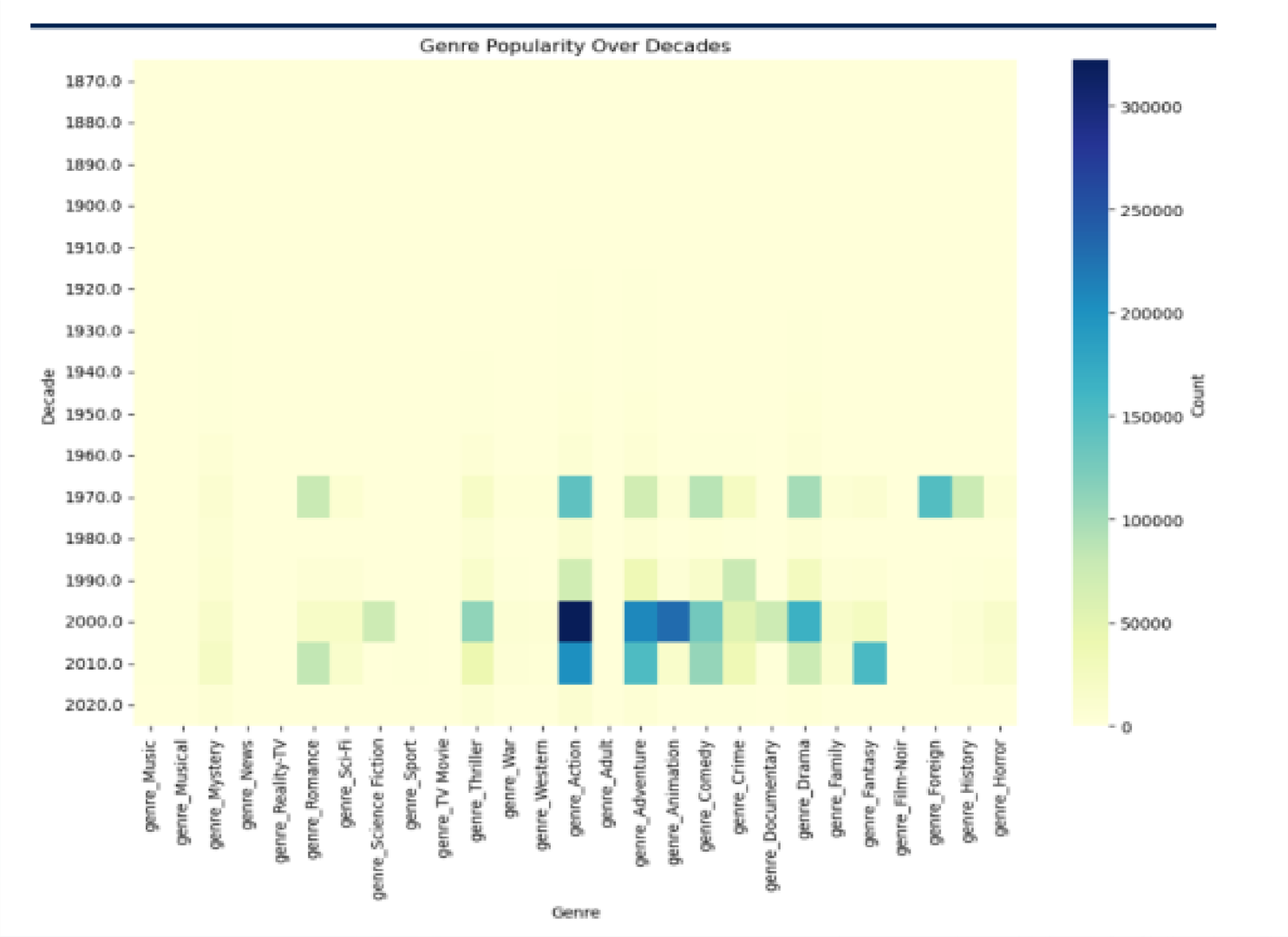
EXPLORATORY DATA ANALYSIS

What is the movie success distribution for each year?

The film industry experienced fluctuations in movie classifications from 1985 to 2012, with most movies falling into the categories "Below Average" and "Disaster." Between 1994 and 1997, a majority of movies underperformed, while between 2007 and 2010, movies were well-observed across all classifications. Between 2012 and 2016, a mix of "Below Average," "Disaster," and "All-Time Blockbuster" movies was seen, with a balanced distribution between "Below Average" and "Disaster" from 2018 to 2023. Most movies grossed below average or disasters, with few all-time blockbusters and very few flops, hits, and averages. Some of the notable outputs are given below:

EXPLORATORY DATA ANALYSIS

Which genre is popular each year?



EXPLORATORY DATA ANALYSIS

Which genre is popular each year?

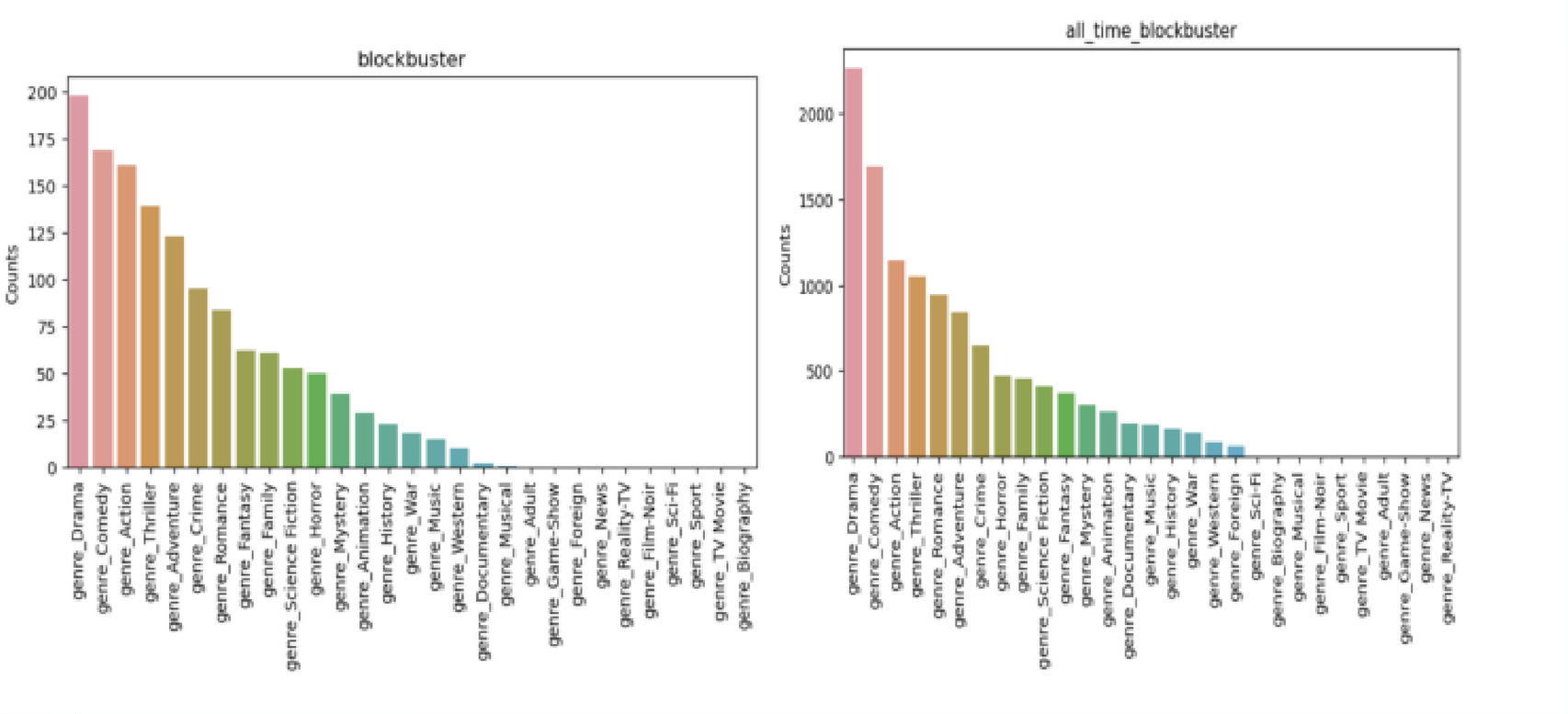
There seems to be a growing trend of Action movies from 1970 to 2010 onwards with the highest in 2000 when many Marvel and DC movies were released. The adventure and animation category also seems to gaining momentum whereas the trend of Foreign and history seems to have decreased.

The top genres appear to be Action, Adventure, Animation, Fantasy, and Comedy

In summary, this project aims to provide a comprehensive analysis of the film industry, providing insight into the factors that determine a film's success. By leveraging two rich data sets and established methods, we hope to open up a wealth of knowledge about the complexities of the film industry.

