FINAL REPORT

Module #4

820 Culminating Project

Submission Date: Mon- 03Apr2023

Student No.: 500129605

Supervisor: Ceni Babaoglu

Anthony DiValentino
adivalentino@torontomu.ca

Mon-03Apr2023

Introduction

Movies are comprised of thousands of creative and financial decisions. Nobody sets out to make an unsuccessful movie. The movie business is fickle, and it is difficult to predict how a single movie will perform at the box office or streaming. Large financial investments and influential partners are required to produce and market a movie. In addition, there are no guarantees that any one movie will be successful enough to the degree required to make the endeavor financially worthwhile for the various investment interests. Movies create a plethora of relevant, relatively high- quality public- facing data from movie meta data to audience and critic engagement data via social media. With the data that is available for movies, and how people interact and respond to movies, this topic lends itself well to the application of machine learning techniques. This report aims to explore if and to what degree public-facing data can be applied to helping predict the financial success of a movie based on various attributes.

The report outlines the results of three machine learning model that were designed to predict the financial success of movies utilizing IMDB data.

The global box office industry was greater than $21 billion in 2022, with over 5000 movies being released worldwide. The average production budget was $81 million. In 2022, the United States produced over 23,000 movies with the next greatest producer of films being the United Kingdom at approximately 4000. https://www.the-numbers.com/movies/production-countries/#tab=year

To support production decision making, this work aims to optimize outcome predictions by deploying a preliminary machine learning model.

Data Description:

The data description section provides an overview of the strategy and techniques that were utilized in the work. It includes a written and visual description of the data used, the sources of the data, the methods used to collect and analyze the data, and the techniques used to make predictions. The data will be analyzed using machine learning algorithms, and other supporting methods. The objective is for the results of the analysis to be used to make predictions about the outcomes of movie successes with the results of the predictions being used to inform decision-making and strategic planning.

The data Is collected entirely from one source. This IMDB data was obtained from Kaggle.com at the following url:

https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

The original dimensions of the data are 5043L x 28W and contains movies from 1916 to 2016. After removing missing values, 3759 remain. As shown in the table below, of the remaining movies, 3649 were released after 1980 and 107 were released prior.

Post 1980	Pre 1980
3649	107

In the table below, the entire dataset is described after removing movie records with missing values. The table also shows the quantity and datatype of each variable.

```
Int64Index: 3756 entries, 0 to 5042
Data columns (total 28 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Color                 3756 non-null   object
 1   Director Name         3756 non-null   object
 2   # Critic Reviews      3756 non-null   float64
 3   Duration              3756 non-null   float64
 4   # Director Likes      3756 non-null   float64
 5   # Actor 1 Likes       3756 non-null   float64
 6   Actor 2 Name          3756 non-null   object
 7   # Actor 1 Likes       3756 non-null   float64
 8   Gross                 3756 non-null   float64
 9   Genres                3756 non-null   object
 10  Actor 1 Name          3756 non-null   object
 11  Movie Title           3756 non-null   object
 12  # Users Voted         3756 non-null   int64
 13  # Cast Likes          3756 non-null   int64
 14  Actor 3 Name          3756 non-null   object
 15  # FB Poster           3756 non-null   float64
 16  Plot Keywords         3756 non-null   object
 17  Movie Link            3756 non-null   object
 18  # Users for Reviews   3756 non-null   float64
 19  Language              3756 non-null   object
 20  Country               3756 non-null   object
 21  Content Rating        3756 non-null   object
 22  Budget                3756 non-null   float64
 23  Title Year            3756 non-null   float64
 24  # Actor 2 Likes       3756 non-null   float64
 25  IMDB Score            3756 non-null   float64
 26  Aspect Ratio          3756 non-null   float64
 27  # Movie Likes         3756 non-null   int64
dtypes: float64(13), int64(3), object(12)
memory usage: 851.0+ KB
```

The below table provides a statistical description of the features of the IMDB dataset from 1916 to 2016 (after records containing missing values have been removed). It includes a summary of the number of records, the mean, median, mode, min, max, and standard deviation for each feature. The below table is designed to provide an overview of the numerical attributes prior to applying modifications and machine learning techniques.

Mon-03Apr2023

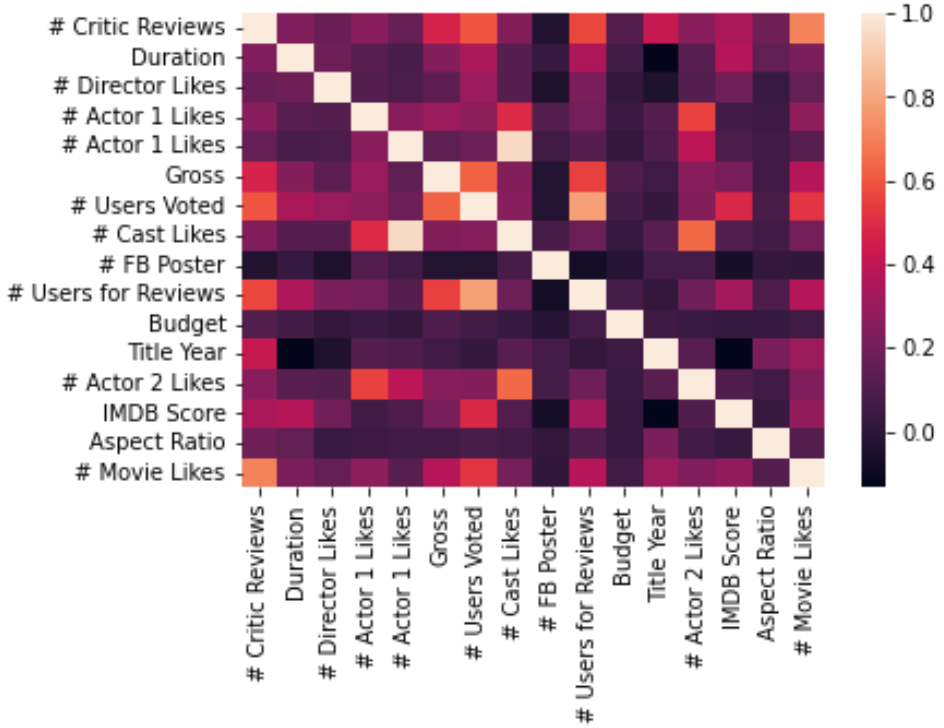
	# Critic Reviews	Duration	# Director Likes	# Actor 3 Likes	# Actor 1 Likes	Gross	# Users Voted	# Cast Likes
count	3756.00	3756.00	3756.00	3756.00	3756.00	3756.00	3756.00	3756.00
mean	167.38	110.26	807.34	771.28	7751.34	52612824.24	105826.73	11527.10
std	123.45	22.65	3068.17	1894.25	15519.34	70317866.91	152035.40	19122.18
min	2.00	37.00	0.00	0.00	0.00	162.00	91.00	0.00
25%	77.00	96.00	11.00	194.00	745.00	8270232.75	19667.00	1919.75
50%	138.50	106.00	64.00	436.00	1000.00	30093107.00	53973.50	4059.50
75%	224.00	120.00	235.00	691.00	13000.00	66881940.75	128602.00	16240.00
max	813.00	330.00	23000.00	23000.00	640000.00	760505847.00	1689764.00	656730.00

