HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Movie Evaluation

GROUP 32
Hoàng Trung Khải - 20225502
Lưu Hoàng Phan - 20225516
Vũ Việt Long - 20225508
Trịnh Duy Phong - 20220065
Nguyễn Phan Thắng - 20225529
Phạm Minh Tiến - 20225555

Table of contents

Introduction

1.Introduction

The success of a movie is a multifaceted phenomenon influenced by various factors, as evidenced by the literature reviewed. This section provides an overview and analysis of existing research on the determinants of movie success, encompassing perspectives ranging from revenue prediction to audience reception.

- The success of a movie cannot be solely measured by box office revenue.
- Factors such as actors, directors, and release timing also play crucial roles.
- Traditional prediction models often consider only a few factors, leading to limited accuracy.
- Analyzing audience sentiment and social media trends can provide valuable insights into a movie's success.
- Challenges include inconsistent data and reliance on pre- or post-release attributes.

Data Pre-processing

3.Data Pre-Processing

3.1 Dataset

- A collection of information about movie from Kaggle

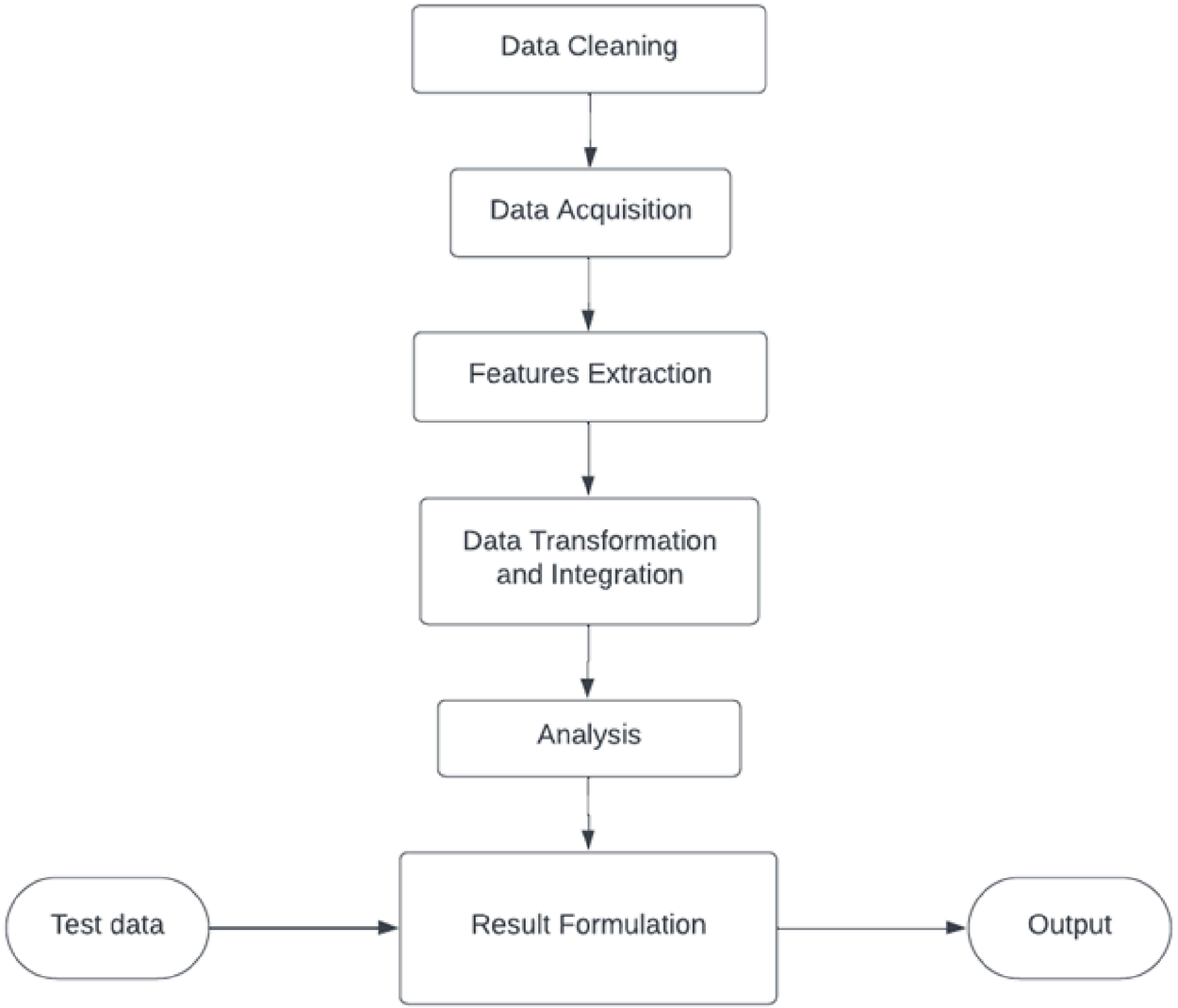
- A total of 91463 movies ,including theirs title, release_date,genre, duration, directors, actors, overview, user_score, number_of_vote, budget, revenue, and restriction.

3.2.Data cleaning

Initially, the dataset contained 91,463 movies.
However, since the focus of the project is on revenue prediction, we decided to exclude all movies without information on budget and revenue.
=> This reduced the dataset to 6,241 movies.