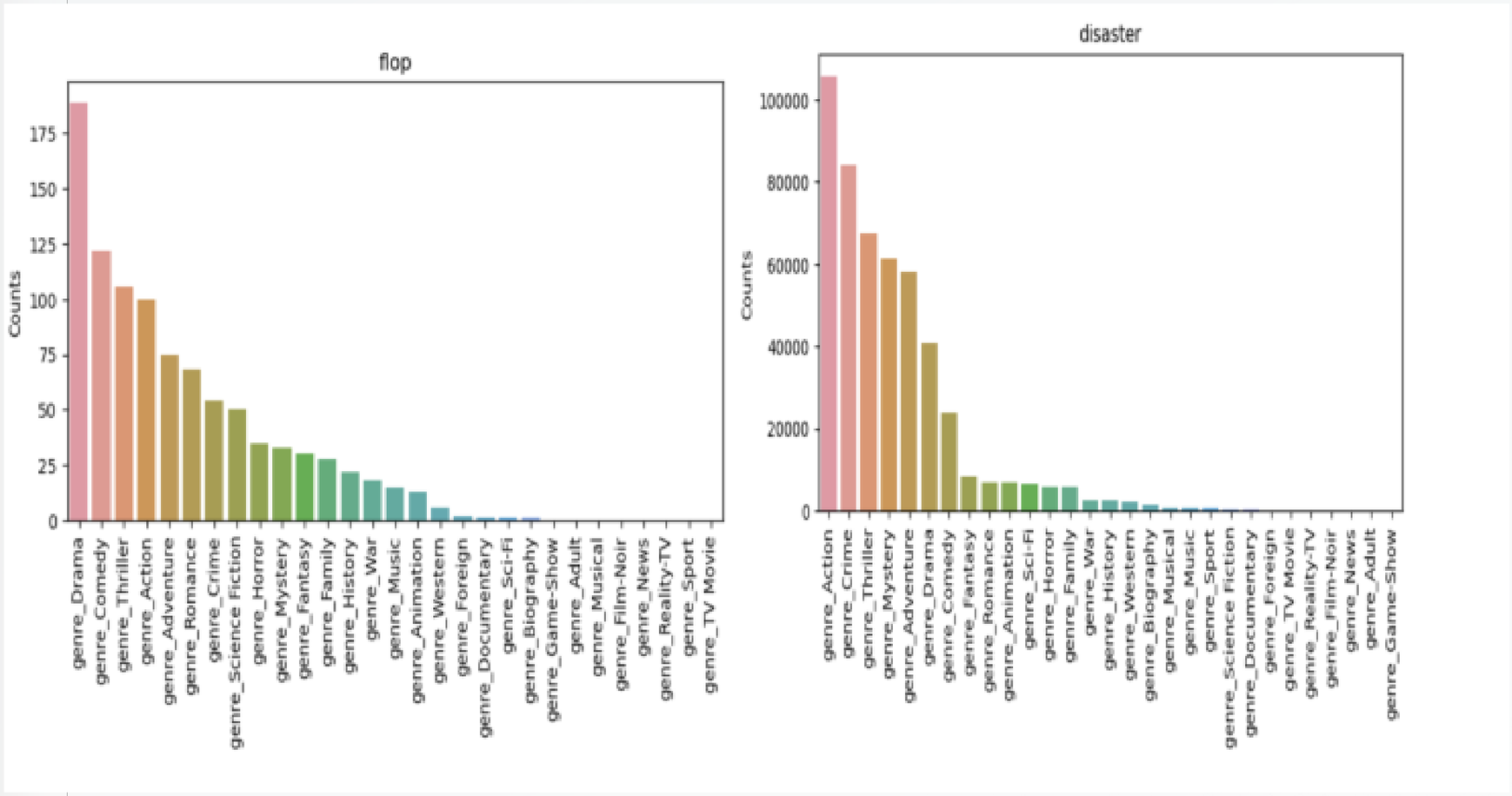
EXPLORATORY DATA ANALYSIS

How does the movie genre contribute to movie classification?



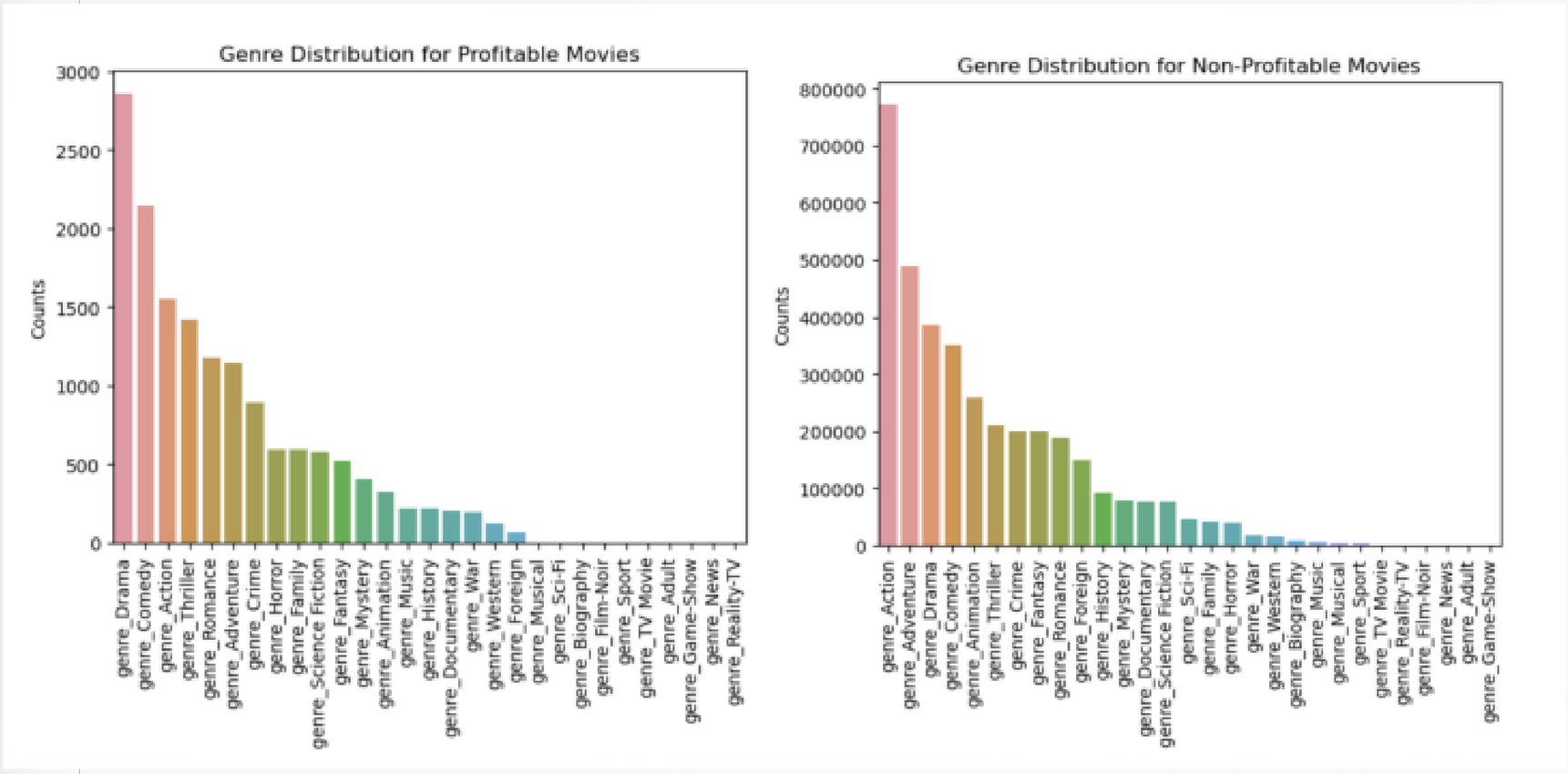
EXPLORATORY DATA ANALYSIS

How does the movie genre contribute to movie classification?



EXPLORATORY DATA ANALYSIS

How does the movie genre contribute to movie classification?