# FB Poster	# Users for Reviews	Budget	Title Year	# Actor 2 Likes	IMDB Score	Aspect Ratio	# Movie Likes
3756.00	3756.00	3756.00	3756.00	3756.00	3756.00	3756.00	3756.00
1.38	336.84	46236849.64	2002.98	2021.78	6.47	2.11	9353.83
2.04	411.23	226010288.48	9.89	4544.91	1.06	0.35	21462.89
0.00	4.00	218.00	1927.00	0.00	1.60	1.18	0.00
0.00	110.00	10000000.00	1999.00	384.75	5.90	1.85	0.00
1.00	210.00	25000000.00	2004.00	685.50	6.60	2.35	227.00
2.00	398.25	50000000.00	2010.00	976.00	7.20	2.35	11000.00
43.00	5060.00	12215500000.00	2016.00	137000.00	9.30	16.00	349000.00

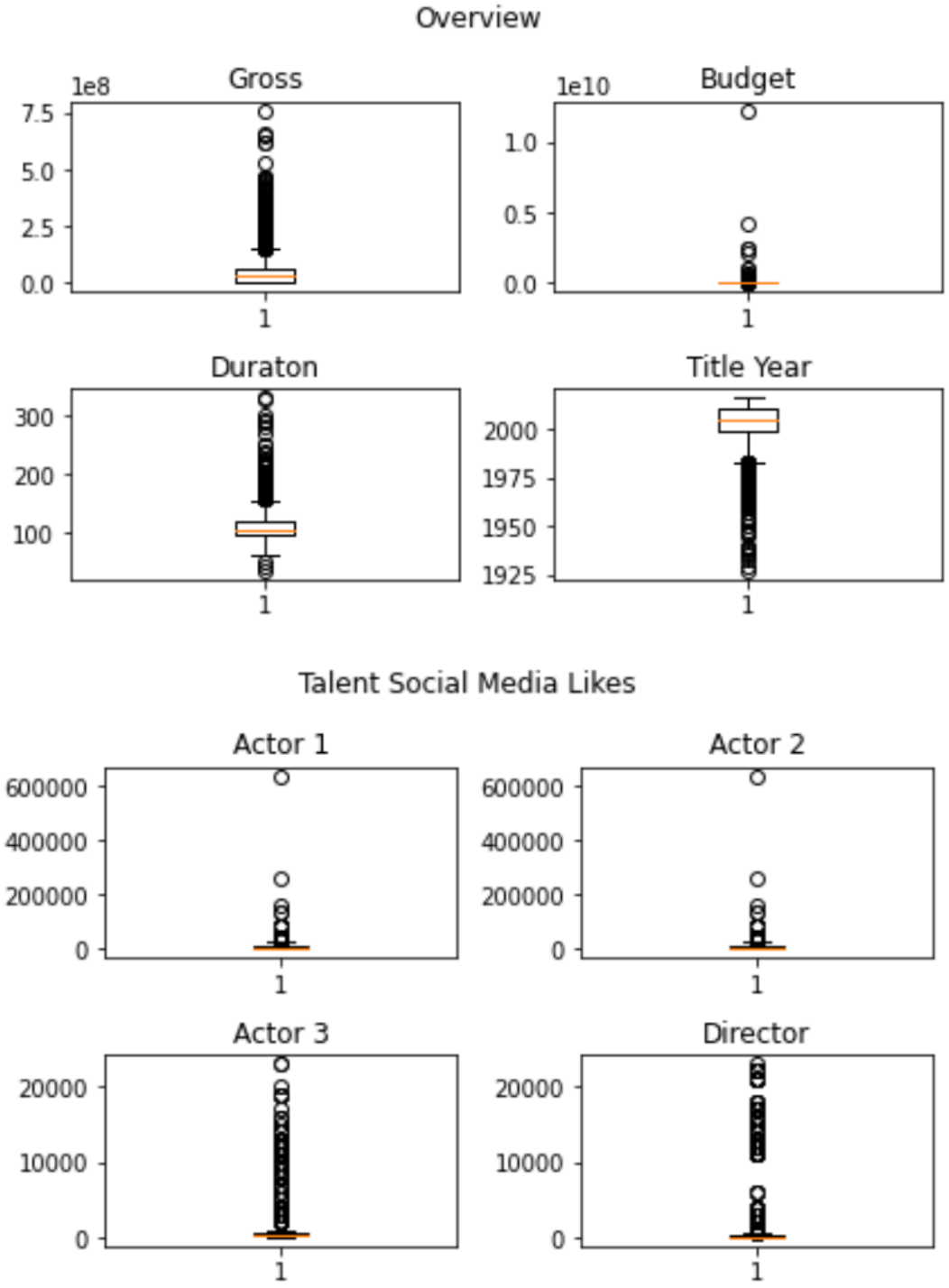
The below correlation table and correlation matrix combination provides an overview into the relationships between the numerical attributes. Utilizing the Pearson method, we can gain insight into which features are positively correlated to one another. The visualizations show that social media attributes are positively correlated with one another in varying degrees. Notably, the number of users who vote on IMDB, and the number of critics who register a a review, are both positively correlated with gross revenue, while the production budget is not significantly correlated with any attribute.