Methodology

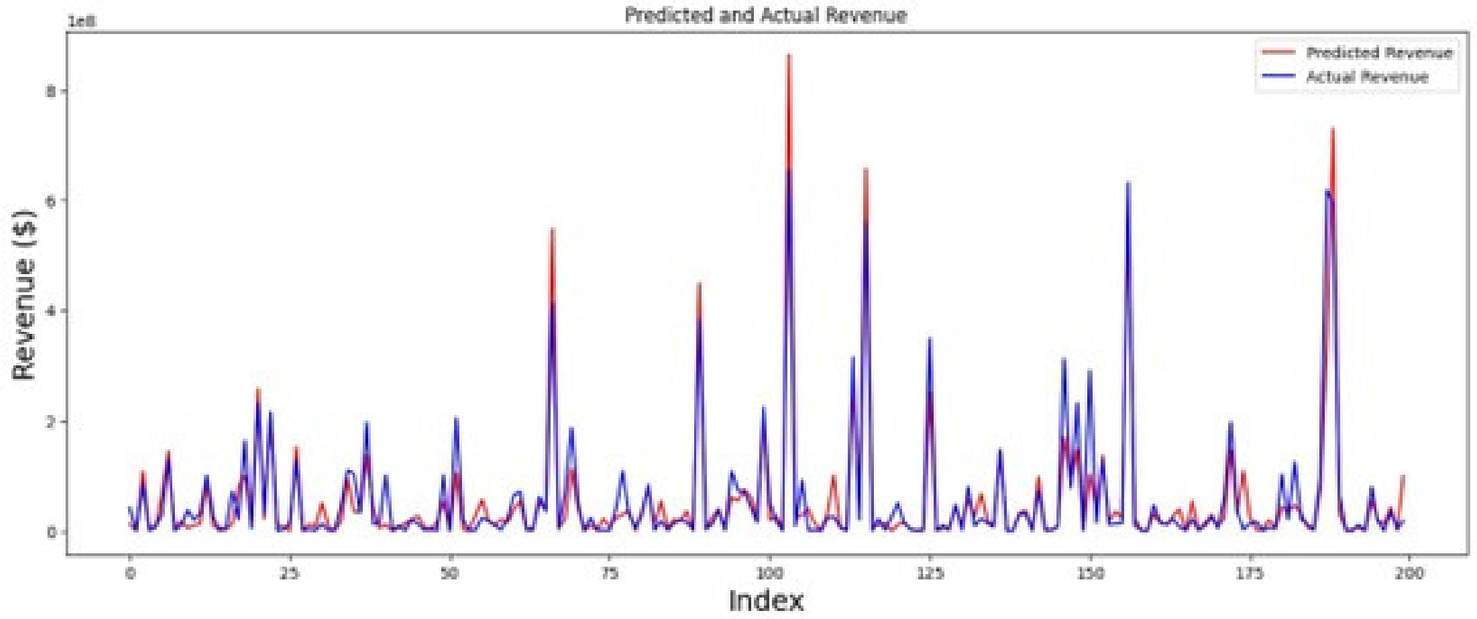
4.1.Work Flow


Data Cleaning
Data Acquisition
Features Extraction
Data Transformation and Integration
Analysis
Test data
Result Formulation
Output

4.2 Algorithms

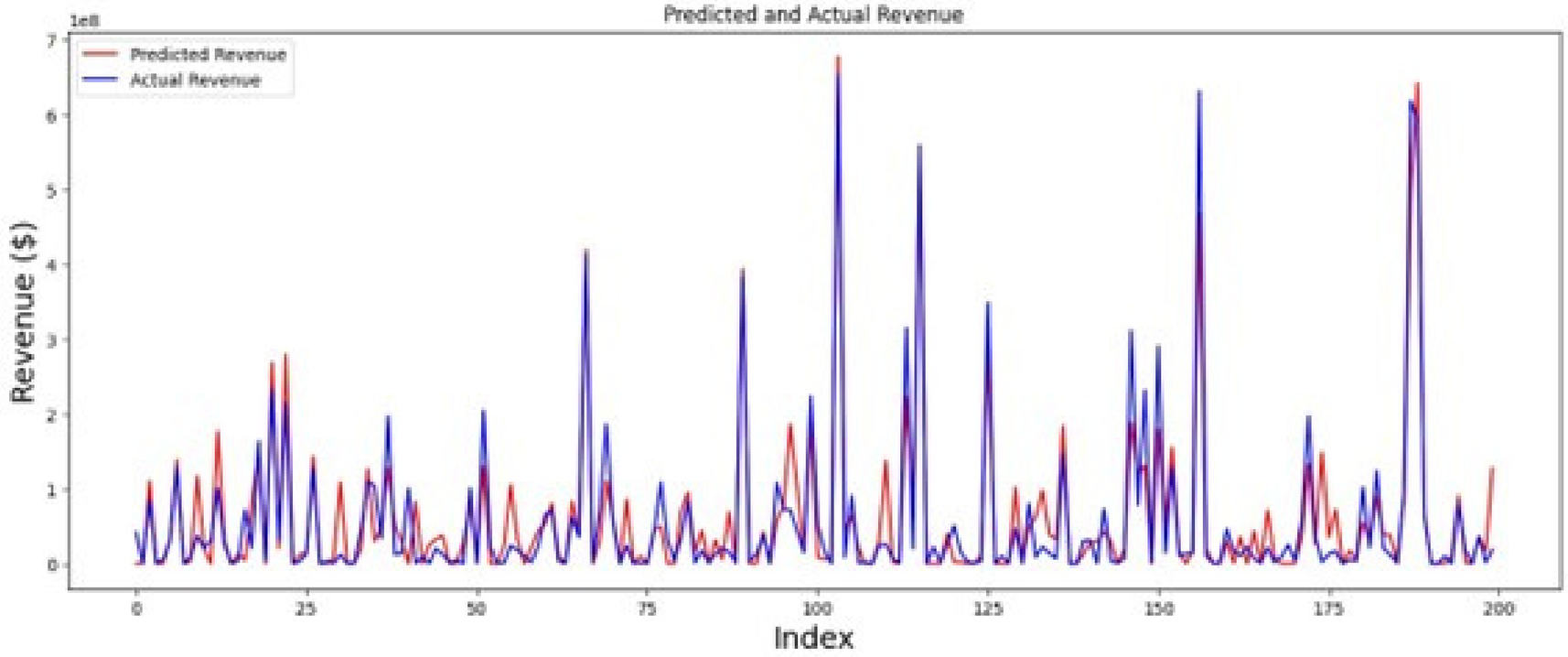
4.2.1.Grid Search

- Grid search is a foundational method in machine learning for optimizing hyperparameters
- Grid search constructs a grid of all possible combinations and evaluates each using a specified performance metric
- The combination yielding the best performance on a validation set is chosen as the optimal configuration



4.2.2.Ridge Regression

- For this algorithm, we used GridSearch to find best lambda (alpha in sklearn library)
- We trained the model as to make R-squared as high as possible and achieved 0.76 for Ridge Regression model. Other metrics such as Mean Absolute Error was 36 million and Root Mean Squared Error was 64 million



4.2.3. Decision Tree Regression.

Decision tree regression is a machine learning technique that constructs regression models in a tree structure
Decision tree regression includes hyperparameters that control the tree's structure and prevent overfitting.

