الجامعة المصرية اليابانية للعلوم و التكنولوجيا
エジプト日本科学技術大学
EGYPT-JAPAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Course Name: Mathematics for Data Science **Course Code:** AID-311

Academic Year: Fall 2023/2024 **Lecturer:** Dr. Ahmed Anter

Movie Production Analysis & Forecasting

Documentation

Student Name: Yousef Ibrahim Gomaa Mahmoud **Student ID:** 320210207

- ## Abstract:

Over the last few decades, movies have become an important media for providing joyous entertainment for the viewers. Likewise, the producers and stakerholders look forward to invest their time and resources to gain the most material and non-material return out of it. As such, predicting revenue (collection) of movies based on their quality before releasing on the big screens has become an apparent necessity. Two of the more common metrics for measuring said quality are net profit, and the viewers' reception. (ratings assigned by moviegoers) Because the current box office revenue projection algorithm ignores the structure of the film industry, the prediction accuracy is not up to par. This research first develops a framework for two-stage collaborative feature processing between humans and machines. First, the regression decision tree technique is used to process all box office features preliminary and automatically remove unnecessary aspects based on box office data. Feature processing and the constructed Neural Networks are combined in the second stage. At this point, different, incompatible feature sets are segregated and the machine-processed features are manually sorted. In addition to the technical aspects, the research addresses the economic and socio-cultural factors influencing movie production. A close examination of industry trends, budgeting constraints, and the evolving landscape of audience preferences provides a holistic understanding of the forces steering the cinematic ship. Therefore, in this work, the goal is to investigate some features of the film industry and analyze them in an attempt to establish what defines a "successful" movie.

1. Introduction:

Throughout this paper and the attached notebook, we explore various aspects of movie production and promotion, aiming to uncover patterns and insights that can contribute to the success of a film. Each column provides valuable information that could hold the key to understanding the dynamics of a movie's performance.

Our task is to analyze a dataset and build a predictive model using different classical machine learning techniques, such as “Decision Tree Regressor” and “Naive Bayes Classifier”, that can help stakeholders make informed decisions about movie production and marketing strategies; it potentially reduces the risk of investors, studios and other stakeholders to select successful film candidates and have them chosen before the production process starts. In addition, because context may differ, it is necessary to have different sources for the data in order to compare results.[5]

Provided that we have already taken the measures of data collection and entry[6], the resultant dataset turned out to compromise of several key parameters as follows:

- **Marketing Expense**
  - The amount of money spent on promotional activities and advertising for the movie.
- **Production Expense**
  - The cost incurred in the creation and production of the movie, including expenses related to filming, sets, and equipment.
- **Multiplex Coverage**
  - The percentage or number of multiplex cinemas where the movie is being screened.
- **Budget**
  - The total financial investment allocated for the movie, encompassing both production and marketing expenses.
- **Movie Length**
  - The duration of the movie in terms of running time, typically measured in minutes.
- **Lead Actor, Lead Actress, Director, Producer and Critic Rating**

    - The evaluation, rating or popularity score assigned to entity that is involved in the movie's production. The critic rating being the score given by movie critics, reflecting the quality and performance of the film.
- **Trailer Views**
    - The number of views the movie trailer has received on various platforms.
- **3D Availability**
    - Indicates whether the movie is available in 3D format or not.
- **Time Taken**
    - The duration of time taken from the start of production to the release of the movie.
- **Twitter Hashtags**
    - The hashtags associated with the movie on Twitter, reflecting social media trends and discussions.
- **Genre**
    - The category or type of the movie, such as action, drama, comedy, etc.
- **Average Age of Actors**
    - The average age of all the actors in the movie.
- **Number of Multiplexes**
    - The total count of multiplex cinemas where the movie is being screened.
- **and Collection**
    - The total revenue generated by the movie, typically measured in terms of box office earnings.

In order to achieve our goal, some preprocessing operations may be required in order to prepare the data so that it could be used in the machine learning process, that is to accurately provide a satisifactory prediction accuracy.

As for the machine learning techniques that need to be involved, each of which has their uses and differences. Some examples of the models being used are:

- Naive Bayesian Classifier.
- Decision Tree (entropy, and error estimation)
- Linear Discriminant Analysis (when used as Classifier)
- Principal Component Analysis
- K-NN (model fitted with different distances)
- Neural Networks Model

Finally, each model is evaluated in terms of accuracy and other metrics, that is to find the best suited one for said task, the results are shown in the conclusion section.

2. Related Work:

2.1. Predicting Gross Movie Revenue:

This paper [1] discusses the challenges of predicting movie box office revenue due to numerous factors like star power, release date, budget, and unpredictable audience reactions. Modern computing and historical movie databases offer a solution. Jeffrey et al. [5] used data to predict US domestic market revenue, employing variables like budget, running time, star power, and MPAA ratings. They found that models incorporating opening weekend business most accurately predicted gross revenue, emphasizing its significance. The study used 311 films from 1998, considering variables such as genre, MPAA ratings, country of origin, star power, production budget, sequel indicators, holiday releases, first weekend screenings, critic ratings, and academy award nominations.

2.2. A Movie Box Office Revenue Prediction Model Based on Deep Multimodal Features

In this paper [2], the author wrote about movie prediction that involves assessing various criteria such as budget, actors, director, producer, set locations, story writer, release day, competing releases, music, release location, and target audience to anticipate revenue and performance. IMDb serves as a key platform for movie ratings and reviews. Analyzing customer reviews on IMDb can gauge satisfaction or dissatisfaction, impacting movie revenue. While data analysis provides powerful insights, it doesn't guarantee the fate of an individual project.[7] Movie success varies, encompassing ticket sales, profit margin, reviews, social impact, franchise potential, or critical awards.[8]

2.3. A Statistical Analysis of Gross Revenue in Movie Industry

This paper [3] explores factors influencing movie gross revenue, including release year, genre, cast, director, and movie length. Utilizing statistical models, linear regression is employed to examine relationships between gross revenue, movie length, and budget. Past research is referenced, highlighting findings on revenue streams, simultaneous releases, viewer satisfaction, and the impact of internet piracy. The dataset comprises 34,168 movies released from the 1880s to the 2010s. Exploratory analysis includes distributions of predictor variables, and regression models are built to assess the relationship between gross revenue, movie length, and budget.[9] Results indicate a positive correlation between budget and movie length, with higher budgets associated with longer runtimes. The study observes trends in movie runtimes over decades, noting an increasing mean and median, with a decreasing inter-quartile range over time.