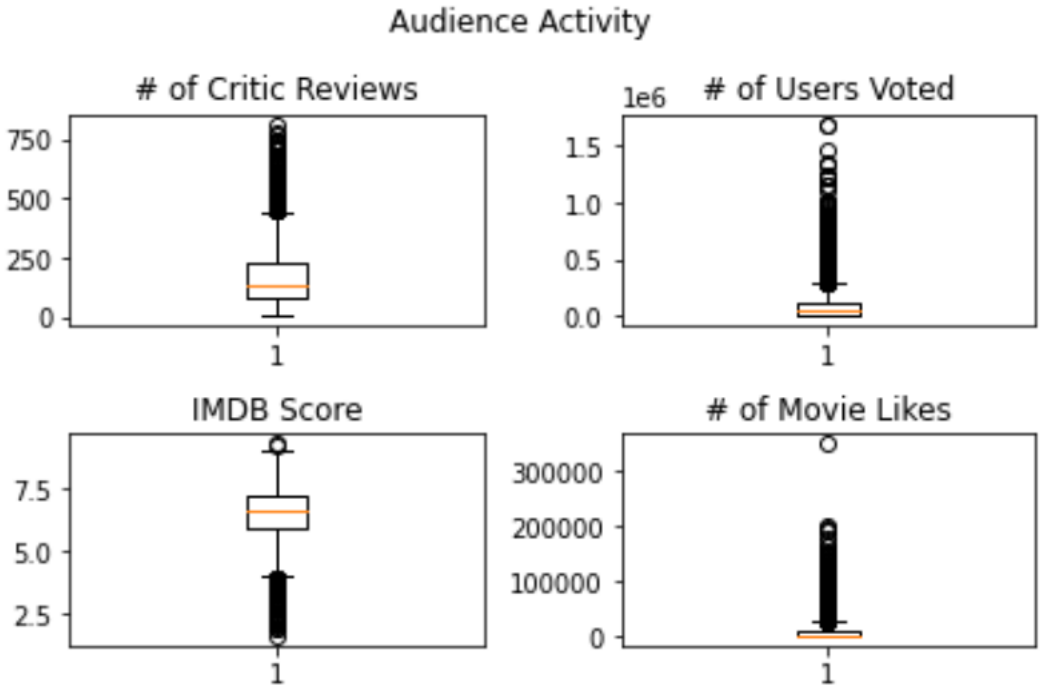
Mon-03Apr2023

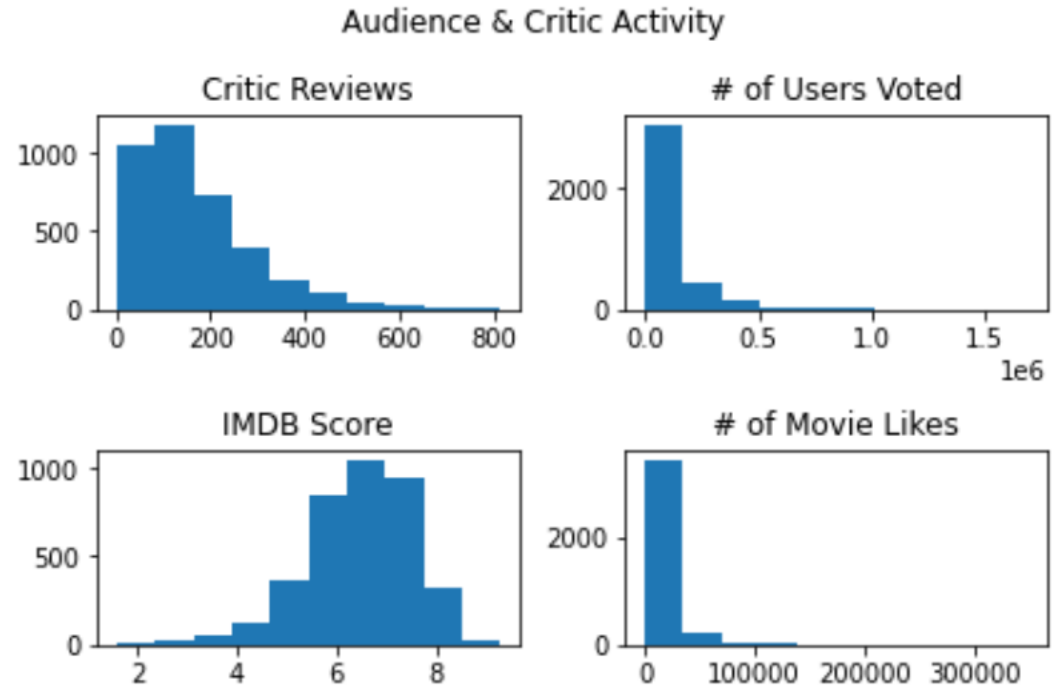
	# Critic Reviews	Duration	# Director Likes	# Actor 3 Likes	# Actor 1 Likes	Gross	# Users Voted	# Cast Likes	# FB Poster	# Users for Reviews	Budget	Title Year	# Actor 2 Likes	IMDB Score	Aspect Ratio	# Movie Likes
# Critic Reviews	1.00	0.23	0.18	0.25	0.17	0.46	0.59	0.24	-0.03	0.56	0.10	0.42	0.25	0.35	0.18	0.71
Duration	0.23	1.00	0.18	0.13	0.08	0.25	0.34	0.12	0.03	0.35	0.07	-0.13	0.13	0.37	0.15	0.22
# Director Likes	0.18	0.18	1.00	0.12	0.09	0.14	0.30	0.12	-0.05	0.22	0.02	-0.04	0.12	0.19	0.04	0.16
# Actor 3 Likes	0.25	0.13	0.12	1.00	0.25	0.30	0.27	0.49	0.11	0.21	0.04	0.12	0.55	0.07	0.05	0.27
# Actor 1 Likes	0.17	0.08	0.09	0.25	1.00	0.14	0.18	0.94	0.06	0.12	0.02	0.10	0.39	0.09	0.06	0.13
Gross	0.46	0.25	0.14	0.30	0.14	1.00	0.62	0.24	-0.03	0.54	0.10	0.05	0.25	0.21	0.06	0.37
# Users Voted	0.59	0.34	0.30	0.27	0.18	0.62	1.00	0.25	-0.03	0.78	0.07	0.02	0.24	0.48	0.08	0.52
# Cast Likes	0.24	0.12	0.12	0.49	0.94	0.24	0.25	1.00	0.08	0.18	0.03	0.13	0.64	0.11	0.07	0.21
# FB Poster	-0.03	0.03	-0.05	0.11	0.06	-0.03	-0.03	0.08	1.00	-0.08	-0.02	0.07	0.07	-0.07	0.02	0.02
# Users for Reviews	0.56	0.35	0.22	0.21	0.12	0.54	0.78	0.18	-0.08	1.00	0.07	0.02	0.19	0.33	0.10	0.37
Budget	0.10	0.07	0.02	0.04	0.02	0.10	0.07	0.03	-0.02	0.07	1.00	0.05	0.04	0.03	0.03	0.05
Title Year	0.42	-0.13	-0.04	0.12	0.10	0.05	0.02	0.13	0.07	0.02	0.05	1.00	0.12	-0.13	0.22	0.31
# Actor 2 Likes	0.25	0.13	0.12	0.55	0.39	0.25	0.24	0.64	0.07	0.19	0.04	0.12	1.00	0.10	0.06	0.23
IMDB Score	0.35	0.37	0.19	0.07	0.09	0.21	0.48	0.11	-0.07	0.33	0.03	-0.13	0.10	1.00	0.03	0.28
Aspect Ratio	0.18	0.15	0.04	0.05	0.06	0.06	0.08	0.07	0.02	0.10	0.03	0.22	0.06	0.03	1.00	0.11
# Movie Likes	0.71	0.22	0.16	0.27	0.13	0.37	0.52	0.21	0.02	0.37	0.05	0.31	0.23	0.28	0.11	1.00



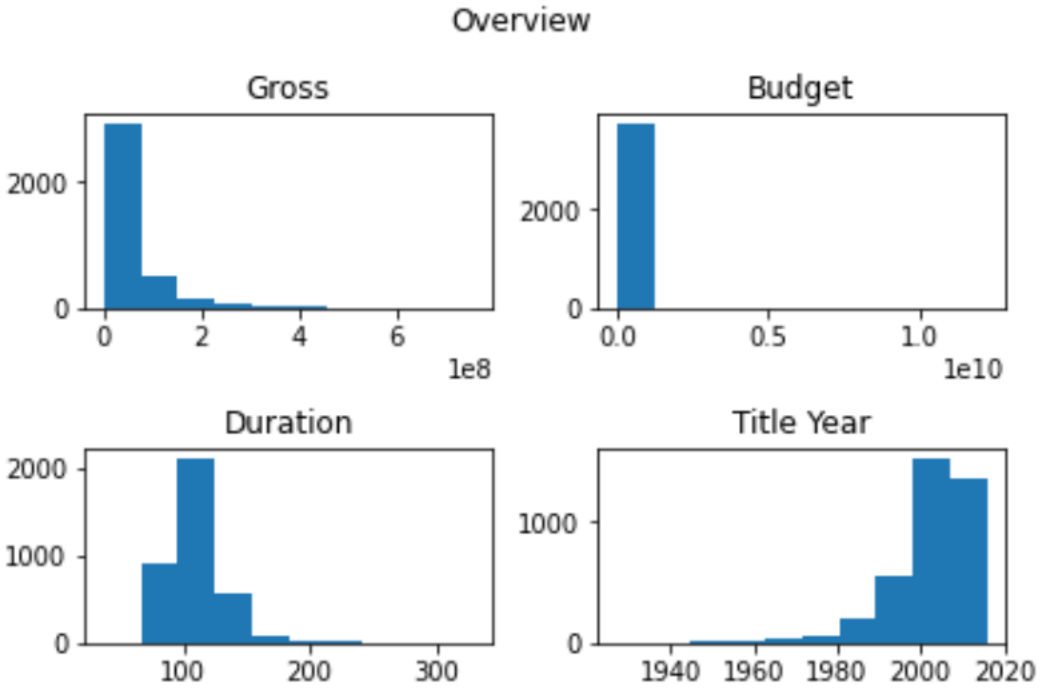
The below compilation of box and whisker plots and histograms are designed to help us visualize the distribution of the dataset's numerical attributes of interest. Utilizing this collection, we can improve understanding of relevant features such as gross revenue, production budget and social media activity.