In evaluating the performance of our predictive model, we conducted a 5-fold cross-validation.
The results of the evaluation are summarized in the table.

Fold	R^2	MAE	RMSE
1	0.77	0.47	1.25
2	0.83	0.53	1.18
3	0.75	0.51	1.28
4	0.81	0.49	1.15
5	0.83	0.48	1.1
Mean	**0.80**	**0.50**	**1.19**

4.2.4. Random Forest Regression

Random Forest Regression is an ensemble learning technique that extends the principles of decision tree regression to mitigate overfitting while maintaining predictive power.
The model's performance depends heavily on hyperparameters, including

The performance of our predictive model was assessed using a 5-fold cross-validation approach. The results are detailed in the table

Fold	R^2	MAE	RMSE
1	0.81	0.45	1.13
2	0.88	0.47	1.0
3	0.8	0.41	1.14
4	0.85	0.41	1.01
5	0.87	0.42	0.96
Mean	**0.84**	**0.43**	**1.05**

4.2.5.Gradient Boosted Decision Trees

Gradient Boosting Decision Trees (GBDT) is an ensemble learning method that falls under boosting, aimed at combining multiple weak models into a stronger one

There are 7 hyperparameters used to optimize GBDT:

- 'learning_rate': Controls the contribution of each tree in the ensemble.
- 'n_estimators': Number of trees in the ensemble.
- 'max_depth': Maximum depth of each decision tree.
- 'min_samples_split': Minimum number of samples required to split an internal node.
- 'min_samples_leaf': Minimum number of samples required at a leaf node.
- 'max_features': Number of features to consider when finding the best split at each node.
- 'subsample': Fraction of samples used for fitting individual base learners.

Table 4.2. GBDT (All Features)

Fold	R2	MAE	RMSE
1	0.81	7431829	56446224
2	0.93	7099429	44010212
3	0.90	5924753	47427212
4	0.92	5821911	53242866
5	0.92	5703825	45652173
Mean	**0.90**	**6396349**	**49355737**

Fold	R^2	MAE	RMSE
1	0.73	12854476	62846127
2	0.74	15468811	64600342
3	0.75	13611195	61646168
4	0.75	13997054	68125112
5	0.74	14085942	62587247
Mean	**0.74**	**14003495**	**63960999**

Table 4.1. GBDT (Pre-released Features)

4.2.6.XGboost

XGBoost, or Extreme Gradient Boosting, is a highly effective machine learning algorithm known for its exceptional performance across various tasks.

In addition to the hyperparameters shared with GBDT such as 'learning_rate', 'n_estimators', 'max_depth', and 'subsample', XGBoost includes unique hyperparameters:

- 'colsample_bytree': Allows for random subsampling of features at each tree-building step, improving model generalization.
- 'gamma': Represents the minimum loss reduction required for further partitioning on a leaf node of the tree.
- 'reg_alpha': Implements Lasso regularization, reducing model complexity by shrinking the magnitude of weights and inducing sparsity in the feature space.
- 'reg_lambda': Utilizes Ridge regularization, penalizing the square of weights to encourage smaller weight values and further reduce model complexity.

Fold	R^2	MAE	RMSE
1	0.73	12854476	62846127
2	0.74	15468811	64600342
3	0.75	13611195	61646168
4	0.75	13997054	68125112
5	0.74	14085942	62587247
Mean	**0.74**	**14003495**	**63960999**

Table 4.3. XGBoost (Pre-released Features)

Fold	R2	Scaled_MAE	Scaled_RMSE	Real_MAE	Real_RMSE
1	0.81	0.44	1.13	6863173	52955007
2	0.96	0.28	0.53	5216215	34450903
3	0.95	0.27	0.55	4410665	36427157
4	0.96	0.24	0.52	4069888	38614170
5	0.96	0.26	0.56	4378360	34023101
Mean	**0.93**	**0.26**	**0.56**	**4987660**	**39294067**

Table 4.4. XGBoost (All Features)

Data Description

4.1.Data Acquisition

Dataset Summary

Feature	Type	Count	Mean	Median	Min	Max	Std. Dev
Duration	Integer	6234	107.86	104	3	310	21.98
User score	Integer	6234	63.41	64	0	100	10.50
Number of Votes	Integer	6234	1561.44	516	0	35686	2898.90
Restriction	Integer	6234	3.61	4	0	5	1.53
Year	Integer	6234	1995.62	2000	1915	2023	17.03
Month	Integer	6234	6.98	7	1	12	3.42
Budget	Integer	6234	23623210	12000000	1	379000000	32986370
Revenue	Integer	6234	63672320	17969840	1	1362000000	126067600
Actor Star Power	Float	6234	576491800	450849500	1	5554789000	638543300
Director Star Power	Float	6234	437609200	102486900	1	9105032000	894363400

4.2. Features Extraction

4.2.1 Release Date

The year and month on revenue and budget is an important aspect to consider in the analysis