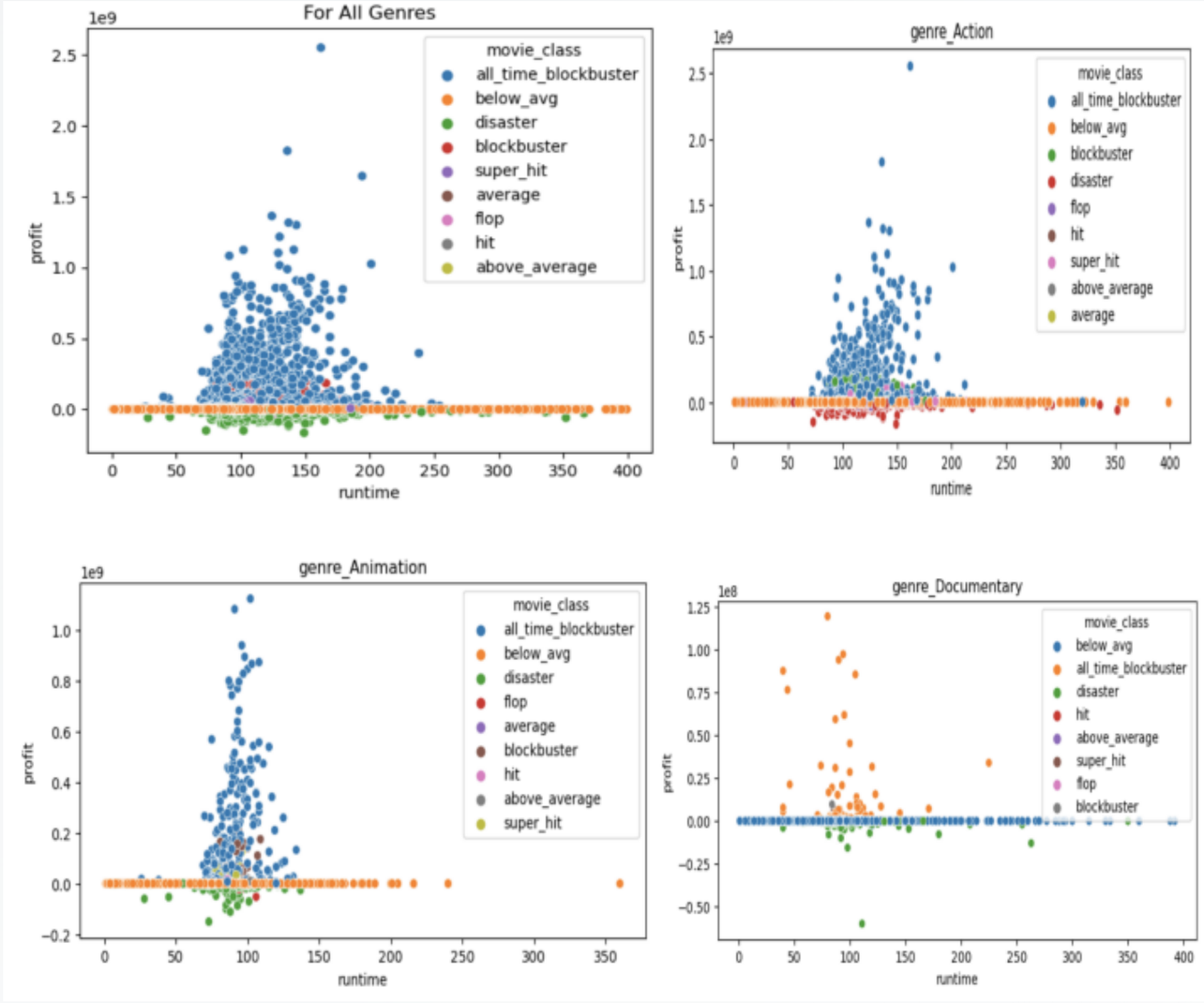
EXPLORATORY DATA ANALYSIS

How does the movie genre contribute to movie classification?

We have also examined the genre distribution for movie categories to assess the influence of genres on the success of a film. Given the extensive range of movie categories available, we will exclusively present the data for the four extreme categories: "all-time blockbuster," "blockbuster," "flop," and "disaster."

EXPLORATORY DATA ANALYSIS

Does the movie runtime contribute to profit?



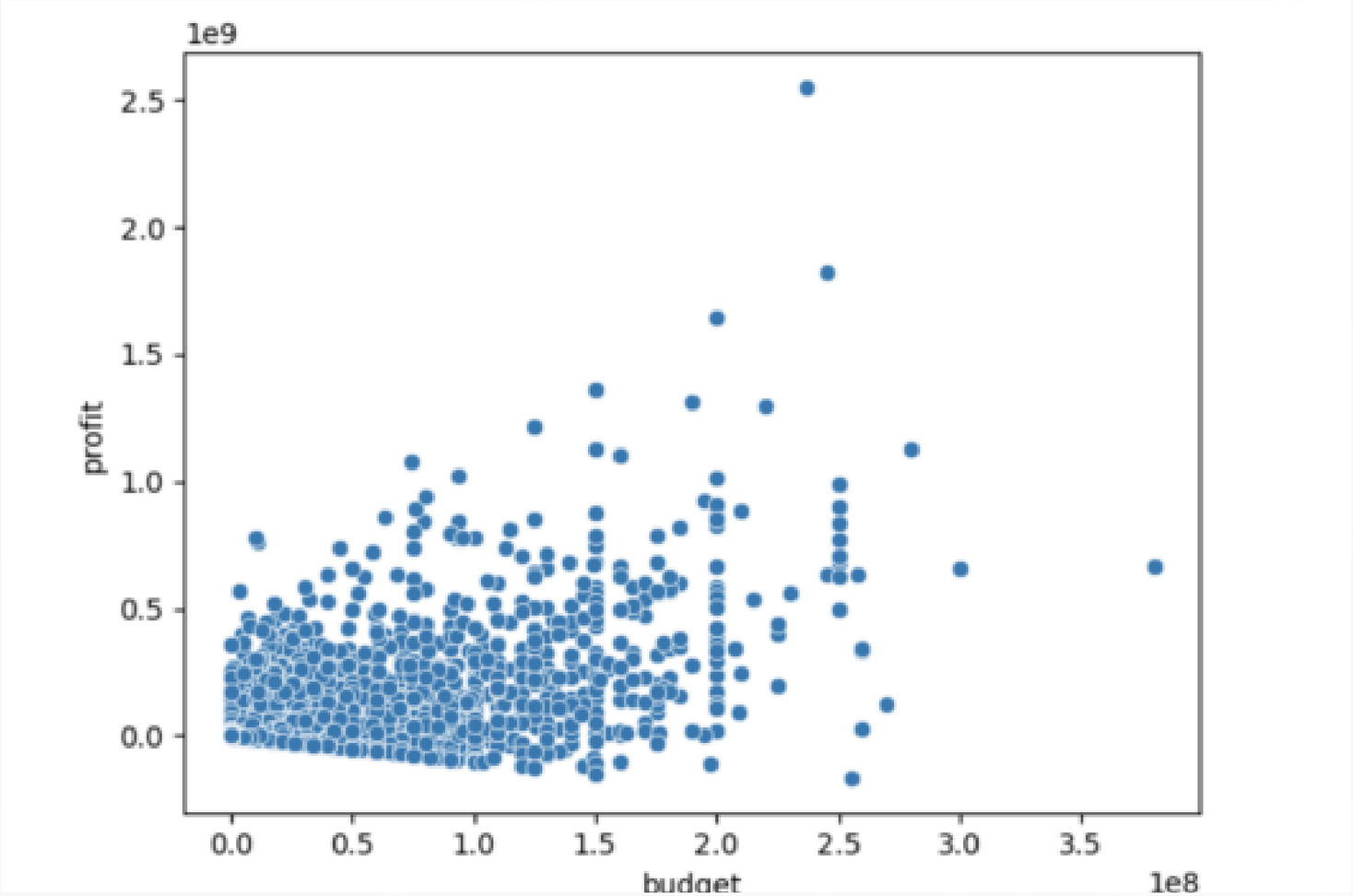
EXPLORATORY DATA ANALYSIS

Does the movie runtime contribute to profit?

We also tried to understand the relationship between the budget allocated and the profit generated. While it is acknowledged that budget is not the primary factor influencing a film's success, it is clear that the profits of some films will increase as their budgets increase. Therefore, it can be inferred that the budget size contributes partly to the overall profit.

EXPLORATORY DATA ANALYSIS

How does the movie budget contribute to movie profit?



EXPLORATORY DATA ANALYSIS

How does the movie budget contribute to movie profit?

We have endeavored to calculate the budget associated with each film course. Based on the graphical presentation, it's clear that the majority of films classified as "below average” and “disasters" actually had relatively low budgets. This observation is consistent with the general view that producing a high-quality film often requires a significant financial investment.

EXPLORATORY DATA ANALYSIS

How does the movie budget contribute to movie classification?

	movie_class	budget_class	movie_counts
0	above_average	high	52
1	above_average	low	20
2	above_average	mid	112
3	all_time_blockbuster	high	603
4	all_time_blockbuster	low	2730
5	all_time_blockbuster	mid	1257
6	average	high	35
7	average	low	19
8	average	mid	64
9	below_avg	high	46
10	below_avg	low	834393
11	below_avg	mid	116
12	blockbuster	high	165
13	blockbuster	low	63
14	blockbuster	mid	234

EXPLORATORY DATA ANALYSIS

How does the movie budget contribute to movie success?