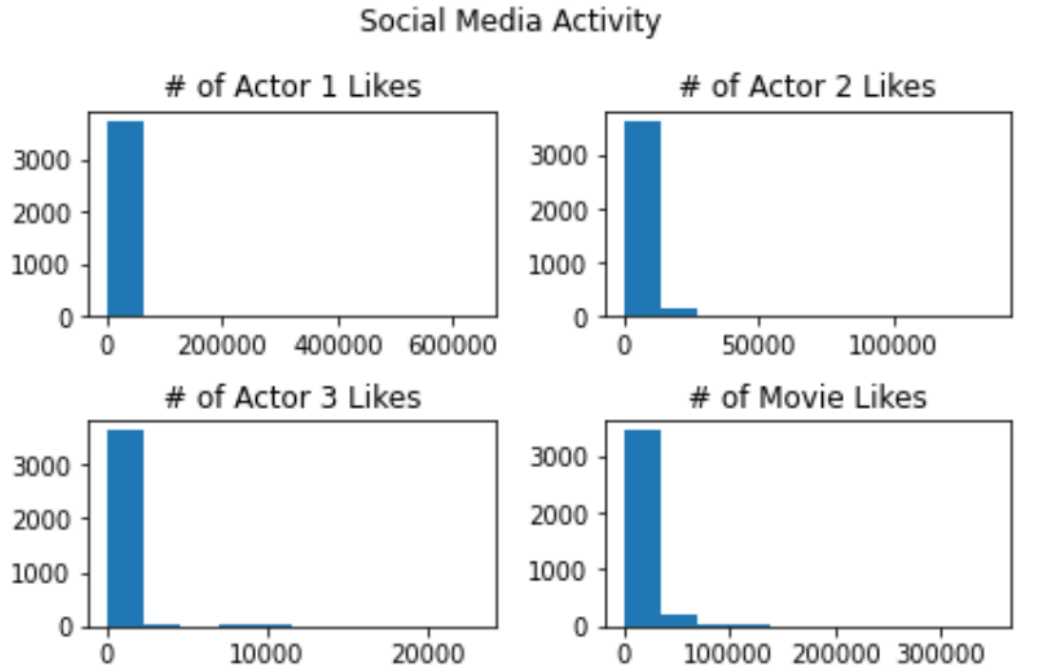

Overview
Gross
1e8
7.5
5.0
2.5
0.0
1
Budget
1e10
1.0
0.5
0.0
1
Duraton
300
200
100
1
Title Year
2000
1975
1950
1925
1
Talent Social Media Likes
Actor 1
600000
400000
200000
0
1
Actor 2
600000
400000
200000
0
1
Actor 3
20000
10000
0
1
Director
20000
10000
0
1


Audience Activity
# of Critic Reviews
750
500
250
0
1
1e6
# of Users Voted
1.5
1.0
0.5
0.0
1
IMDB Score
7.5
5.0
2.5
1
# of Movie Likes
300000
200000
100000
0
1

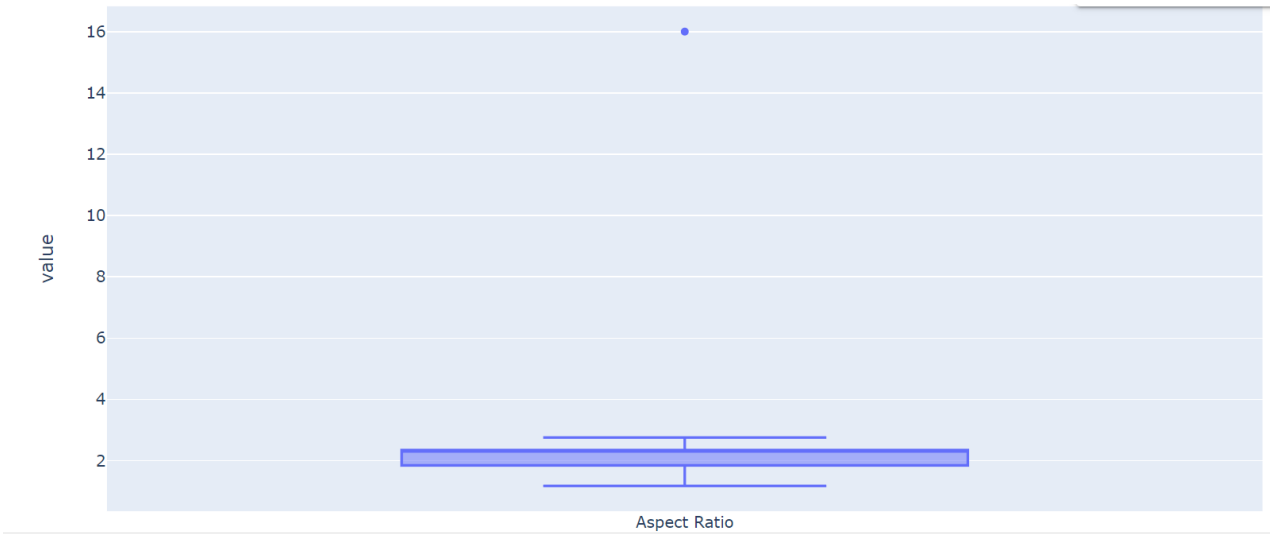

Audience & Critic Activity
Critic Reviews
1000
500
0
0
200
400
600
800
# of Users Voted
2000
0
0.0
0.5
1.0
1.5
1e6
IMDB Score
1000
500
0
2
4
6
8
# of Movie Likes
2000
0
0
100000
200000
300000