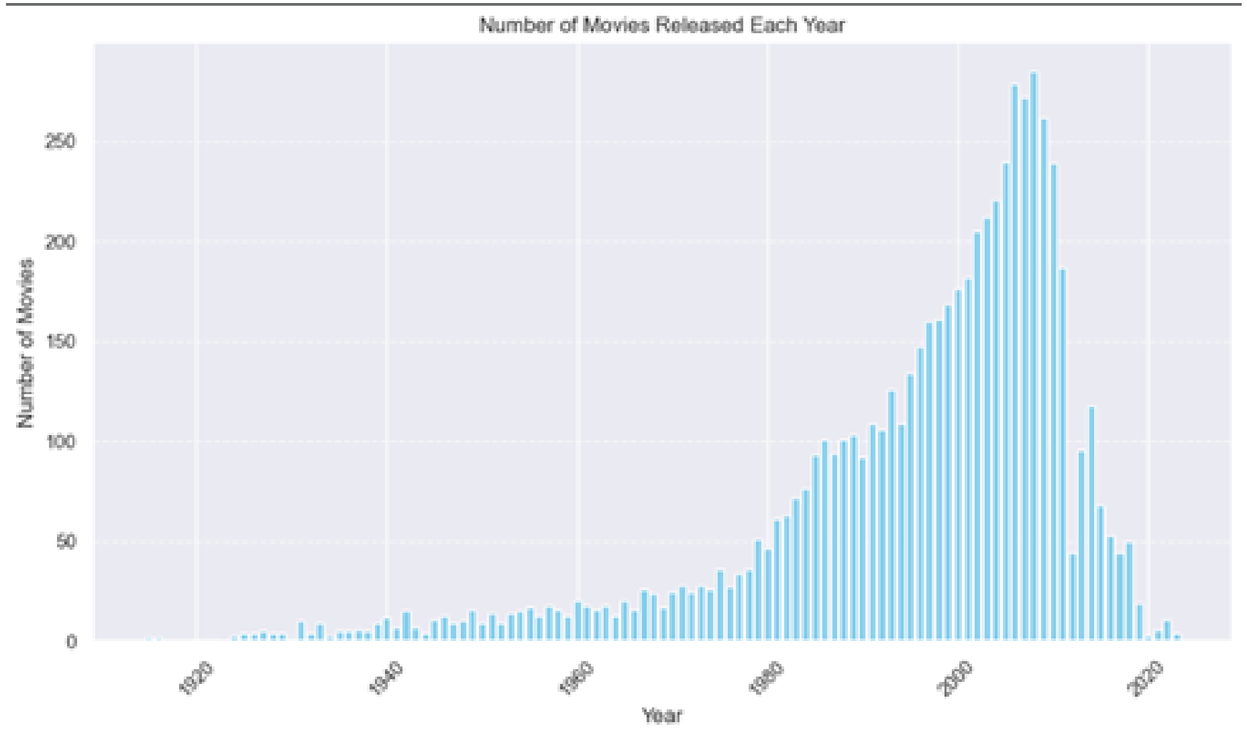


Figure 3.1. Number of Movies released each year

4.2.2.Star Power

- This paper delves into the significance of star power in shaping the success of movies, focusing on actors/actresses and directors
- In this context, a star is defined as an actor/actress whose previous movies have achieved notable box office success, indicating their established appeal among audiences.
- Essentially, the more successful projects associated with an actor/actress/director, the higher their perceived popularity and potential for future success.
-

4.2.3.Budget

Budget is a key factor in predicting a movie's success before its release. Movies with bigger budgets, covering both production and marketing, often generate more excitement and anticipation

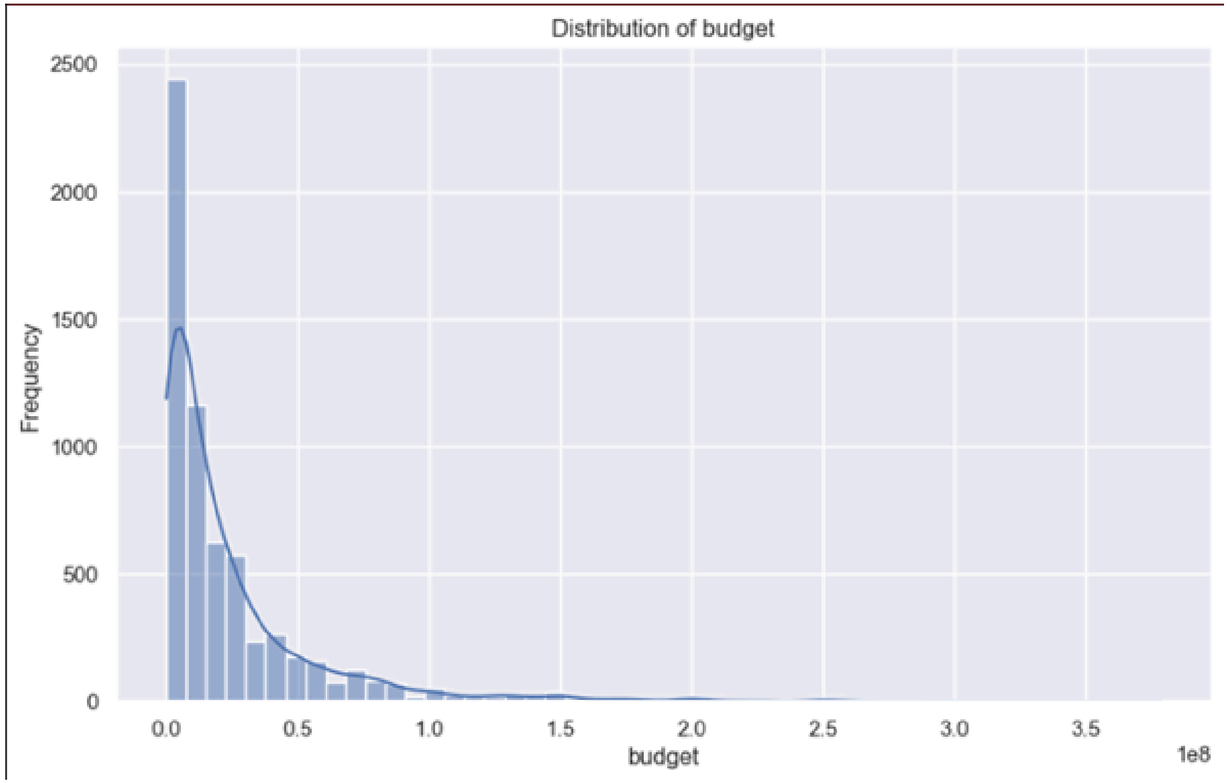


Figure 3.2. Distribution of Budget

4.2.4.Restrictions

Content restrictions, such as age ratings and censorship guidelines, can influence the audience demographics and the movie's accessibility in various markets