2.4. Movie Revenue Prediction Based on Purchase Intention Mining Using YouTube Trailer Reviews

This paper [4] went through with explaining the state of the U.S. motion picture industry, as with an average investment of $60 million per film, it faces uncertain profitability. Previous studies exploring movie revenue prediction have yielded unsatisfactory results. This paper addresses the challenge by examining movie reviews, specifically focusing on extracting purchase intention, sentiment analysis of YouTube trailer reviews, and movie metadata for early revenue prediction. The study uses a dataset from 29 movies released in 2016 and 2017, extracting metadata and financial information from Box Office Mojo. The proposed approach achieves an effective purchase intention mining accuracy of 0.90, demonstrating correlations between purchase intention, sentiment ratios, and movie revenue. The article contributes by introducing purchase intention mining for revenue prediction, creating a lexicon, providing a dataset for public use, and comparing machine learning algorithms for movie revenue prediction using reviews.[10] The paper concludes with a review of previous studies, the proposed method, experimental results, and future implications.

3. Methodology

The analysis methods used are divided into two-stages; the pre-processing analysis, during which the incorrect and/or missing data is handled, the outliers are removed and the duplicates are dropped, as well as binning/label encoding processes are held, and the post-processing analysis, during which both the skewness, kurtosis are measured and tests revolving around the hypothesises of means are done, such as “Z-Test”/”T-Test,” “ANOVA,” and “Chi Square Test.”

Afterwards, the machine learning techniques are fitted using a training set and tested/validated on a separate set, both of which are generated, transformed and scaled using the built-in functions of “SciKit Learn.” Later on, we use the evaluation metrics to measure how well the model performs on new data, including accuracy, mean absolute error, mean square error, precision score, recall score, f1_score and more. Note that the evaluation step differs from one technique (classification and regression, both of which are utilized throughout this paper) to another.

First, before we explain our machine learning methods, considering how big our dataset is, some of the data may be highly correlated in a way that negatively affects our evaluation later on, as such, some dimensionality reduction is due.

4. Proposed Model

4.1. Importing Dataset

This phase is self-explanatory, as we import the dataset from the proposed comma separated values file (“.csv” extension) into a pandas dataframe, which will be used throughout the whole process.

4.2. Pre-processing

During the pre-processing phase, the imported dataset is investigated using some standard methods:

- A command evalutes the mean, min, max, standard deviation, and the quartiles of the dataset.
- A command returns the count of null values in the dataset, and another that returns the duplicates, if there exists a duplicate value in a numerical feature, it replaces it with the mean. Otherwise, it replaces it with the mode.
- Next, we evaluate the skew and kurtosis of the data to determine how the values are distributed throughout the system.

- Afterwards, we use Z-Test and ANOVA (Analysis of Variance), which are statistical tests used in different scenarios, to make inferences about population parameters or to compare means across different groups.
- Some data visualization can significantly help in understanding how the data is structured and distributed, such as pie charts, histograms and scatter plots, which is dependent on the type of data being visualized. (categorical/numerical and univariate/bivariate)

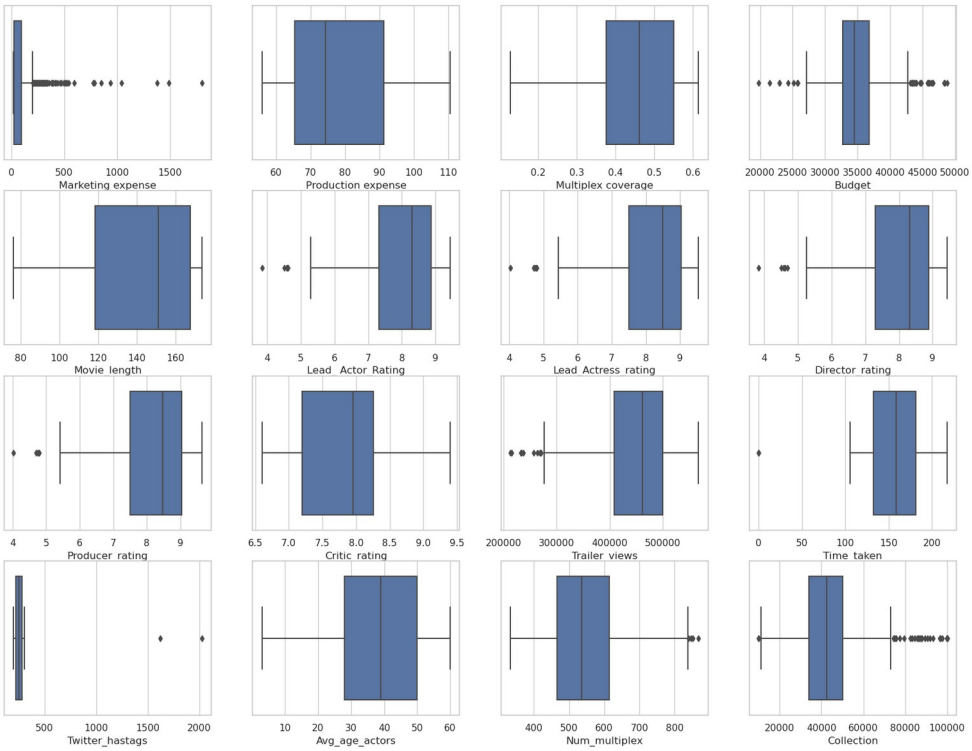

Figure 1. *Box Plot of Dataset Features*

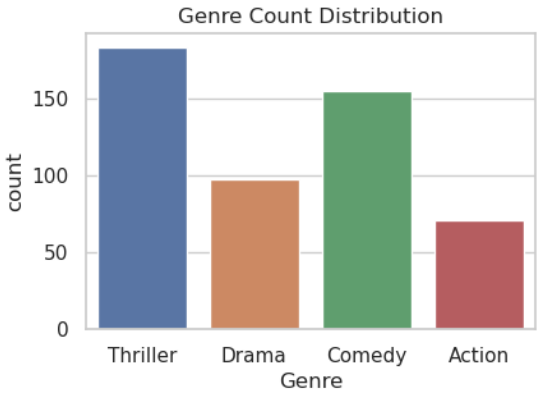

Figure 2. *Genre Count Distribution*

4.3. Feature Selection

During the feature selection phase, we evaluate which data would prove to be useful for prediction later on, as highly correlated data may cause the models' accuracies to decline and increase computation overhead unnecessarily, this can be achieved using a correlation matrix which calculates the linear correlation coefficient between each feature and the other using the formula.