Mon-03Apr2023

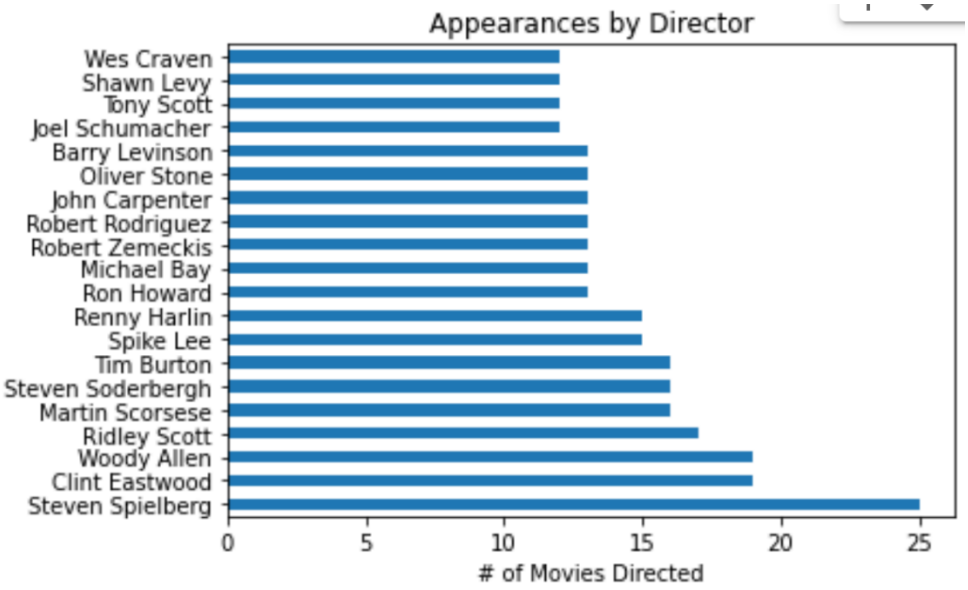




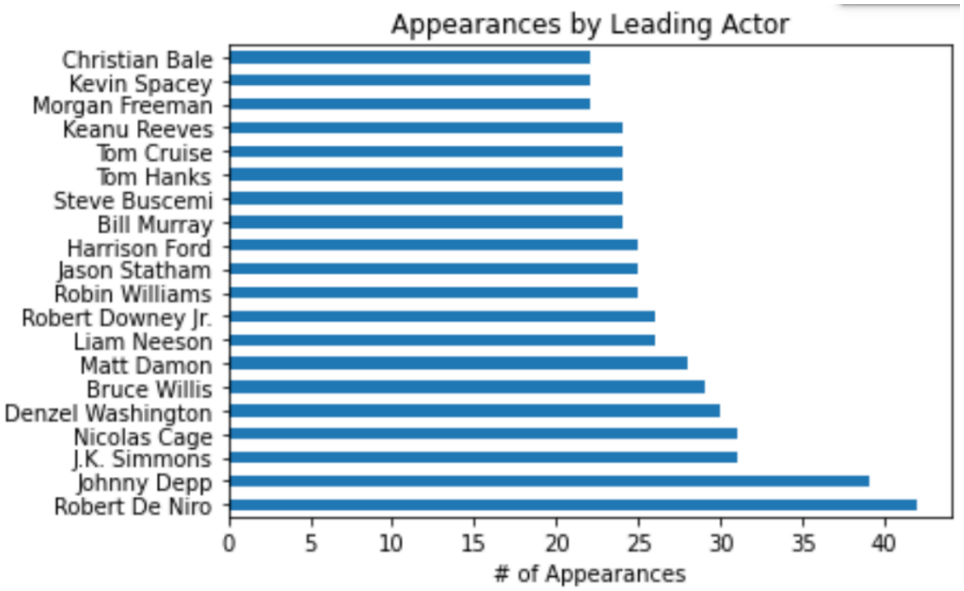
Mon-03Apr2023

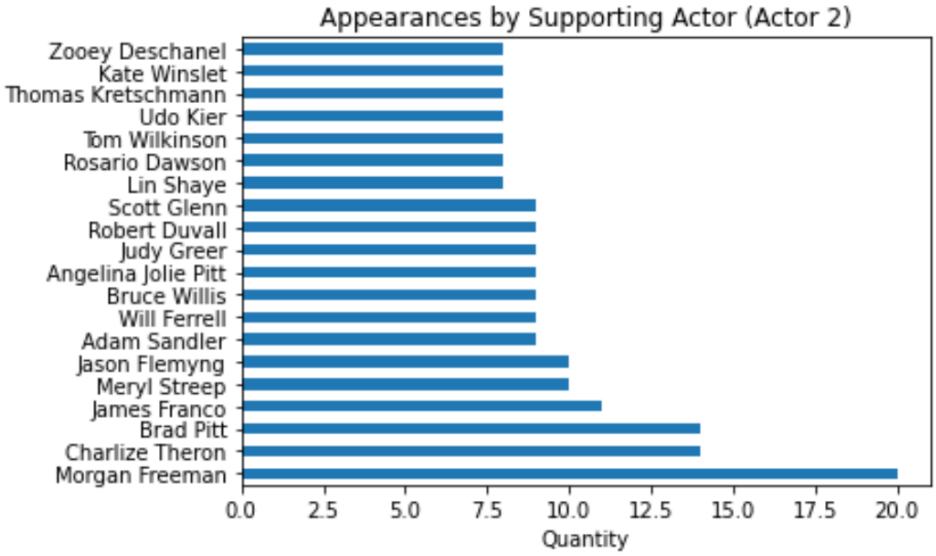


The below compilation of bar plots is designed to help us visualize the categorical attributes of interest. Utilizing this collection of charts, we can gain insight into movie metadata such as personnel- related data. In addition, the below visualizations can assist with understanding which genre and content ratings are most prevalent in the dataset.