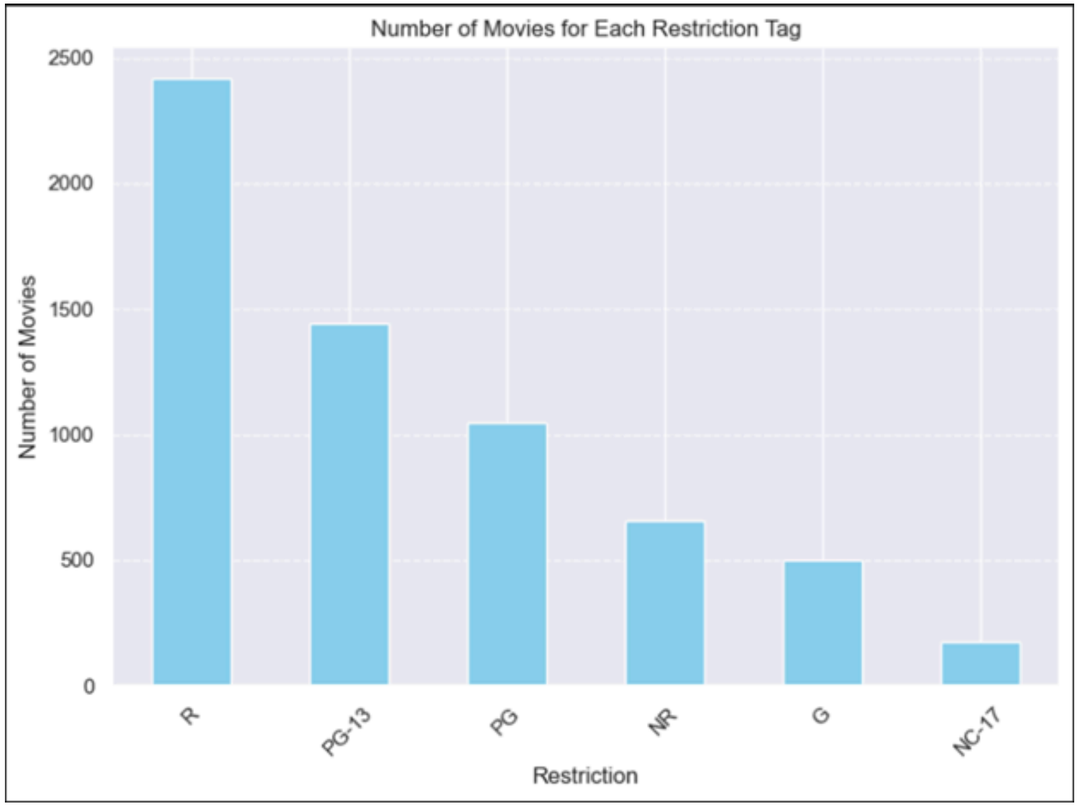


Figure 3.3. Number of Movies each Restriction Tag

4.2.5.Number of votes & User Score

- User scores are pivotal in evaluating audience reception toward a movie
- This mean score serves as a valuable metric for assessing overall performance and audience satisfaction.
- This metric is crucial for understanding the movie's popularity, level of viewer engagement, and overall reception.
- a higher number of votes contributes to assessing the movie's success and reception accurately.

4.2.6.Genres

- Genres categorize movies into groups that reflect the primary focus or tone of the film
- Common genres include action, comedy, drama, horror, romance, science fiction, and thriller, among others

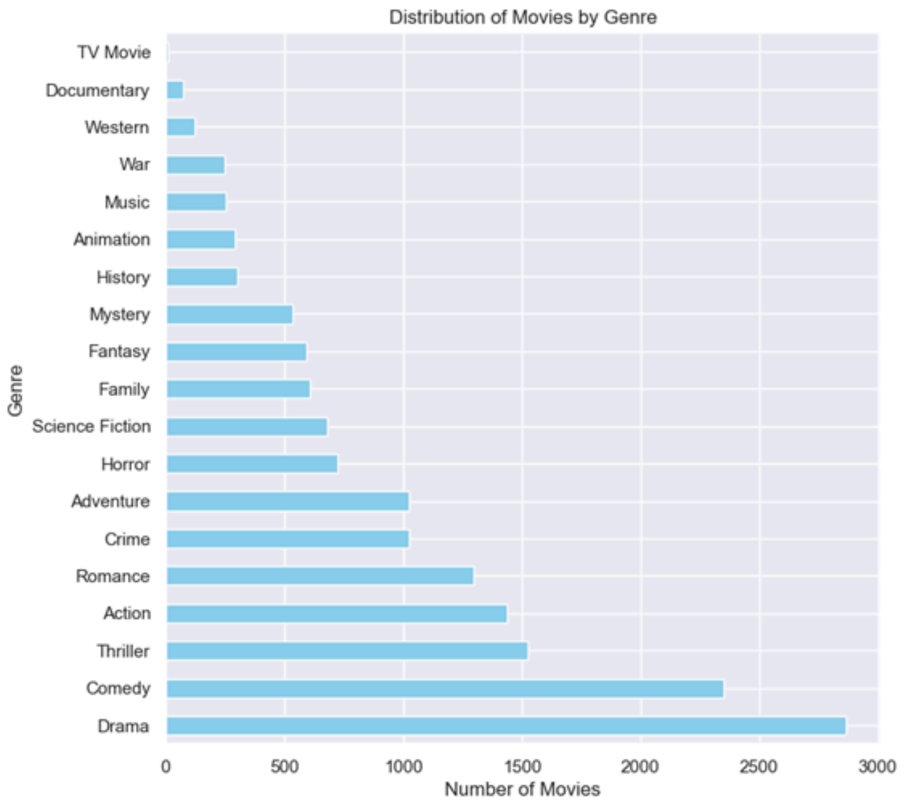


Figure 3.4. Distribution of Movies by Genre

4.2.7.Duration

- The duration of a movie refers to its length in terms of running time, typically measured in minutes
- Shorter films may be more suitable for certain genres or storytelling formats, such as animation or experimental cinema, while longer durations are often associated with epic narratives or in-depth character development
- The duration of a movie can also influence its ticket price. Generally, longer movies may have higher ticket prices due to the perceived value of a more extended entertainment experience.