16	disaster	low	136206
17	disaster	mid	1949
18	flop	high	78
19	flop	low	49
20	flop	mid	214
21	hit	high	36
22	hit	low	32
23	hit	mid	83
24	super_hit	high	108
25	super_hit	low	48
26	super_hit	mid	178

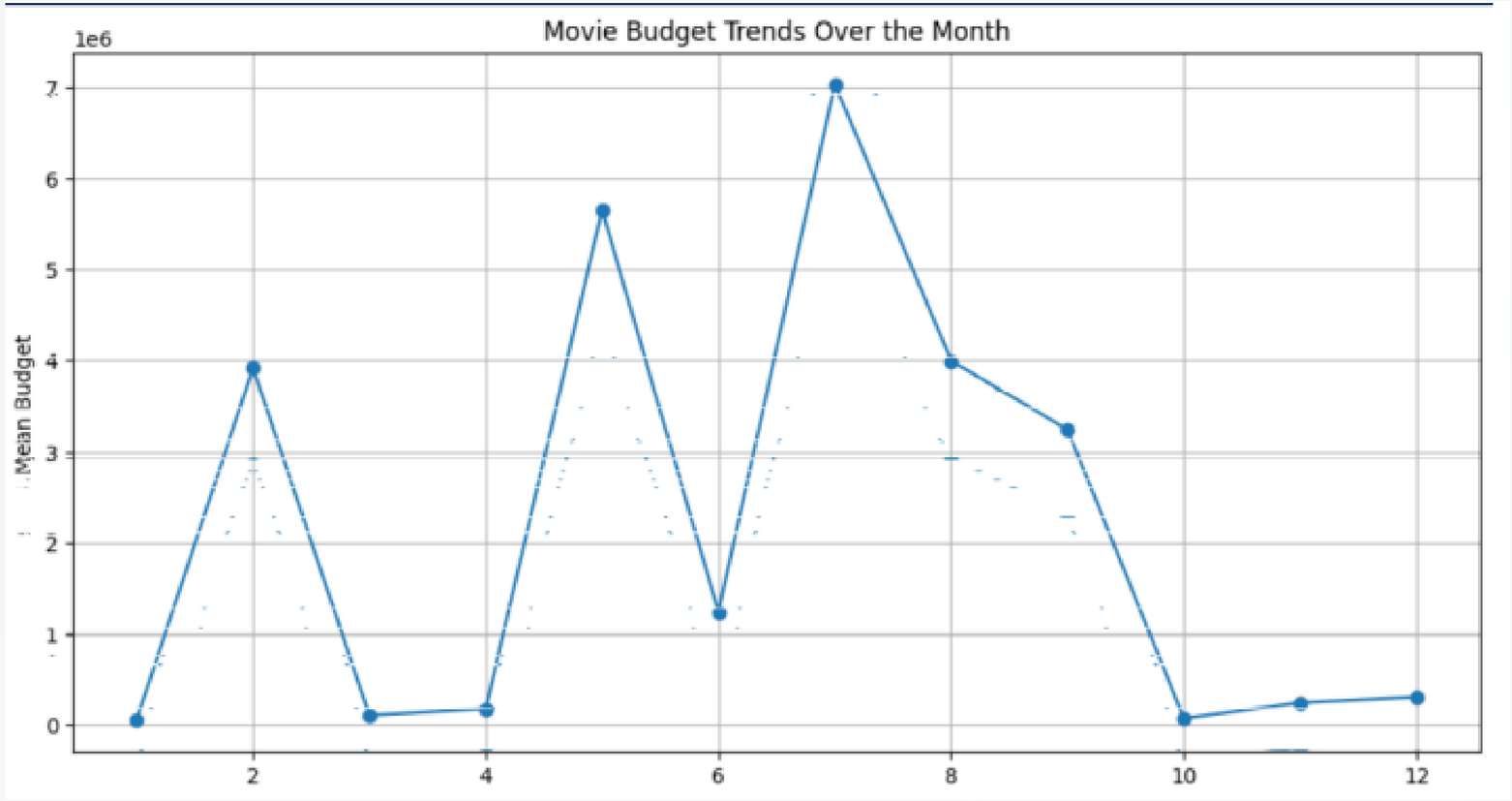
EXPLORATORY DATA ANALYSIS

How does the movie budget contribute to movie success?

According to the plot above, we can observe the movie budget trends through the months. The movie budget of the movies released in July seemed to have the biggest budget. While the ones in January seemed to have the least amount of budget(appears to be 0). This might be true because in January many people go on vacations and do not visit the movies as much while in July most schools and colleges have leisure time so more big-budget releases are targeted then.

EXPLORATORY DATA ANALYSIS

How is the movie budget through the months?



EXPLORATORY DATA ANALYSIS

How is the movie budget through the months?

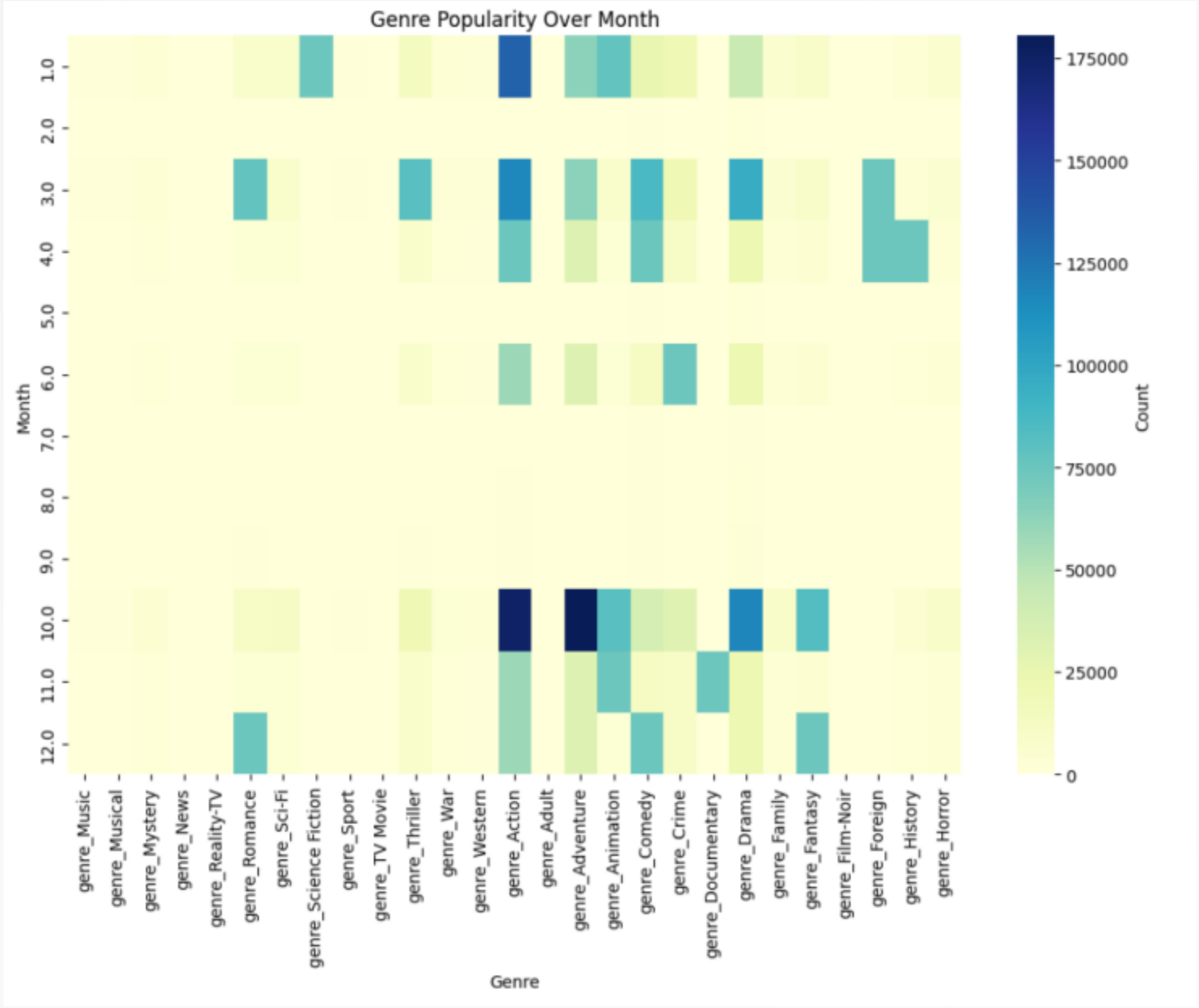
Next, we create a scatter plot between movie runtime and profit to see the relationship between movie runtimes and profits they made. We create scatter plots for each genre.

Deductions

- Overall, a movie's runtime should be within a certain range.
- For some genres, the runtimes are highly related to the profit.

EXPLORATORY DATA ANALYSIS

Which genre is popular for each month?



EXPLORATORY DATA ANALYSIS

Which genre is popular for each month?

There seems to be a specific movie genre that is popular for a particular month. Action along with Adventure genres seems to be the most popular, especially for October. Dramas are also popular in October. Action movies are also popular for January and March while Romance movies are the most popular for January. Hence, the specific movie genres should target these release months (for example: Action movies should be targeted for October release along with Adventure.

This analysis provides ground for understanding the pattern between movie release months based on genre and their success probabilities.