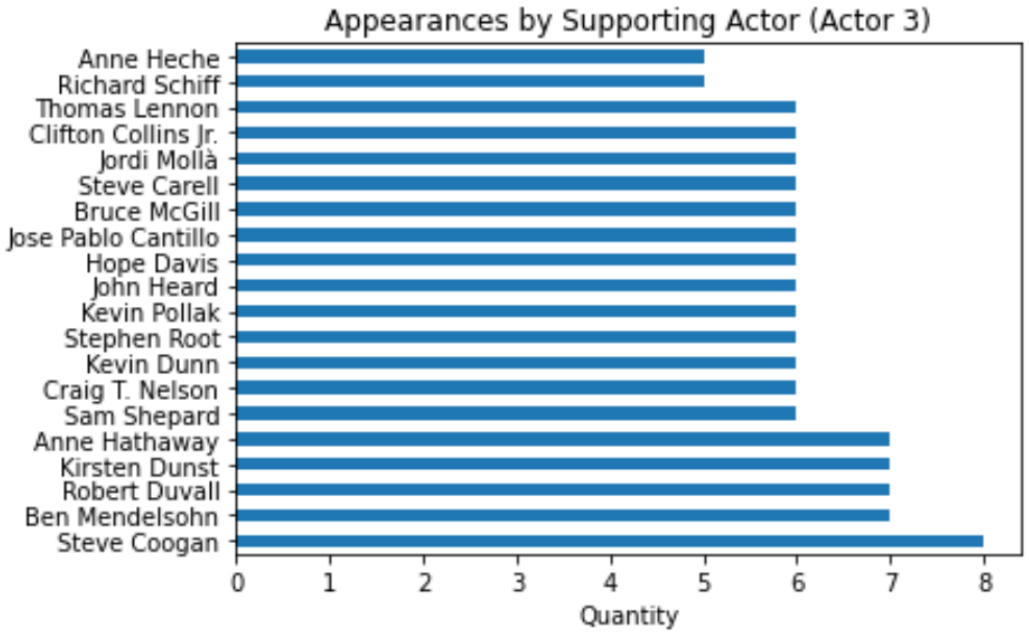


Mon-03Apr2023

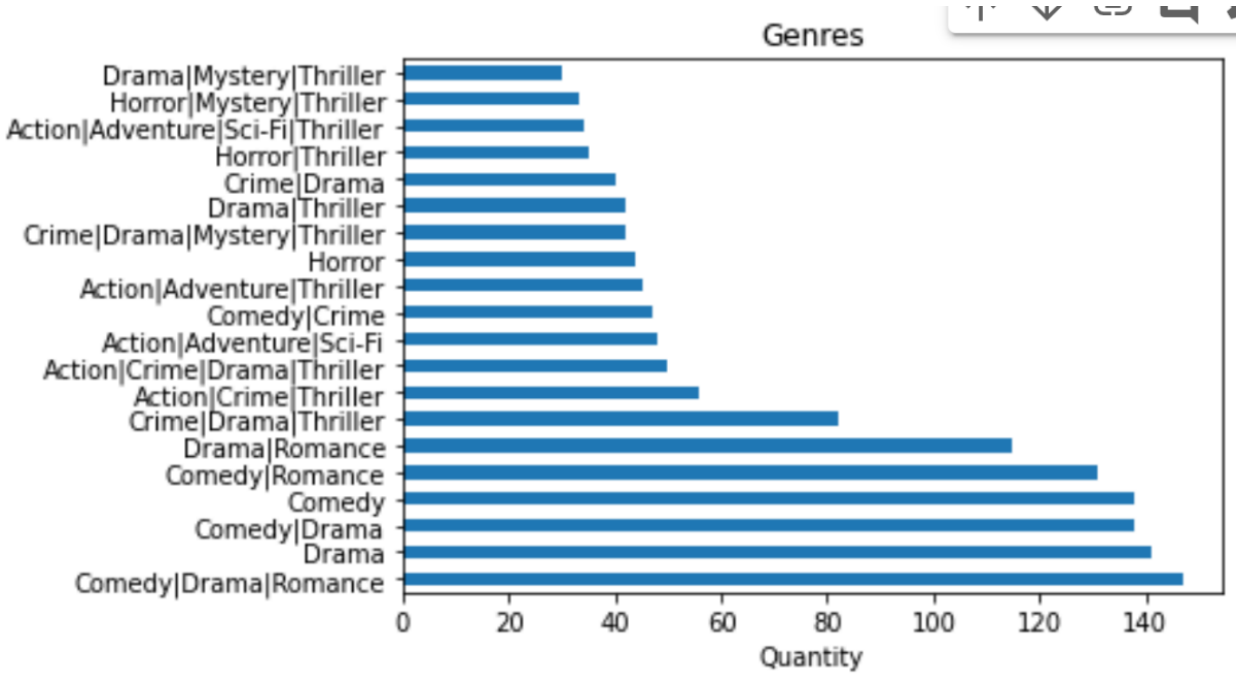


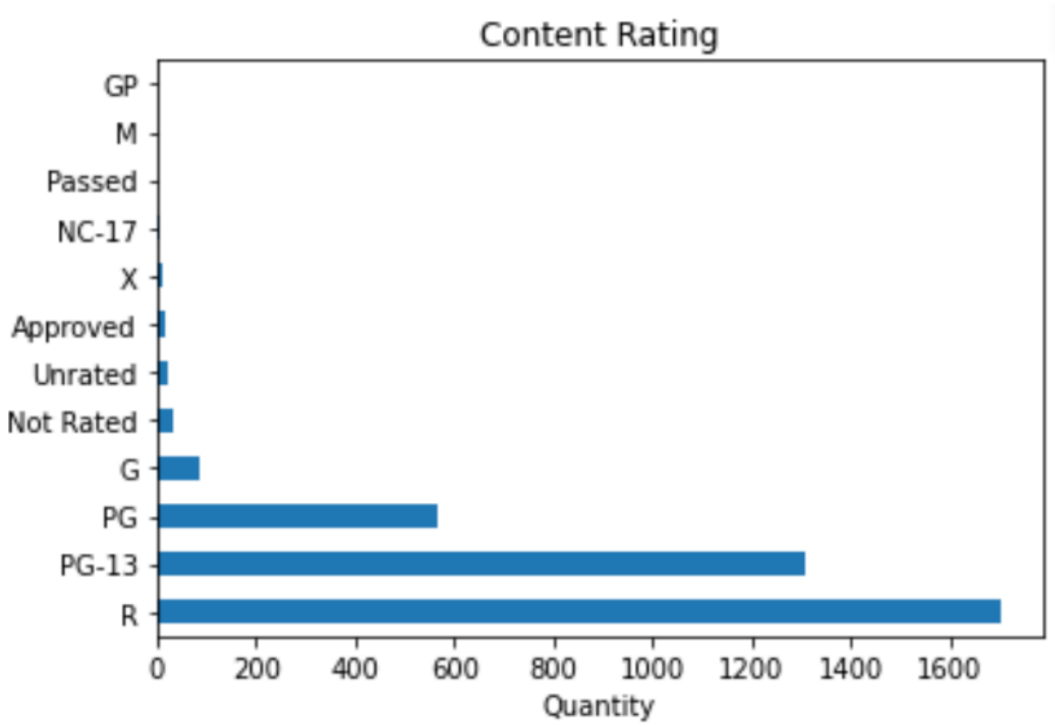


Mon-03Apr2023



Mon-03Apr2023





Mon-03Apr2023

The below tables show the content rating and genre categories with their corresponding gross revenue sum and average IMDB score.

Table 1: Content Rating

Content Rating	Gross	IMDB Score
X	186588116.00	6.54
Unrated	133245086.00	6.95
R	54654725975.00	6.64
Passed	33010612.00	7.13
PG-13	90206375077.00	6.27
PG	43781187904.00	6.30
Not Rated	108020991.00	6.97
NC-17	26861222.00	6.37
M	125108900.00	7.45
GP	43800000.00	6.70
G	7551015555.00	6.50
Approved	763828398.00	7.47

Genre

Genres	Gross	IMDB Score
Action|Adventure|Sci-Fi	9290409191.00	6.65
Comedy|Romance	6290103991.00	5.94
Comedy	6180810524.00	5.85
Comedy|Drama|Romance	5000414623.00	6.49
Action|Adventure|Thriller	4612288238.00	6.75
...	...	...
Crime|Drama|Film-Noir|Mystery|Thriller	7927.00	7.70
Comedy|Fantasy|Thriller	6643.00	6.40
Thriller	2468.00	4.80
Comedy|Fantasy|Musical|Sci-Fi	2436.00	5.70
Crime|Documentary	1111.00	7.70

745 rows × 2 columns

Methodology:

This section outlines the methodology used to build our machine learning model. Below, the data collection methods, data preparation and manipulation methods, testing techniques and evaluation metrics that were utilized are described.

The software application that was used was Google Collaboratory, leveraging Python libraries such as sci-kit learn, pandas, matplotlib, plotly, and seaborn. The commented source code is stored as a .ipynb file and can be accessed via the corresponding GitHub repository. Below is a table outlining each core library and corresponding application within the IDE for the purpose of this work.

Library	**Application**
GitHub repository: https://github.com/adivalentino/820.git	
Sci-kit learn	Machine learning algorithm processors, Testing and evaluation tools, preprocessing tools, visualizations,
Pandas	Structuring the data, exploring the data, modifying the data
Matplotlib/Plotly/ Seaborn	Visualizing the data

Mon-03Apr2023

Os	Import Google Drive to IDE

Publicly available IMDB data obtained from Kaggle.com was utilized to build the model. Data for genres, release dates, production budgets, box office revenue, ratings, reviews, and social media engagement activity were collected, analyzed, and applied. Exploratory data analysis was performed after the dataset was cleaned to remove missing values, while outliers were left in the dataset. Feature engineering techniques were applied to engineer a new categorical feature based on values of existing features. In addition, features that were thought to be irrelevant to this project were removed from the dataset.

Three different models were used to predict the financial success of movies. The algorithms were Gaussian Naïve Bayes, Support Vector Machine and Decision Tree. For each model, the same train-test split partitioning was applied to create the required subsets. Pre-processing strategy was applied to encode text- based values with a corresponding numerical value. In addition, feature scaling was applied to provide a normalized range of variables and ensure that all features are given equal weightage in the analysis, while reducing the risk of a particular feature having numerical dominance over another.