4.3.Data Integration and Transformation

4.3.1.Restriction

- Firstly, we integrate the various restriction ratings present in the dataset into a standardized MPAA rating system
- Secondly,we employ label encoding to represent the MPAA ratings numerically

MPAA Restriction	Label Encode
G	0
NC-17	1
NR	2
PG	3
PG-13	4
R	5

Table 3.2. Restriction

4.3.2.Release Date

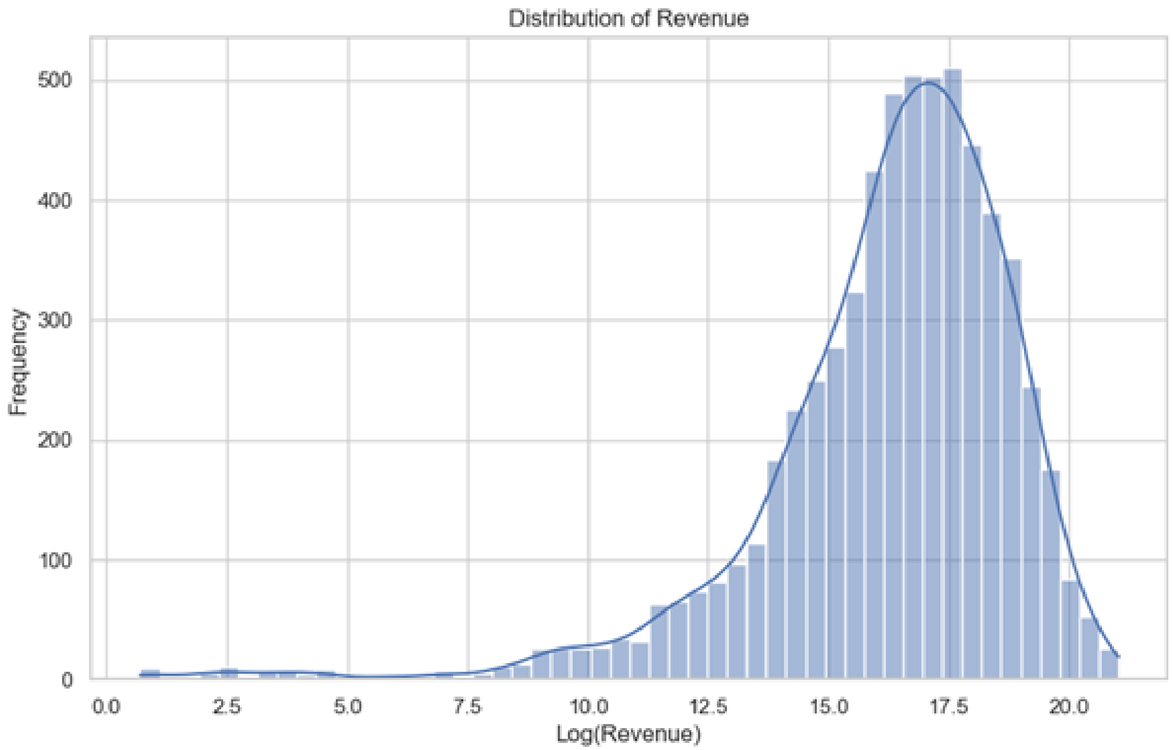
- This feature is seperated into 2 different, year and month.
- Since the release month represents a cyclical pattern, exploring its relationship with movie income enables us to identify any seasonal trends or variations in audience preferences.
- Otherwise, year represents some overarching patterns such as changes in audience preferences or shifts in industry dynamics.

4.3.3.Actors/Actresses & Directors Star Power

- Calculate the total revenue of movies for each actor/actress or director: For each actor/actress, sum the revenue of all movies they have participated in. Similarly, we also calculate total director’s movie revenue.
- Calculate the mean revenue of all actors/actresses or directors involved in the film. This provides the average revenue contribution of all actors/actresses or directors participating in the movie.

4.4.4.Revenue

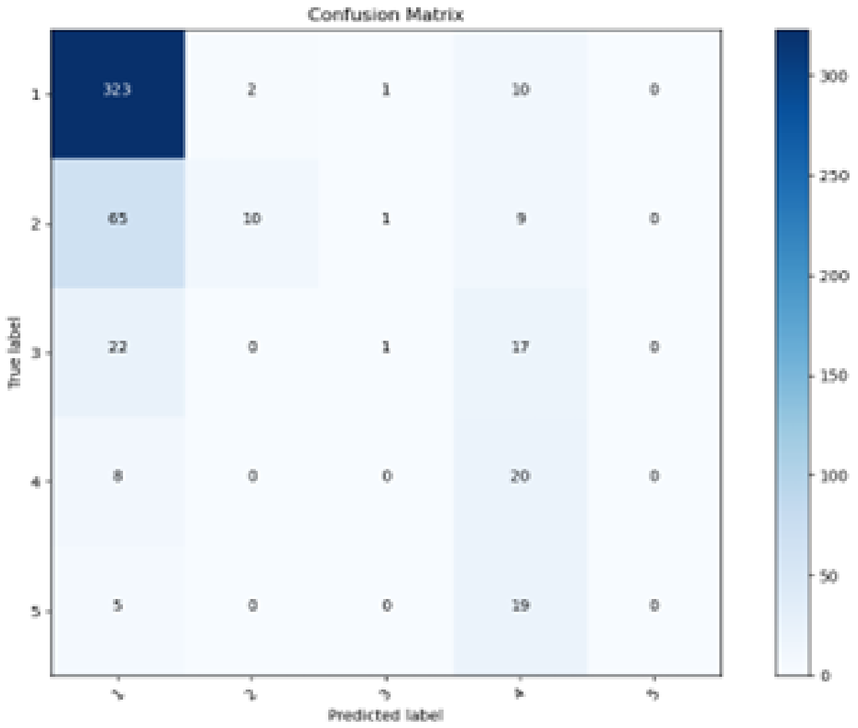
Revenue data spans a wide range of values, from very small to very large numbers (from 1 to over 1 billion).Taking the logarithm of the revenue values can help normalize the data, making it easier to compare and analyze.