FEATURE ENGINEERING

01 Computation of Star Power

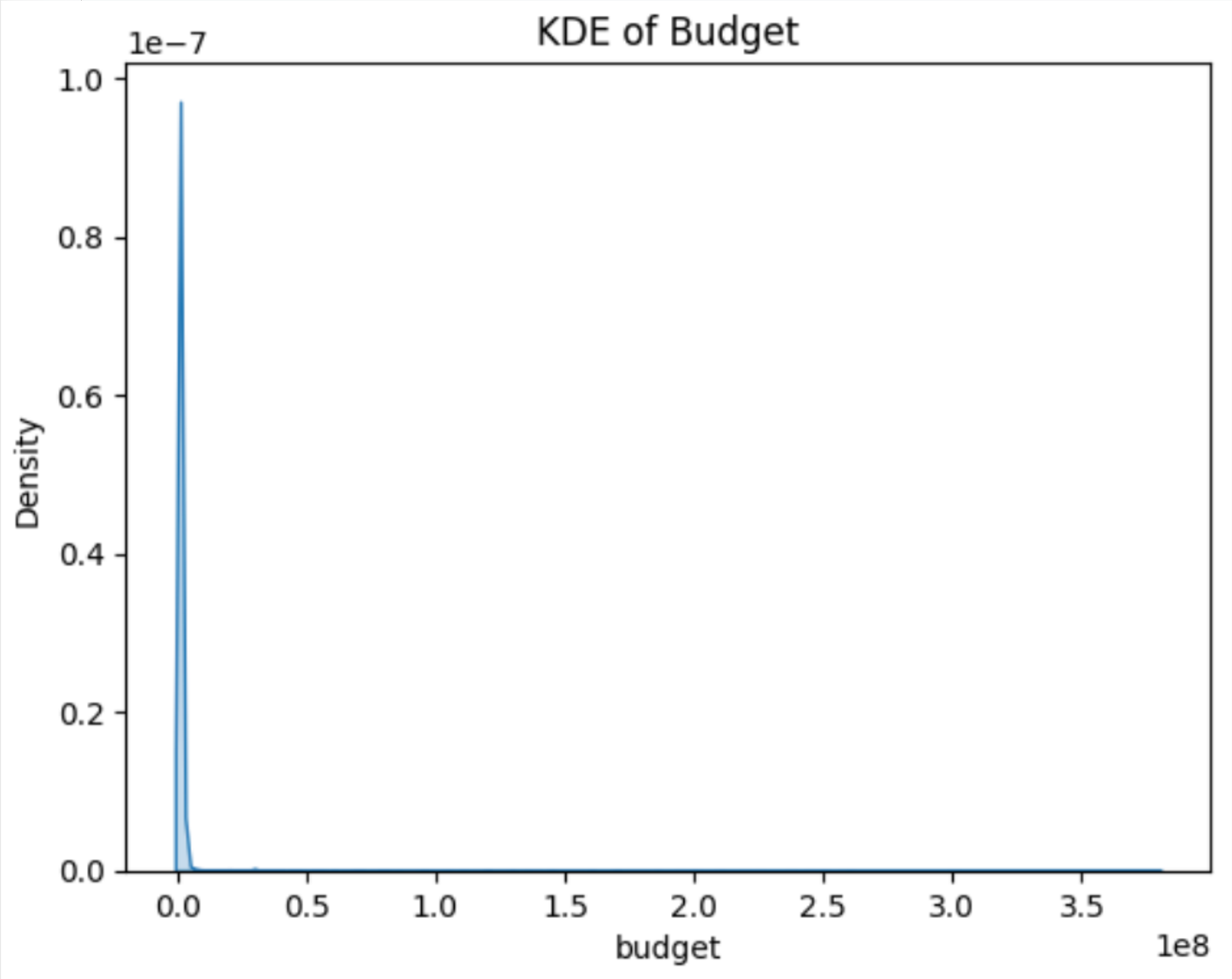
02 Classification of movie gross collection based on the reference

03 Extraction of year and month from Release Date

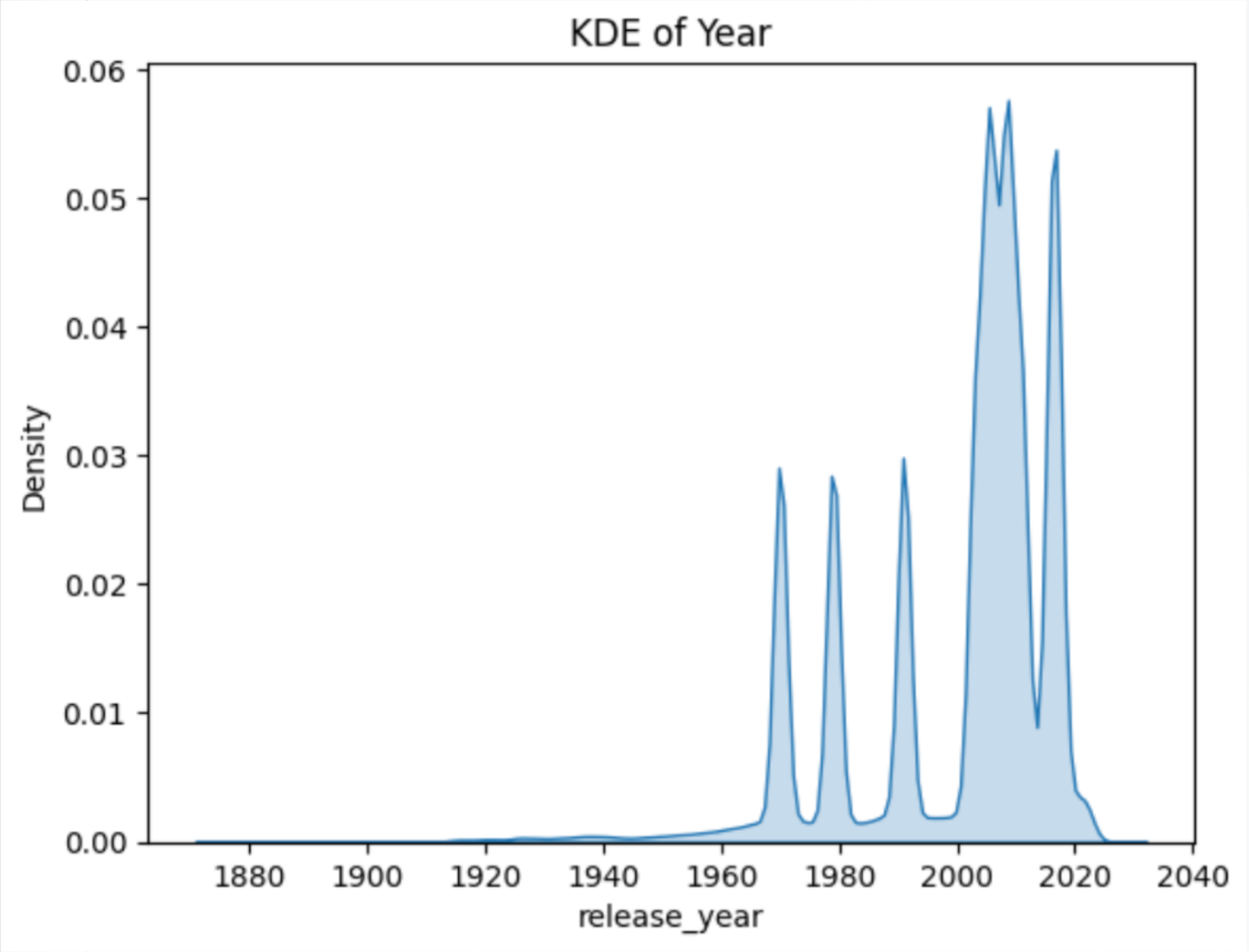
DATA PRE-PROCESSING AND FEATURE SELECTION

Based on our EDA we have the following columns: 'budget', 'release_year', 'release_month',' runtime',' certificate','star_power', and 'director_power' as prominent in the prediction which is set as X. We also set our label [‘movie_class’] as Y. We then split the dataset into X_train, X_test, y_train, and y_test in the ratio of 0.8:0.2