Two testing and evaluation techniques were applied. The first was utilizing a standard 80/20 test-train data split, with the second being a 10-fold cross- validation method. The subsets were trained with each model and evaluated against the test sets using accuracy scores to compare the performance of the model relative to one another.

The models were trained on data from over 3756 movies and were tested on various subsets of the data. The model was then used to predict the gross revenue class of a set of movies that have already been released. The results of the models are presented in this report, accompanied by visualizations to display the outcomes. Additionally, the report provides an analysis of the model performance on the data. Finally, the report includes an assessment of the model's strengths and future considerations for improving the work.

The methodology and approach are further detailed below under the following headings:

a. Data Preparation
b. Feature Engineering
    a. Label Encoding
    b. Scaling and normalization
c. Testing and Evaluation

Data Preparation:

Outlining a critical step in the machine learning process, this report is designed to provide a technical overview of the methods utilized to prepare the data set for machine learning applications. The scope of the work outlined below includes, removing records with missing values, extracting unwanted features from the original dataset, and creating a new categorical variable to support our movie revenue classification problem. The report aims at showing the feature engineering methods such as

normalization, encoding and data cleaning that were applied to the data prior to applying the selected machine learning algorithms.

The data was obtained from Kaggle.com. The full URL is accessible in the data description section.

The below table outlines the before and after removing missing values from the original dataset. The left portion of the table count the quantity of missing values in the original dataset, while the right portion of the table confirms there aren't any missing values after the cleaning technique is applied.

Color 19	Color 0
Director Name 104	Director Name 0
# Critic Reviews 50	# Critic Reviews 0
Duration 15	Duration 0
# Director Likes 104	# Director Likes 0
# Actor 1 Likes 23	# Actor 1 Likes 0
Actor 2 Name 13	Actor 2 Name 0
# Actor 1 Likes 7	# Actor 1 Likes 0
Gross 884	Gross 0
Genres 0	Genres 0
Actor 1 Name 7	Actor 1 Name 0
Movie Title 0	Movie Title 0
# Users Voted 0	# Users Voted 0
# Cast Likes 0	# Cast Likes 0
Actor 3 Name 23	Actor 3 Name 0
# FB Poster 13	# FB Poster 0
Plot Keywords 153	Plot Keywords 0
Movie Link 0	Movie Link 0
# Users for Reviews 21	# Users for Reviews 0
Language 12	Language 0
Country 5	Country 0
Content Rating 303	Content Rating 0
Budget 492	Budget 0
Title Year 108	Title Year 0
# Actor 2 Likes 13	# Actor 2 Likes 0
IMDB Score 0	IMDB Score 0
Aspect Ratio 329	Aspect Ratio 0
# Movie Likes 0	# Movie Likes 0

Feature Engineering:

To utilize the information that is most relevant to the problem at hand, feature extraction techniques were applied. The following features were removed from the dataset to prepare for training the various machine learning algorithms.

a. Plot Keywords
b. Movie Link
c. Movie Title

New Feature Created:

Mon-03Apr2023

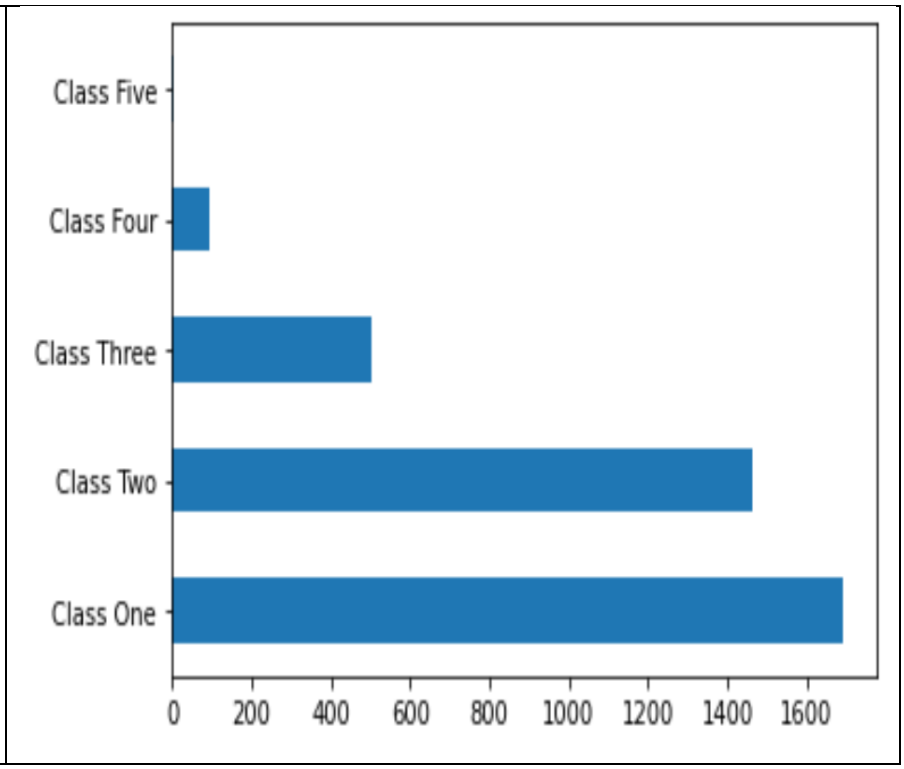
To support our classification problem, a new feature was created. The gross revenue variable was assigned a corresponding revenue category between 1 and 5. The below table shows the first 20 movies with their newly assigned corresponding gross revenue category. In addition, the gross revenue category ranges are displayed.

	Movie Title	Classes
0	Avatar	Class Five
1	Pirates of the Caribbean: At World's End	Class Four
2	Spectre	Class Three
3	The Dark Knight Rises	Class Four
5	John Carter	Class Two
6	Spider-Man 3	Class Four
7	Tangled	Class Three
8	Avengers: Age of Ultron	Class Four
9	Harry Potter and the Half-Blood Prince	Class Four
10	Batman v Superman: Dawn of Justice	Class Four
11	Superman Returns	Class Three
12	Quantum of Solace	Class Three
13	Pirates of the Caribbean: Dead Man's Chest	Class Four
14	The Lone Ranger	Class Two
15	Man of Steel	Class Four
16	The Chronicles of Narnia: Prince Caspian	Class Three
17	The Avengers	Class Five
18	Pirates of the Caribbean: On Stranger Tides	Class Three
19	Men in Black 3	Class Three
20	The Hobbit: The Battle of the Five Armies	Class Four

	Revenue	Class
0	$0 - 24.99M	Class 1
1	$25 - 99.99M	Class 2
2	$100 - 249.99M	Class 3
3	$250 - 499.99M	Class 4
4	$500 - 1.0B	Class 5

Mon-03Apr2023

	Classes
Class One	1692
Class Two	1463
Class Three	502
Class Four	93
Class Five	6



Label Encoding:

Label encoding was utilized to ensure the text- based data can be leveraged within our machine learning model. The below tables show a sample of the dataset after label encoding and numerical scaling techniques were applied to text- based data.