RESULTS AND EVALUATION

Classification Algorithms

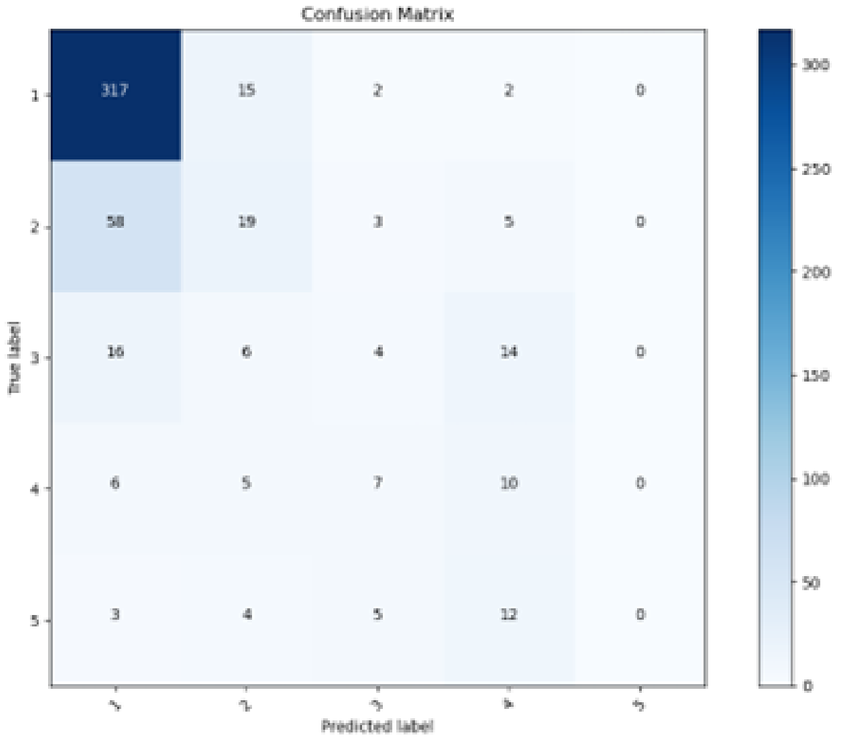
Decision Tree



	precision	recall	f1-score	support
0	0.76	0.96	0.85	336
1	0.83	0.12	0.21	85
2	0.33	0.03	0.05	40
3	0.27	0.71	0.39	28
4	0.00	0.00	0.00	24
accuracy			0.69	513
macro avg	0.44	0.36	0.30	513
weighted avg	0.68	0.69	0.62	513

Figure 4.1. The Grid Search provided us with the combination (None, 6, 9)

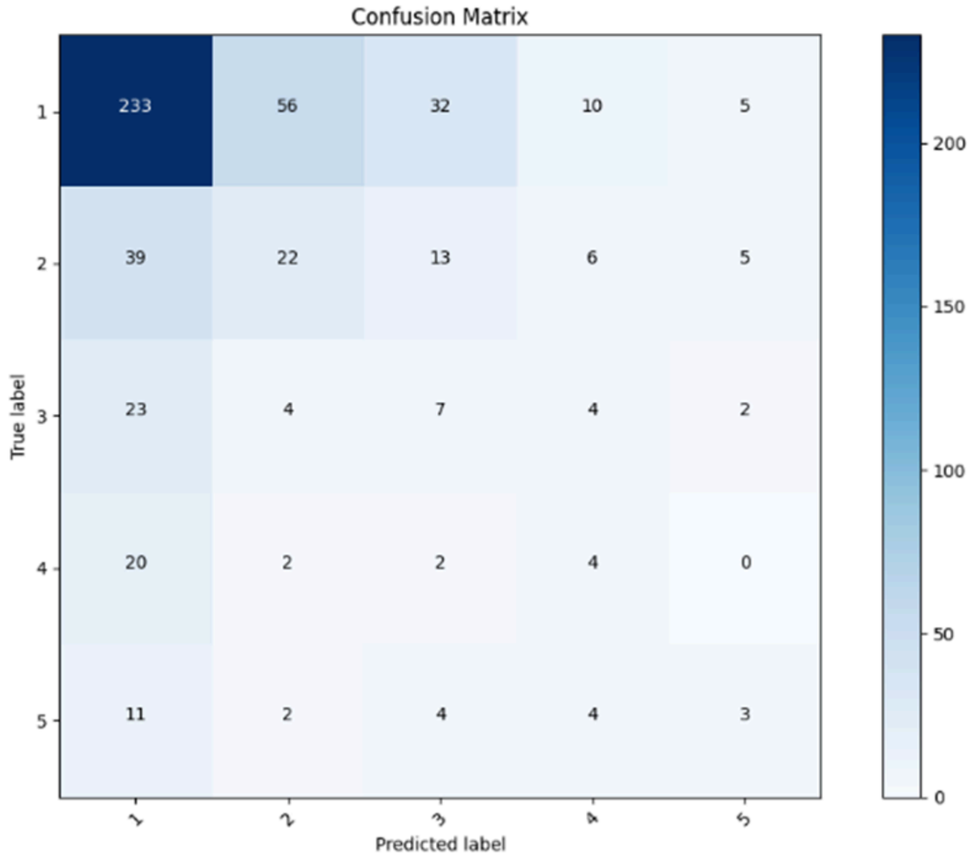
Random Forest



	precision	recall	f1-score	support
0	0.79	0.94	0.86	336
1	0.39	0.22	0.28	85
2	0.19	0.10	0.13	40
3	0.23	0.36	0.28	28
4	0.00	0.00	0.00	24
accuracy			0.68	513
macro avg	0.32	0.32	0.31	513
weighted avg	0.61	0.68	0.64	513

Figure 4.2. The Grid Search provided us with the combination (150, 6,None, 9)