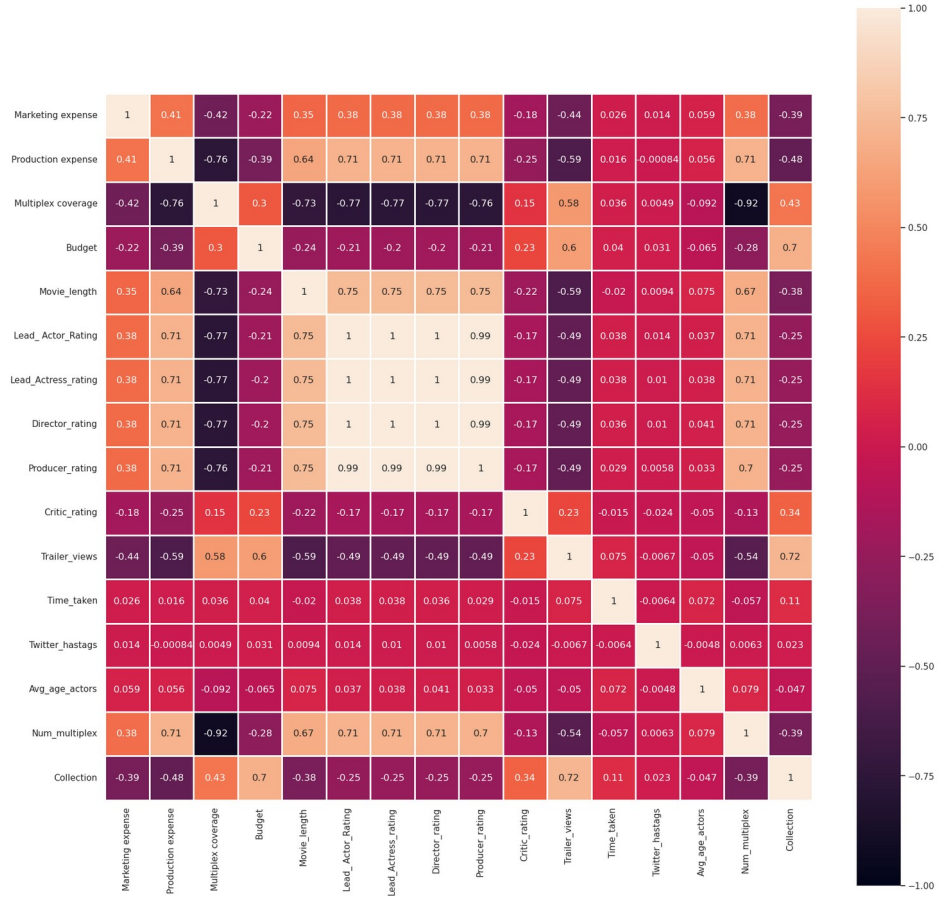

Figure 3. *Correlation Matrix of Dataset*

Where ***r = 0*** represents no correlation, ***+r*** presents positive correlation and ***-r*** represents negative correlation.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{([n\sum x^2 - (\sum x)^2]\cdot[n\sum y^2 - (\sum y)^2])}}$$

4.4. Feature Reduction

Based on the earlier phase, we can choose to reduce the dimensionality of the data by using Linear Discriminant Analysis, Principal Component Analysis and Singular Value Decomposition, which can improve the evaluation accuracies of the models by choosing the data best fit to represent the model before training.

a) Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised dimensionality reduction technique that is commonly used to simplify the complexity in high-dimensional data while retaining trends and patterns. It works by transforming the original features into a new set of uncorrelated variables called principal components. These principal components are ordered by the amount of variance they represent in the data. First, we determine the number of features to retain, which is a hyperparamater to be tuned. As we should choose the number of components that make up a high percentage of the total variance in the data. In other words, we choose the components that best present the data. Then, each principal component is calculated as a linear combination of the original features (eigen vectors), then we achieve the reduced dimensions by multiplying the original data by the matrix of principal components.

b) Singular Value Decomposition

Another dimensionality reduction technique is the Singular Value Decomposition (SVD). SVD decomposes a matrix into three other matrices, which can then be used to represent the original matrix with reduced rank. In the context of this dataset, SVD is often used to perform the decomposition and achieve dimensionality reduction by dropping the values with the least variance in the decomposed matrices. Similar to the PCA technique, we can choose the amount of singular values to retain.

4.5. Classification/Regression Methods

a) Naive Bayesian Classifier

One of the classification techniques used is the “Naive Bayesian” classifier, which uses the binned version of the dataset (as shown in **Figure 1**) as it is a probabilistic classification algorithm based on Bayes' theorem, which assumes that all the features are independent (which may be proven true according to the correlation matrix that is illustrated in the preprocessing step.) In this case, the target feature is the ‘Collection’ which, as previously mentioned in Section 1, represents the revenue collected. Although the formula can be written as follows:

$$P(Collection = c \mid features) = \frac{P(features \mid Collection = c) \cdot P(Collection = c)}{P(features)}$$

Instead, we will use the built-in “Scikit Learn” module GaussianNB.