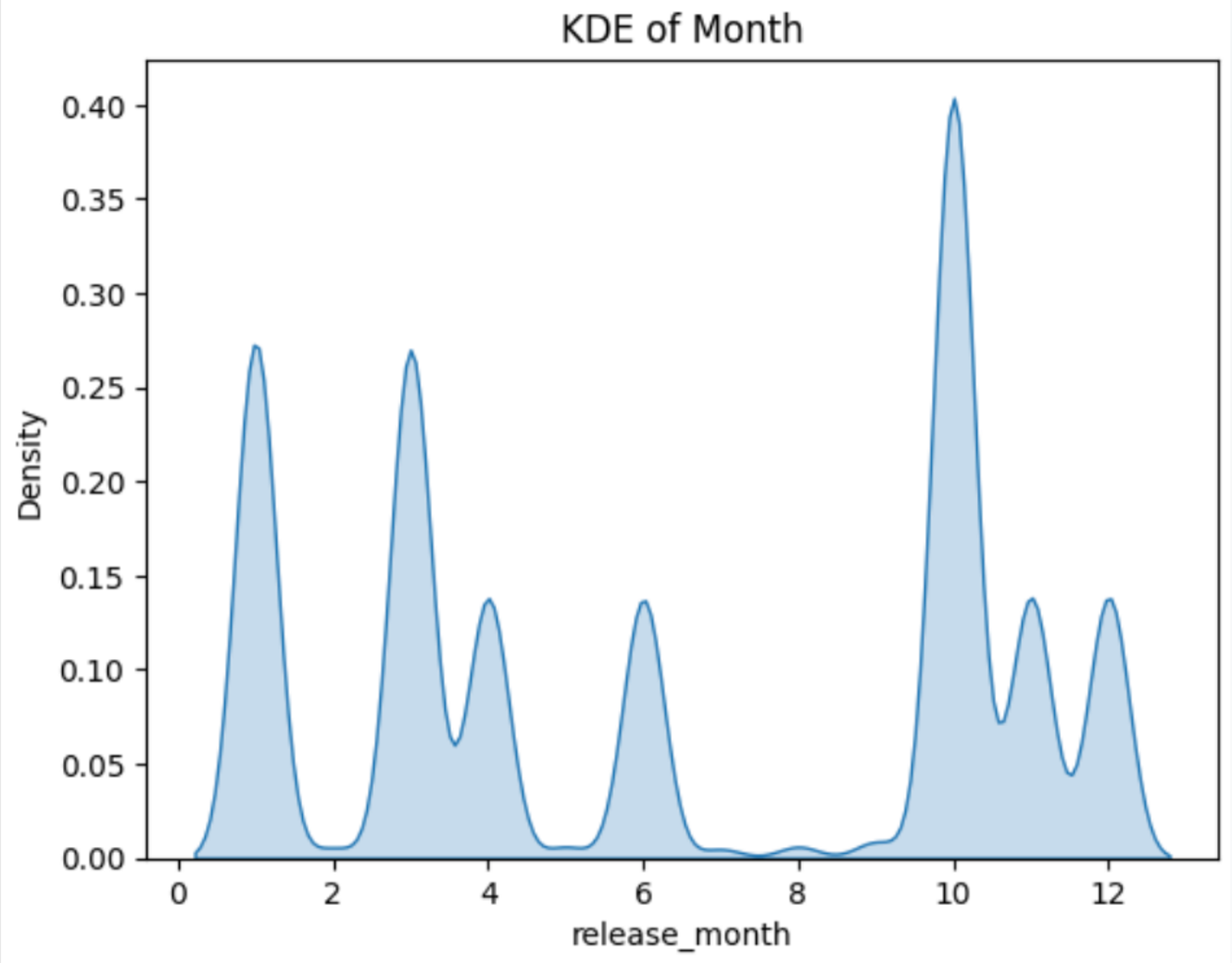
DISTRIBUTION OF BUDGET



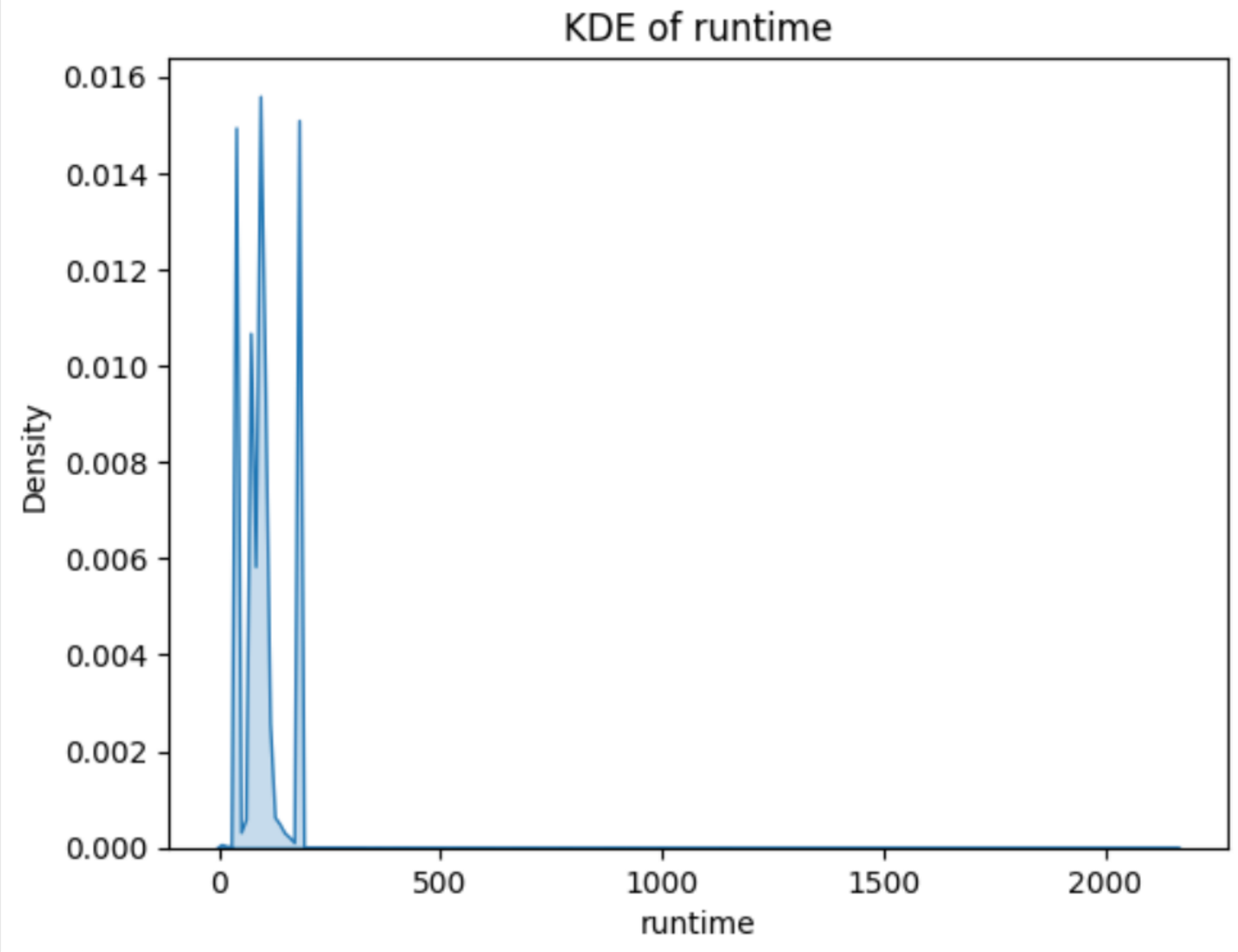
DISTRIBUTION OF RELEASED YEAR



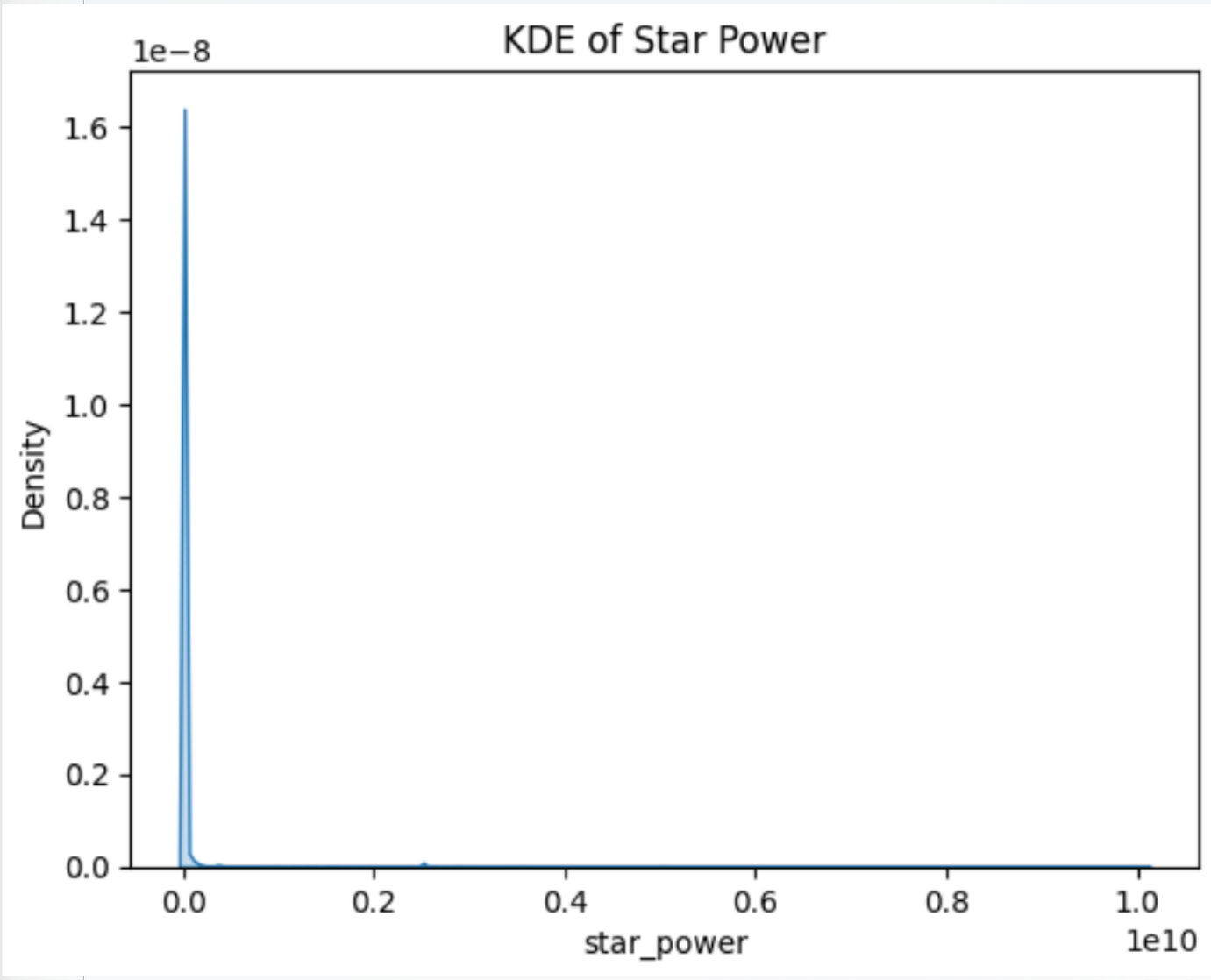
DISTRIBUTION OF RELEASED MONTH



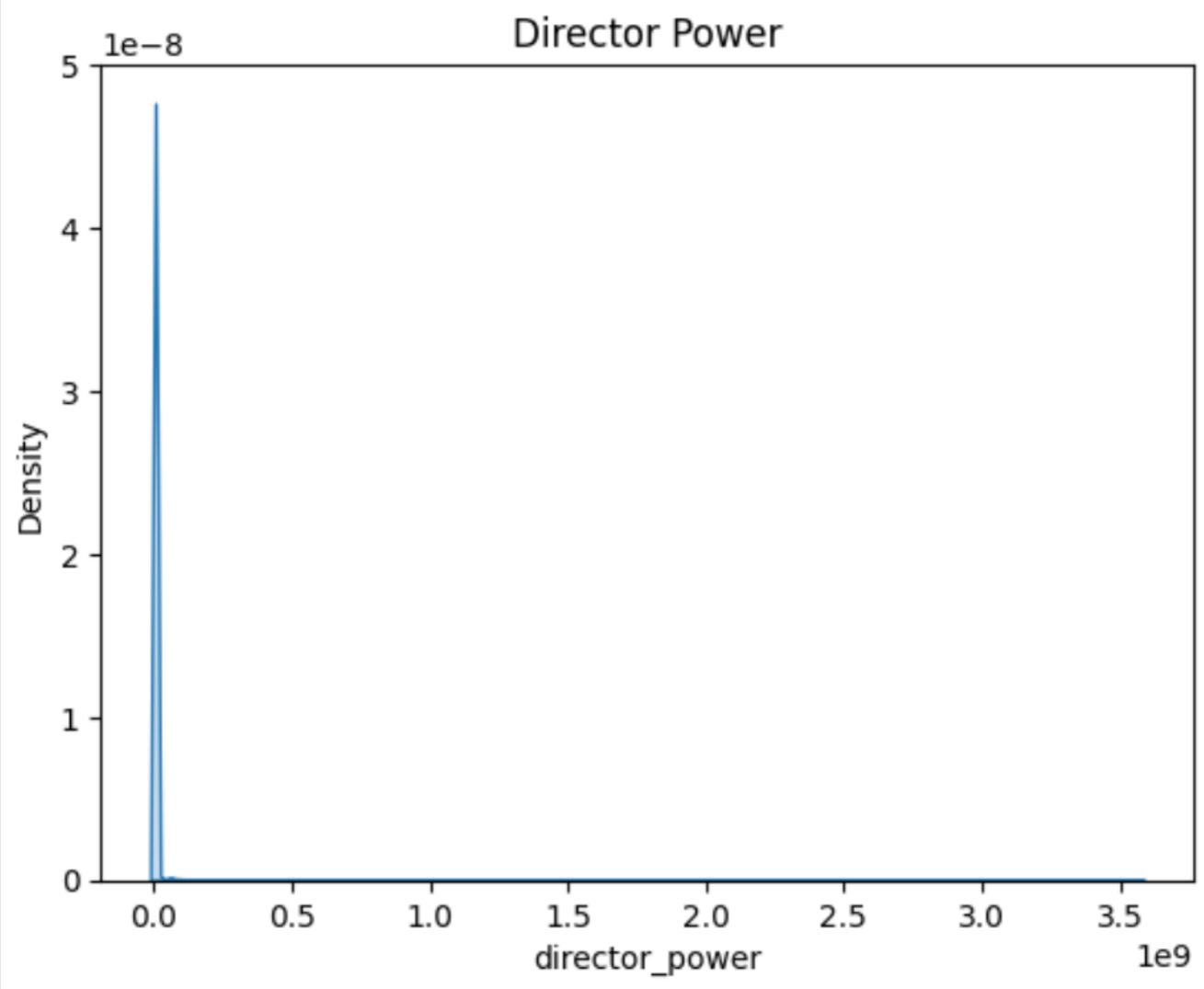
DISTRIBUTION OF RELEASED RUNTIME



DISTRIBUTION OF STAR POWER



DISTRIBUTION OF DIRECTOR POWER



DATA PRE-PROCESSING AND FEATURE SELECTION

- Analysis of KDE plot to understand the imputation and scaling requirements of the columns

- For numeric transformers such as 'budget', 'director_power', and 'star_power,' an Iterative Imputer was chosen.

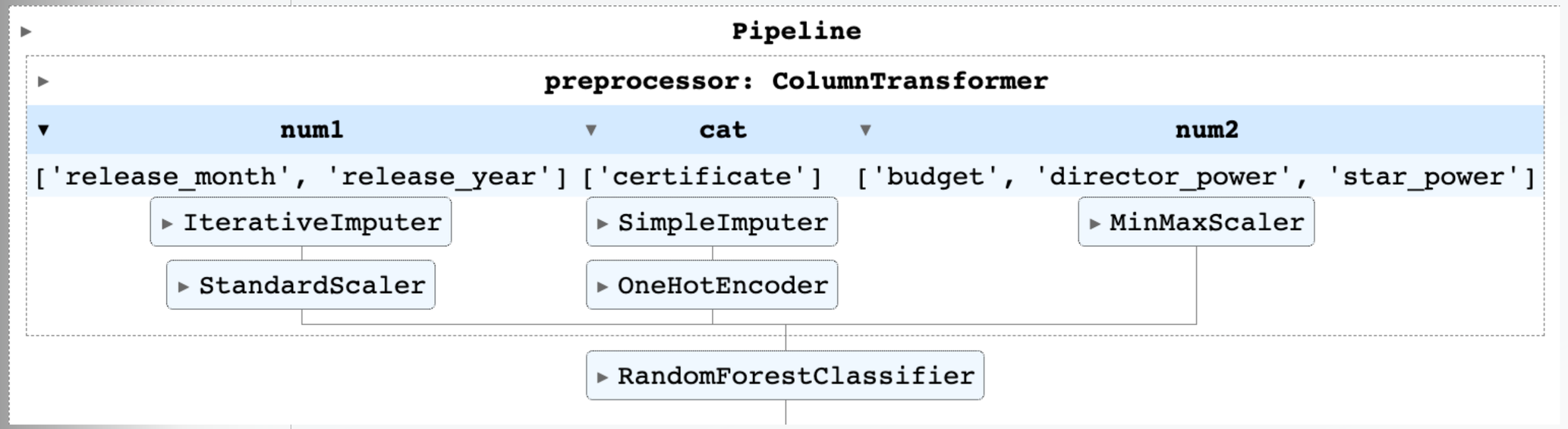
- Distribution of 'budget,' 'director_power,' and 'star_power' was skewed, making Min-Max scaling suitable

- For the columns 'release_month' and 'release_year' the Standard Scaler was chosen
- Addressing null values in categorical columns, specifically 'certificate,' and "Not Rated" was used to fill in missing values.

- Additionally One-hot encoding was then applied to this column

- This was all fitted in a pipileine named preprocessor to prevent data leakage

OUR PRE-PROCESSING PIPELINE



MODELING

Model Selection

Algorithm Name	Accuracy	F1 Macro Score	F1 Weighted Score
LogisticRegression	0.88	0.14	0.84
RandomForestClassifier	0.993 (best)	0.297 (best)	0.992 (best)
KNeighbour Classifier	0.989	0.268	0.988
Gradient Boosting Classifier	0.991	0.245	0.990

MODELING

Hyperparameter Tuning

```
param_grid = {
    'classifier__n_estimators': [50, 100, 200],
    'classifier__max_depth': [None, 10, 20],
    'classifier__min_samples_split': [2, 5, 10],
    'classifier__min_samples_leaf': [1, 2, 4],
}
```

MODELING

Hyperparameter Tuning

```
classifier_params = {
    'n_estimators': 200,
    'max_depth': None,
    'min_samples_split': 10,
    'min_samples_leaf': 2,
}
```

MODELING

ANN

We also tried defining, compiling, and training a neural network model using the **Keras** library for a classification task. The classification task involves predicting one of nine classes. For which a sequential neural network model is defined with several dense layers and a softmax output layer was defined. The model was compiled using the **Adam optimizer**, categorical cross-entropy loss, and additional metrics such as accuracy and TensorFlow's recall metric. The training involved 100 epochs with a batch size of 204800.