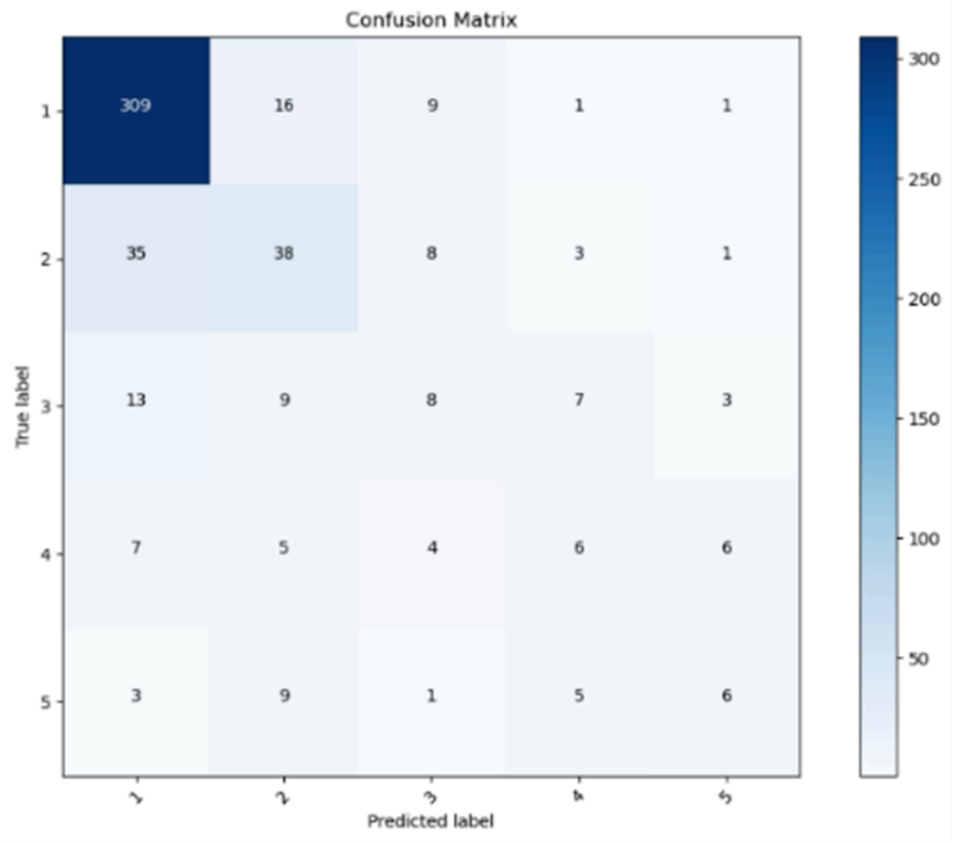
SVC kernal RBF



	precision	recall	f1-score	support
0	0.71	0.69	0.70	336
1	0.26	0.26	0.26	85
2	0.12	0.17	0.14	40
3	0.14	0.14	0.14	28
4	0.20	0.12	0.15	24
accuracy			0.52	513
macro avg	0.29	0.28	0.28	513
weighted avg	0.54	0.52	0.53	513

Figure 4.3. The Grid Search provided us with the combination (100, 0.3, ‘rbf’)

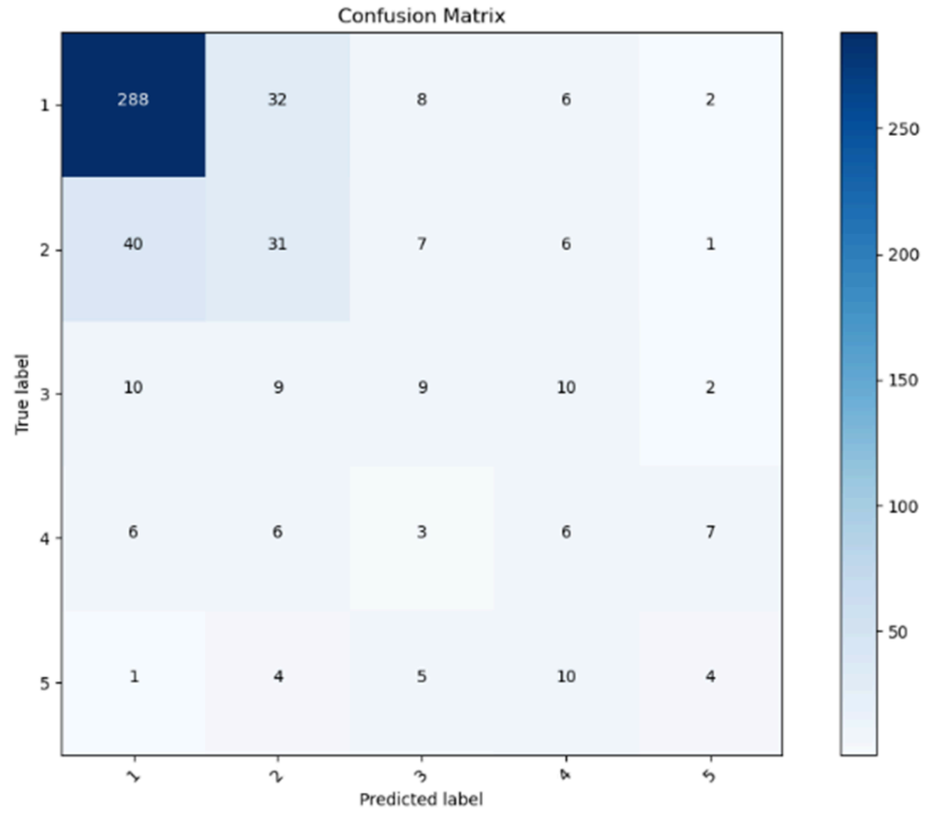
XGBoost Classification



	precision	recall	f1-score	support
0	0.84	0.92	0.88	336
1	0.49	0.45	0.47	85
2	0.27	0.20	0.23	40
3	0.27	0.21	0.24	28
4	0.35	0.25	0.29	24
accuracy			0.72	513
macro avg	0.45	0.41	0.42	513
weighted avg	0.69	0.72	0.70	513

Figure 4.4. The Grid Search provided us with the combination (0.1, 3, 3, 200, 0.6)

Neural Network



	precision	recall	f1-score	support
0	0.83	0.86	0.85	336
1	0.38	0.36	0.37	85
2	0.28	0.23	0.25	40
3	0.16	0.21	0.18	28
4	0.25	0.17	0.20	24
accuracy			0.66	513
macro avg	0.38	0.37	0.37	513
weighted avg	0.65	0.66	0.65	513

Figure 4.5. Optuna provided us with the combination {‘number layer’: 3; layer_size_0: 58, layer_size_1: 35; dropout_rate_1: 0,43; layer_size_2: 32; dropout_rate_2: 0.09}

HUST

THANK YOU !