Budget_Category	Movie_Length_Category	Actor_Rating_Category	Actress_Rating_Category	Director_Rating_Category	Producer_Rating_Category	Critic_Rating_Category	Collection_Category
C	Long	Good	Excellent	Excellent	Good	Bad	Failure
C	Short	Bad	Bad	Bad	Bad	Good	Success
B	Medium	Bad	Bad	Bad	Bad	Bad	Failure
A	Long	Excellent	Excellent	Excellent	Excellent	Bad	Failure
D	Short	Bad	Bad	Bad	Bad	Excellent	Success

Figure 4. *Sample of Categorized/Binned Features*

b) Decision Tree Classifier

Another one of the classification techniques used is the “Decision Tree” classifier, which also uses the binned data. However, the decision tree classifier is a predictive model that maps features (or attributes) to outcomes by recursively partitioning the data into subsets based on the values of the features. It recursively chooses the feature that provides the best split (using the Entropy metric for this case) and continues until either a maximum depth is specified or has already extracted the pure leaves of the tree. (No further expansion)

c) Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classification and dimensionality reduction technique that finds the linear combinations of features that best separate different classes in the data. It aims to maximize the separation between classes while minimizing the variance within each class; it tries to find a projection of the data into a lower-dimensional space where the classes are well-separated. Therefore, it can be particularly useful when dealing with multiclass classification problems, although with our dataset, that is not the case.

d) Neural Network

This model is a little different than the classical machine learning techniques, as we are using “Tensorflow” and its modules to develop an Artificial Neural Network (ANN) of simple architecture to do this task. An ANN is a computational model inspired by the way biological neural networks in the human brain work. For this task, we can assume a binary classification task, that is because our target contains 2 classes, “Success” and “Failure,” as well as a sigmoid activation function in the output layer, or we can use it to perform regression on the encoded dataset. With enough fine-tuning, this model can improve enough to provide one of the better accuracies in terms of classification or regression.

e) K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple classification technique that has a time complexity of O(1) for training, and O(N) for predicting as it does not train the data, but instead memorizes them. It identifies the k-nearest neighbors (value ‘k’ is given) where the majority

class is assigned as the predicted class for the target. It utilizes different distance metrics, such as Euclidean distance, Manhattan distance, and the Cosine distance in order to evaluate the k-nearest neighbors. In our case, we are going to use all of the previously mentioned distance metrics for our analysis study.

4.6. Evaluation Metrics

Evaluation metrics are crucial for assessing the performance of machine learning models, whether they are employed for classification or regression tasks.

a) Classification Evaluation Metrics:

- **Accuracy:**

It is the ratio of correctly predicted instances to the total instances.

- **Precision:**

It is the ratio of correctly predicted positive observations to the total predicted positives.

- **Recall (Sensitivity or True Positive Rate):**

It is the ratio of correctly predicted positive observations to all the actual positives.

- **F1 Score:**

It is the harmonic mean of precision and recall.

- **Area Under the Receiver Operating Characteristic (ROC-AUC):**

It represents the area under the ROC curve, which measures the trade-off between sensitivity and specificity. (difference)

b) Regression Evaluation Metrics:

- **Mean Absolute Error (MAE):**

  The average absolute differences between predicted and actual values.

- **Mean Squared Error (MSE):**

  The average of the squared differences between predicted and actual values.

- **Root Mean Squared Error (RMSE):**

  The square root of the MSE, providing the error in the same units as the target variable.

4.7. Proposed Model Illustration:

The high-resolution model is attached in the following page.