RESULTS AND DISCUSSION

ANN

	precision	recall	f1-score	support
all_time_blockbuster	0.02	0.11	0.04	37
blockbuster	0.08	0.51	0.14	918
super_hit	0.00	0.00	0.00	24
hit	1.00	0.97	0.98	166912
above_average	0.04	0.18	0.07	92
average	1.00	0.97	0.98	27677
below_avg	0.05	0.18	0.07	68
flop	0.01	0.07	0.01	30
disaster	0.03	0.18	0.05	67
accuracy			0.96	195825
macro avg	0.25	0.35	0.26	195825
weighted avg	0.99	0.96	0.98	195825

RESULTS AND DISCUSSION

Random Forest

	precision	recall	f1-score	support
all_time_blockbuster	0.00	0.00	0.00	37
blockbuster	0.44	0.24	0.31	918
super_hit	0.00	0.00	0.00	24
hit	1.00	1.00	1.00	166912
above_average	0.00	0.00	0.00	92
average	0.98	0.99	0.99	27677
below_avg	1.00	0.01	0.03	68
flop	0.00	0.00	0.00	30
disaster	0.00	0.00	0.00	67
accuracy			0.99	195825
macro avg	0.38	0.25	0.26	195825
weighted avg	0.99	0.99	0.99	195825

RESULTS AND DISCUSSION

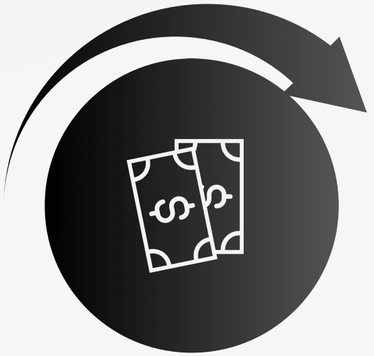
ANN

Class Imbalance

The imbalanced distribution of classes in the dataset can adversely affect the training of ANNs. The model might prioritize accuracy by focusing on the majority class, neglecting the minority classes

Limited Representational Capacity:

The chosen architecture and size of the ANN might not have sufficient representational capacity to effectively learn complex patterns in minority classes, resulting in poor performance.



Complexity and Non-Linearity

ANNs, especially deep architectures, are highly flexible and can capture intricate relationships within the data. However, this flexibility may lead to overfitting, particularly when dealing with imbalanced classes or noisy features.

Sensitivity to Initialization and Hyperparameter

ANNs are sensitive to the choice of initial weights and hyperparameters. If not properly tuned, the model may struggle to generalize well to minority classes, leading to lower precision, recall, and F1 scores

RESULTS AND DISCUSSION

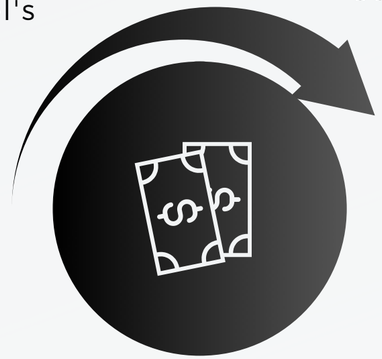
RANDOM FOREST

Handling Imbalanced Data

Random Forests inherently handle imbalanced datasets by considering different subsets of features and samples in each tree. This can mitigate the impact of imbalanced classes on the model's performance.

Ensemble Advantage

Random Forests are ensemble methods that aggregate predictions from multiple decision trees. This ensemble approach often provides robustness against overfitting and tends to generalize well to different classes.



Class Weighting:

The use of class weights, particularly in the Random Forest model, helps address class imbalance by assigning higher weights to minority classes during training. This allows the model to pay more attention to under-represented classes.

Tree-Based Structure

The tree-based structure of Random Forests allows them to capture non-linear relationships in the data effectively. Each decision tree contributes to the overall decision, helping the model make nuanced predictions

RECOMMENDATIONS

Hyperparameter Tuning for ANN

Exploring techniques such as data augmentation or resampling to balance class distribution, aiding both models in learning patterns from minority classes.

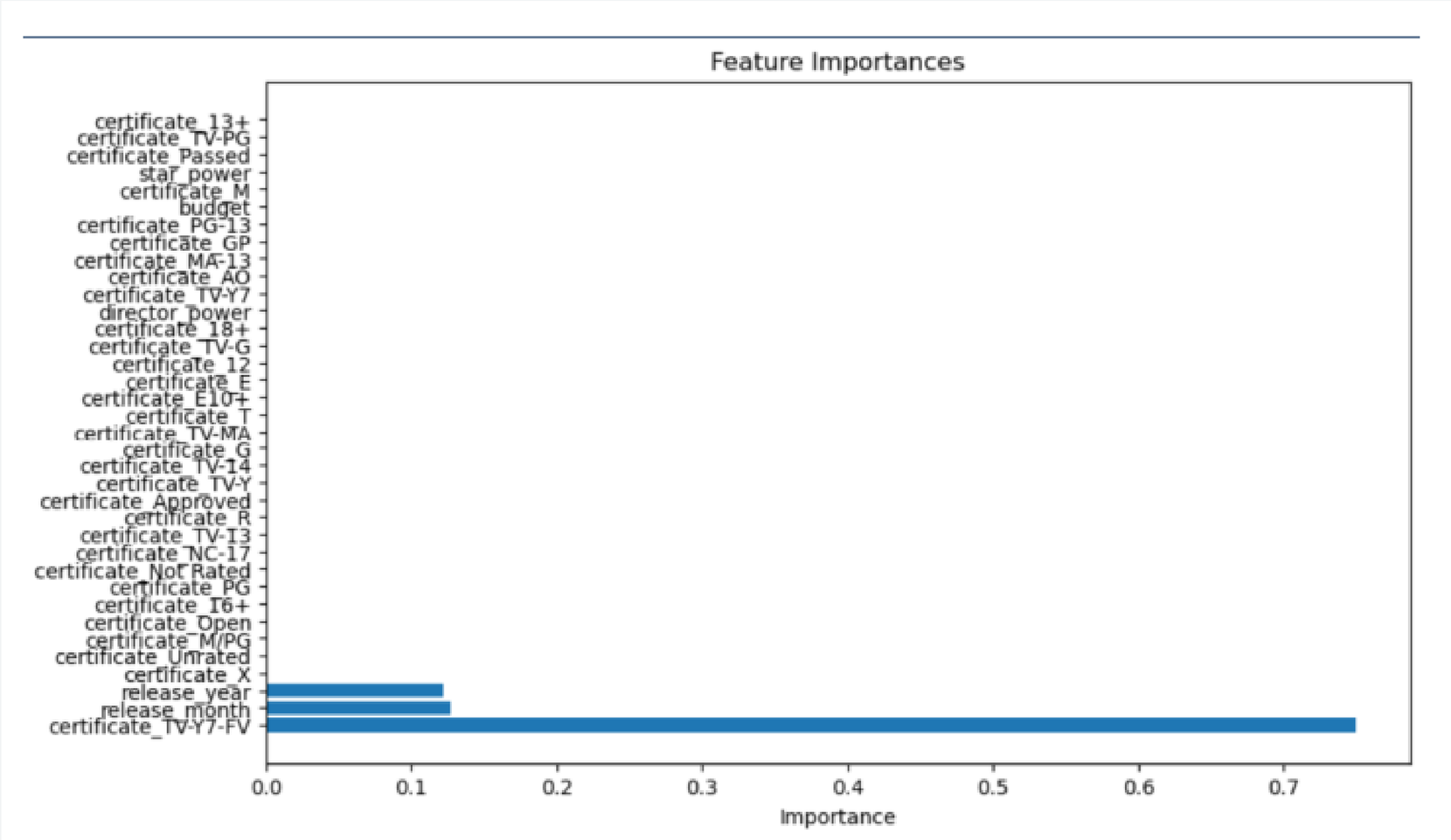
Considering combining predictions from both models using ensemble strategies. This can leverage the strengths of each model and potentially enhance overall performance.

RECOMMENDATIONS

Experiment with adjusting classification thresholds to balance precision and recall based on the specific goals of the classification task.

Cross-Validation and Robust Evaluation

FEATURE IMPORTANCE



FUTURE WORK

Recommendation System for Genre-Specific Features

NLP for Movie Plot Analysis

Data Collection for Rare Classes

FUTURE WORK

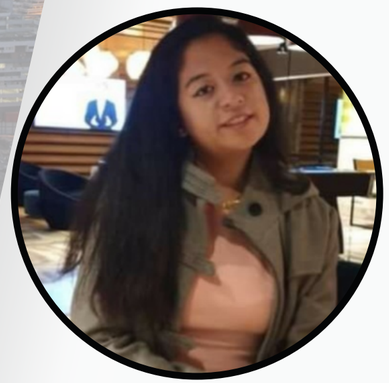
Temporal Analysis for Release Month

Collaboration with Industry Experts

User Feedback and Iterative Improvement

OUR TEAM

Myo Thiha



Rakshya Lama Moktan