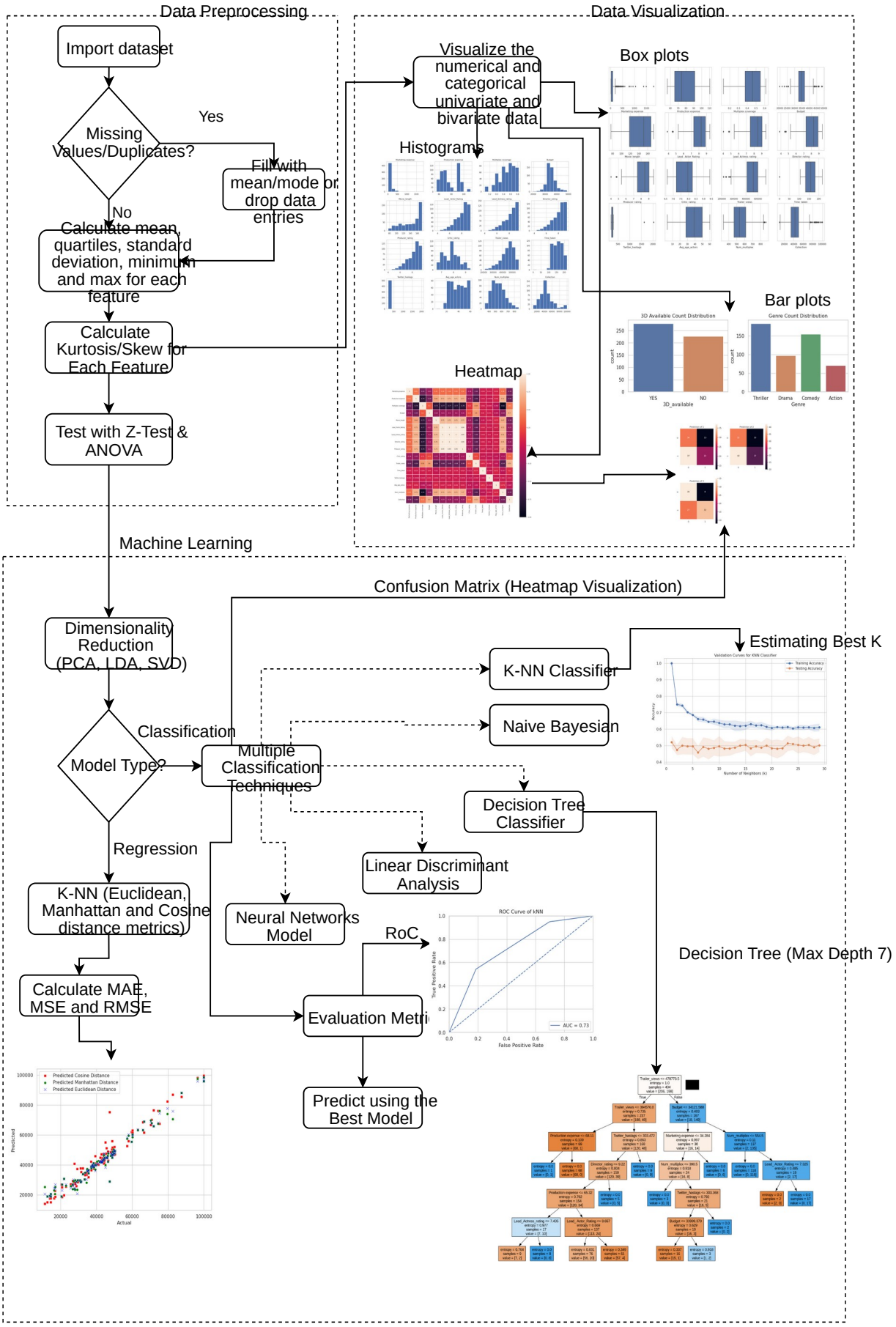

Illustration 1. *Proposed Model*

5. Results and Discussion

5.1. Dataset Description

The dataset encompasses various parameters related to movies, including marketing expenses (money spent on promotional activities and advertising), production expenses (costs involved in creating the film), multiplex coverage (percentage or count of cinemas screening the movie), budget (total financial investment for production and marketing), movie length (duration in minutes), evaluations and ratings for lead actors, actresses, directors, producers, and critics, trailer views, 3D availability, time taken from production to release, Twitter hashtags reflecting social media trends, movie genres, average age of actors, the number of multiplexes screening the movie, and collection (total revenue measured in box office earnings). This comprehensive dataset provides insights into diverse facets of the movie industry, aiding analysis and decision-making processes.

The dataset consists of 506 entries and 18 features, 12 of whom are of type ‘float64’, 4 are of ‘int64’, and 2 are of type ‘object.’

Feature	Datatype
Marketing expense	float64
Production expense	float64
Multiplex coverage	float64
Budget	float64
Movie_length	float64
Lead_ Actor_Rating	float64
Lead_Actress_rating	float64
Director_rating	float64
Producer_rating	float64
Critic_rating	float64
Trailer_views	int64
3D_available	object
Time_taken	float64
Twitter_hastags	float64
Genre	object
Avg_age_actors	int64
Num_multiplex	int64
Collection	int64

Table 1. *Dataset Datatypes*

5.2. Resultant Tables from Methods

a) Z-Test

The first step in Z-test is to formulate the hypothesis, then calculate the z-score according to the formula:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Then determine the Z-score critical to the significance level α, if the calculated Z-score is higher than the critical value, the hypothesis is rejected. However, in our case, we fail to reject the null hypothesis, indicating no significance difference between the means.

Z-Score	0.12842754827482128
Critical Z-Score	1.6448536269514722
Result	Fail to Reject Null Hypothesis
p-value	0.44890531564904823
Result	Fail to Reject Null Hypothesis

Table 2. *Z-Test Analysis Table*

b) ANOVA

This method is used to determine whether there are any statistical difference between the means of three or more independent groups by comparing the variance. The first step is to formulate the hypothesis, then calculate the sample means, variances of each group, then calculate sum of squares of each group and between groups, degree of freedom and mean squares respectively. From the mean squares, calculate the f-score which we can use to infer our decision. In our case, we reject the null hypothesis and the f-score is higher than the critical f-score, indicating that at least one group mean is different.

F-Statistic	12.610774279494422
P-value	3.671096649089372e-10
Result	Reject the Null Hypothesis

Table 3. *ANOVA Analysis Table*

5.3. K-Nearest Neighbors

As explained in the propose model, we will use K-nearest neighbors which assigns the most frequent class that is assigned to the neighbors to the target, and evaluate the best value for K in order to achieve the best possible accuracy, that is through validation curves.

In addition, we shall use some of the dimensionality reduction techniques in order to further improve the resultant accuracy.

- **KNN (with PCA) (K=3)**

Accuracy	79.41176470588235%
Precision	0.8974358974358975
Recall	0.6730769230769231
F-measure	0.7692307692307692

- **KNN (with LDA) (K=3)**

Accuracy	85.29411764705883%
Precision	0.9512195121951219
Recall	0.75
F-measure	0.8387096774193549

- **KNN (with SVD) (K=3)**

Accuracy	77.45098039215686%
Precision	0.8222222222222222
Recall	0.7115384615384616
F-measure	0.7628865979381444

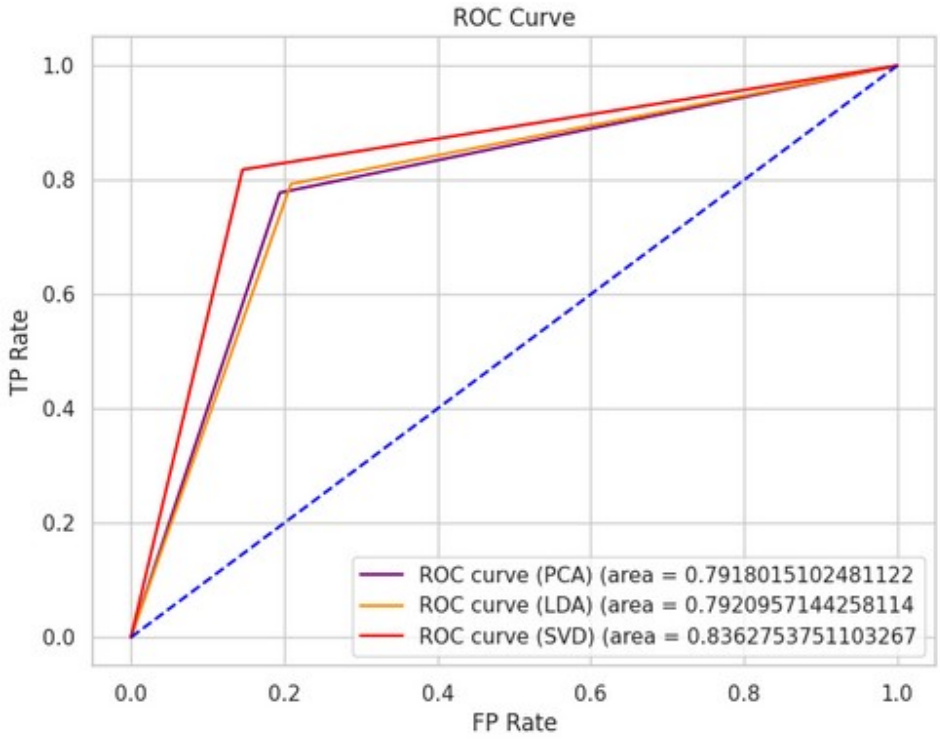

Figure 5. *RoC of KNN Classifiers*

- **KNN (K=3)**

Accuracy	0.7745098039215687
Precision	0.8222222222222222
Recall	0.7115384615384616
F-measure	0.7628865979381444

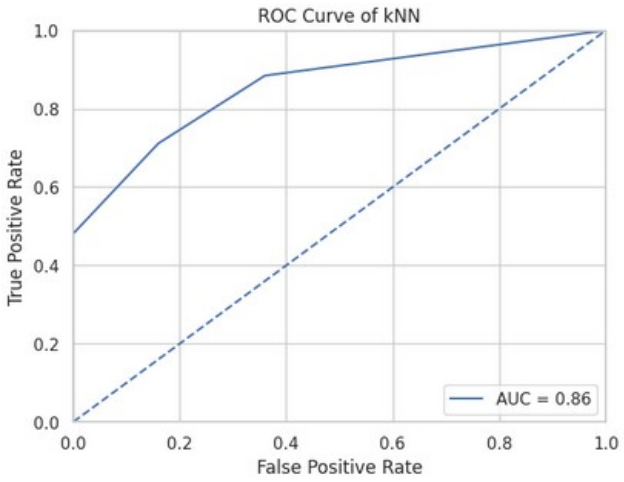

Figure 6. *Roc of Standard KNN*

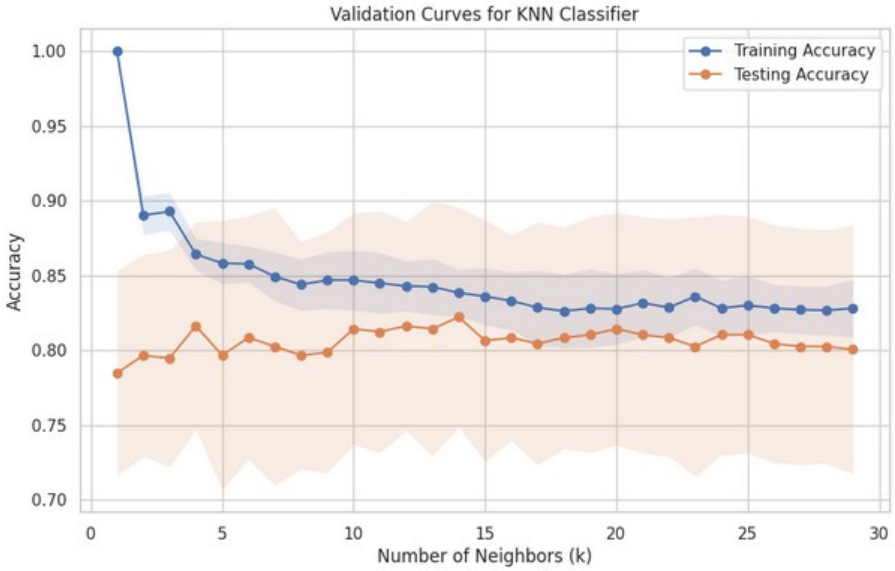

Figure 7. *Validation Curves for KNN*

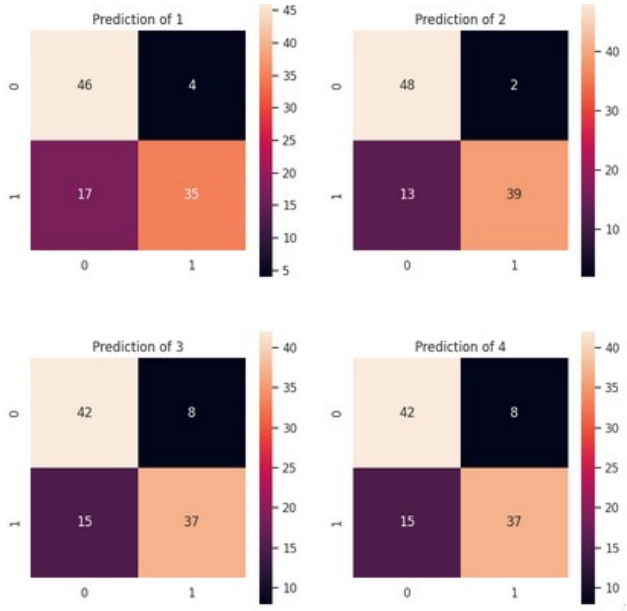

Figure 8. *Confusion Matrix using Heatmap for all KNNs*

b) KNN (Regressor) (K=7)

- **Euclidean Distance**

MAE	6898.0392156862745
MSE	118131412.56502606
RMSE	10868.827561656595

- **Manhattan Distance**

MAE	7085.994397759105
MSE	125210108.04321724
RMSE	11189.732259675262

- **Cosine Distance**

MAE	7223.529411764704
MSE	115955214.08563428
RMSE	10768.250279670987

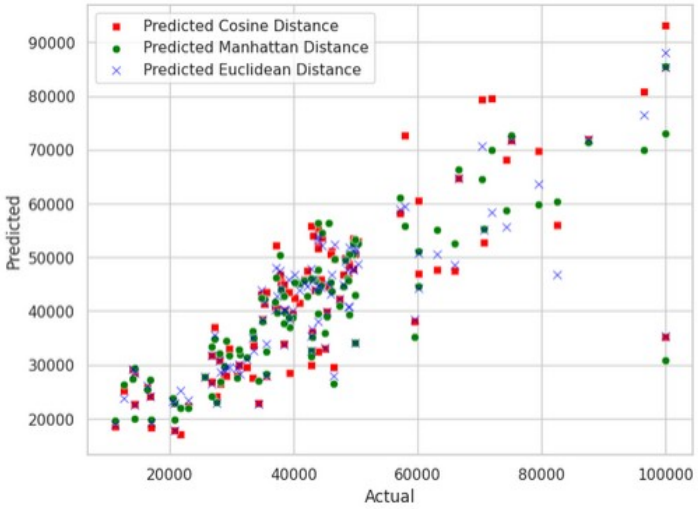

Figure 9. *Scatter Plot of all KNNs Regressors*

c) Naive Bayes

For this model, we need to ensure that the features of the dataset are purely independent on each other. For each feature, it starts by calculating the probability distributions of each feature given the other, based on conditional probability's Bayes' theorem, which is written as follows:

$$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$$

Then, during prediction, it chooses based on the posterior, the one that yields the highest probability as the predicted class for a given instance.

Accuracy	0.7549019607843137
Precision	0.7647058823529411
Recall	0.75
F-measure	0.7572815533980582

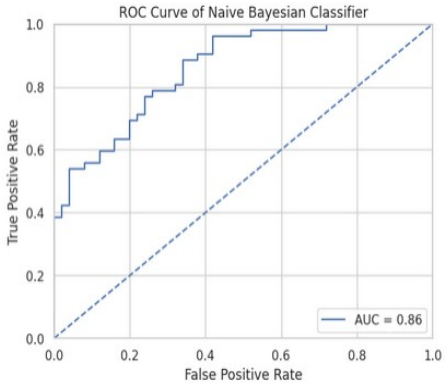

Figure 10. *RoC of Naive Bayes*

d) Decision Tree

In the Decision Tree classifier, the first step is to establish the root node, which consists of all the dataset's observations. Then, the algorithm selects the best feature to split the dataset. In our model, we will use "Entropy" as the criteria, during which we shall choose the feature that generates the most information gain, then for each child node the steps would repeat until it reaches the terminal, which is the leaf nodes of the tree, those that represent the final outcome or decision that would be taken.

Accuracy	0.7745098039215687
Precision	0.9393939393939394
Recall	0.5961538461538461
F-measure	0.7294117647058823

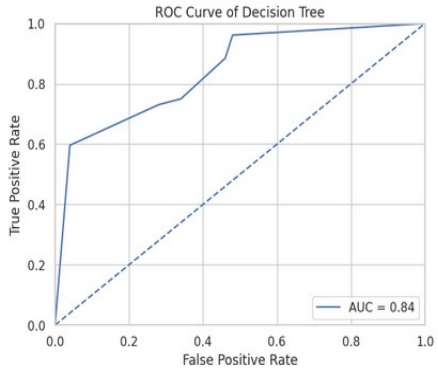

Figure 11. *RoC of Decision Tree*

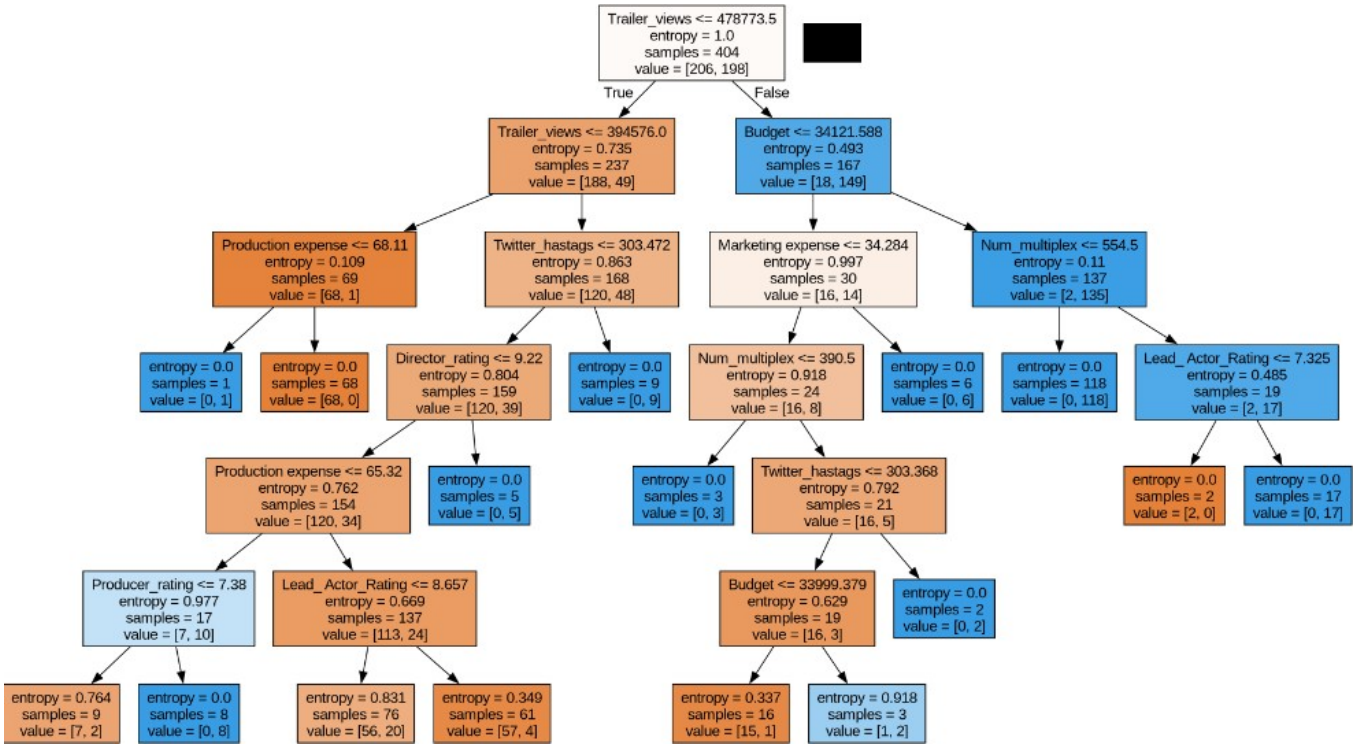

Figure 12. *Decision Tree Graph of Depth 7*

e) Neural Networks

Generally, neural networks models consist of 3 types of layers:

- Input Layer: In this layer, each node represents a feature, and the number of nodes are determined by the input shape (number of features) of the data.
- Hidden Layers: Nodes in this layer perform computations on the input data through activation functions.
- Output Layer: Nodes in this layer produce the final output, either regression or classification.

As such, in our model, a 1024 filter size dense layer (fully connected) acts as our first layer. Then, it passes the output to a 512 filter size dense layer, then to another 256 filter size dense layer and then to the final output layer of size 1. For a total of 672769 parameters to train.

Accuracy	0.8333333134651184

```
Model: "sequential_14"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense_56 (Dense)            (None, 1024)              16384

 dense_57 (Dense)            (None, 512)               524800

 dense_58 (Dense)            (None, 256)               131328

 dense_59 (Dense)            (None, 1)                 257

=================================================================
Total params: 672769 (2.57 MB)
Trainable params: 672769 (2.57 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
Epoch 1/10
41/41 [==============================] - 2s 7ms/step - loss: 0.4401 - accuracy: 0.8050 - val_loss: 0.3486 - val_accuracy: 0.8765
Epoch 2/10
41/41 [==============================] - 0s 4ms/step - loss: 0.3150 - accuracy: 0.8576 - val_loss: 0.3227 - val_accuracy: 0.8642
Epoch 3/10
41/41 [==============================] - 0s 4ms/step - loss: 0.2859 - accuracy: 0.8762 - val_loss: 0.2786 - val_accuracy: 0.8889
Epoch 4/10
41/41 [==============================] - 0s 4ms/step - loss: 0.2816 - accuracy: 0.8793 - val_loss: 0.3455 - val_accuracy: 0.8889
Epoch 5/10
41/41 [==============================] - 0s 4ms/step - loss: 0.2354 - accuracy: 0.8978 - val_loss: 0.3201 - val_accuracy: 0.8889
Epoch 6/10
41/41 [==============================] - 0s 4ms/step - loss: 0.2111 - accuracy: 0.8978 - val_loss: 0.4341 - val_accuracy: 0.8765
Epoch 7/10
41/41 [==============================] - 0s 4ms/step - loss: 0.1740 - accuracy: 0.9164 - val_loss: 0.3892 - val_accuracy: 0.9012
Epoch 8/10
41/41 [==============================] - 0s 4ms/step - loss: 0.1770 - accuracy: 0.9288 - val_loss: 0.3742 - val_accuracy: 0.8889
Epoch 9/10
41/41 [==============================] - 0s 4ms/step - loss: 0.1390 - accuracy: 0.9226 - val_loss: 0.5426 - val_accuracy: 0.8765
Epoch 10/10
41/41 [==============================] - 0s 4ms/step - loss: 0.2167 - accuracy: 0.8885 - val_loss: 0.5432 - val_accuracy: 0.8765
4/4 [==============================] - 0s 3ms/step - loss: 0.5644 - accuracy: 0.8333
Test Accuracy: 0.8333333134651184
```

Figure 12. *ANN using Tensorflow Module*

6. Conclusion and Future Work

The project aimed to forecast a movie's collection using several methods: Naive Bayes, Decision Tree, Artificial Neural Networks, K-NN and . In the exploratory factor analysis, several factors were identified from eighteen variables for regression modeling, while some others were fit for classification, some of which were binned and encoded accordingly. The study suggests that predicting gross revenue during production is not very accurate. The developed models are imperfect, as they do not consider various variables like plot, social media sentiment, stardom, and awards. The use of more advanced techniques, such as random forest, may improve revenue predictions in the future.

Best model in terms of accuracy: **K-NN post-LDA → 0.8529411764705883**

Model	Accuracy
K-NN (PCA)	0.7941176470588235
K-NN (LDA)	0.8529411764705883
K-NN (SVD)	0.7745098039215686
K-NN	0.7745098039215687
LDA	0.8235294117647058
NAIVE BAYES	0.7549019607843137
DECISION TREE	0.7843137254901961
NEURAL NETWORKS	0.8333333134651184

Table 6. Table of All Models' Accuracies

7. References

[1] Dey, S. (2018). Predicting Gross Movie Revenue. *arXiv preprint arXiv:1804.03565*.

[2] Madongo, C. T., & Zhongjun, T. (2023). A movie box office revenue prediction model based on deep multimodal features. *Multimedia Tools and Applications*, 1-29.

[3] Chen, A. W. (2018). A statistical analysis of gross revenue in movie industry. *International Journal of Business Management and Economic Research (IJBMER)*, *9*(3), 1276-1280.

[4] Ahmad, I. S., Bakar, A. A., & Yaakub, M. R. (2020). Movie revenue prediction based on purchase intention mining using YouTube trailer reviews. *Information Processing & Management*, *57*(5), 102278.

[5] Wang, D., Wu, Y., Gu, C., Wang, Y., Zhu, X., Zhou, W., & Lin, X. M. (2022). A movie box office revenues prediction algorithm based on human-machine collaboration feature processing. *Journal of Engineering Research*.

[6] Hao, B. (2023). The Analysis of the Factors that Influence the Film Revenue. *Highlights in Science, Engineering and Technology*, *47*, 154-159.

[7] Wang, Z., Zhang, J., Ji, S., Meng, C., Li, T., & Zheng, Y. (2020). Predicting and ranking box office revenue of movies based on big data. *Information Fusion*, *60*, 25-40.

[8] Gao, Z., Malic, V., Ma, S., & Shih, P. (2019). How to make a successful movie: factor analysis from both financial and critical perspectives. In *Information in Contemporary Society: 14th International Conference, iConference 2019, Washington, DC, USA, March 31–April 3, 2019, Proceedings 14* (pp. 669-678). Springer International Publishing.

[9] Kharb, L., Chahal, D., & Vagisha. (2020). Forecasting Movie Rating Through Data Analytics. In *Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part II 5* (pp. 249-257). Springer Singapore.

[10] Murschetz, P. C., Bruneel, C., Guy, J. L., Haughton, D., Lemercier, N., McLaughlin, M. D., ... & Bakhtawar, B. (2020). Movie Industry Economics: How Data Analytics Can Help Predict Movies' Financial Success. *Nordic Journal of Media Management*, *1*(3), 339-359.