	Color	Director Name	# Critic Reviews	Duration	# Director Likes	# Actor 3 Likes	Actor 2 Name	# Actor 1 Likes	Gross	Genres	...	# Users for Reviews
0	1	620	723.00	178.00	0.00	855.00	1002	1000.00	760505847.00	91	...	3054.00
1	1	538	302.00	169.00	563.00	1000.00	1592	40000.00	309404152.00	85	...	1238.00
2	1	1395	602.00	148.00	0.00	161.00	1795	11000.00	200074175.00	107	...	994.00
3	1	251	813.00	164.00	22000.00	23000.00	381	27000.00	448130642.00	243	...	2701.00
5	1	62	462.00	132.00	475.00	530.00	1837	640.00	73058679.00	105	...	738.00

5 rows × 25 columns

Language	Country	Content Rating	Budget	Title Year	# Actor 2 Likes	IMDB Score	Aspect Ratio	# Movie Likes	Classes
9	43	7	237000000.00	66	936.00	7.90	1.78	33000	Class 5
9	43	7	300000000.00	64	5000.00	7.10	2.35	0	Class 4
9	42	7	245000000.00	72	393.00	6.80	2.35	85000	Class 3
9	43	7	250000000.00	69	23000.00	8.50	2.35	164000	Class 4
9	43	7	263700000.00	69	632.00	6.60	2.35	24000	Class 2

Normalization:

The dataset contains numerical data that varies greatly. To ensure all values are in the same relative range, normalization techniques were applied to the data. The below tables show a sample of the data after it has been manipulated into a consistent manner.

	Color	Director Name	# Critic Reviews	Duration	# Director Likes	# Actor 3 Likes	Actor 2 Name	# Actor 1 Likes	Gross	Genres	...	# Users for Reviews
0	1.00	0.37	0.89	0.48	0.00	0.04	0.46	0.00	1.00	0.12	...	0.60
1	1.00	0.32	0.37	0.45	0.02	0.04	0.73	0.06	0.41	0.11	...	0.24
2	1.00	0.84	0.74	0.38	0.00	0.01	0.82	0.02	0.26	0.14	...	0.20
3	1.00	0.15	1.00	0.43	0.96	1.00	0.17	0.04	0.59	0.33	...	0.53
4	1.00	0.04	0.57	0.32	0.02	0.02	0.84	0.00	0.10	0.14	...	0.15

5 rows × 25 columns

Language	Country	Content Rating	Budget	Title Year	# Actor 2 Likes	IMDB Score	Aspect Ratio	# Movie Likes
0.27	0.98	0.64	0.02	0.90	0.01	0.82	0.04	0.09
0.27	0.98	0.64	0.02	0.88	0.04	0.71	0.08	0.00
0.27	0.95	0.64	0.02	0.99	0.00	0.68	0.08	0.24
0.27	0.98	0.64	0.02	0.95	0.17	0.90	0.08	0.47
0.27	0.98	0.64	0.02	0.95	0.00	0.65	0.08	0.07

Mon-03Apr2023

Evaluation and Results:

This section provides an overview of the machine learning testing and evaluation techniques that were applied within our movie revenue classification project. In the below sections, methods used for splitting the data into training and test sets are outlined. In addition, 3- different machine learning algorithms are applied against multiple splitting methods. Confusion matrixes were generated to visualize the performance of each machine learning algorithm.

The evaluation and results section are designed to provide a description of the methods and techniques utilized to assess the performance of our predictive model. In addition, it will set the foundation for our comparative analysis with past works that comes up later in the report. The models were evaluated using accuracy scores to assess the model's ability to accurately predict the financial success of a movie.

Model performance was assessed through comparison with the existing models and alternate approaches that were used in this work. The results of the evaluation and comparison will be used to measure the effectiveness of our predictive model and inform performance improvement decisions-making.

Experiment #1- Naïve Bayes

Using Gaussian Naïve Bayes with 10- fold cross validation, the model performed very poorly, predicting multiclass categorization at approximately a 30% rate. The below array displays each of accuracy score of each of the 10 tests that were taken.

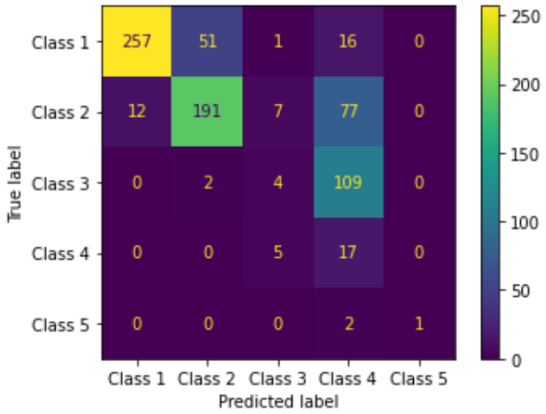
```
array([0.30263158, 0.31578947, 0.28 , 0.25333333, 0.24 , 0.33333333, 0.28
, 0.29333333, 0.28 , 0.32 ]
```

Cross- Validated: 0.2898

However, when using an 80/20 test-train split on the same data, the model predicted at a 62.5% rate.

80/20 Split: 0.625

In the confusion matrix below, the predictions generated by naïve bayes with traditional 80/20 split are evaluated against the class variable.



Mon-03Apr2023

Experiment #2- Support Vector Machine

Using Support Vector Machine, the model performed significantly better than Gaussian Naïve Bayes. SVM performed it' multiclass categorization at approximately an 80% rate in its first test while incrementally decreasing in accuracy for each subsequent test that was performed. The first two tests in the cross- validation step resulted in 78% and 82%, respectively, while the final test resulted in an accuracy score of 69%. The below array displays the accuracy scores of each of the 10 tests that were taken.

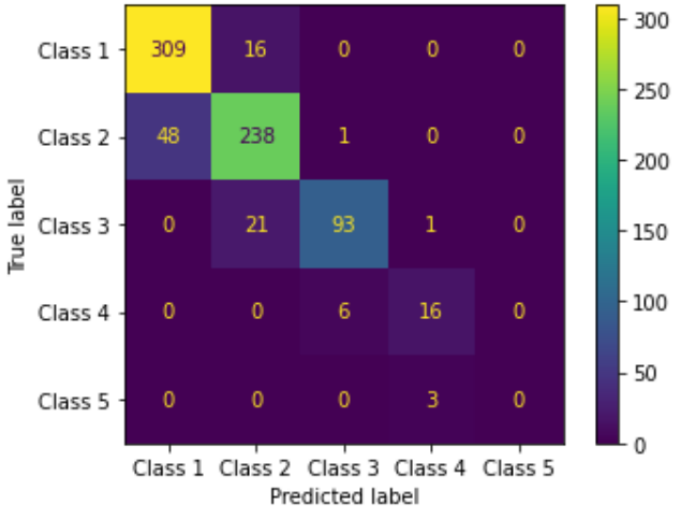
```
array([0.78947368, 0.82894737, 0.72 , 0.69333333, 0.70666667, 0.62666667,
0.61333333, 0.82666667, 0.70666667, 0.69333333])
```

Cross- Validated: 0.7205

Using an 80/20 test-train split on the same data, the model predicted at an 87.5% rate.

80/20 Split: 0.8723

In the confusion matrix below, the predictions generated by support vector machine with traditional 80/20 split are evaluated against the class variable.



Experiment #3- Decision Tree

Using **Decision Tree**, the model performed better than both Support Vector Machine and Naïve Bayes. The accuracy scores for the 10 tests ranged from 98-100%. The below array displays the accuracy scores of each of the 10 tests that were taken.

```
array([1. , 0.98684211, 1. , 1. , 1. , 1. , 1. , 1. , 1. , 1. ])
```

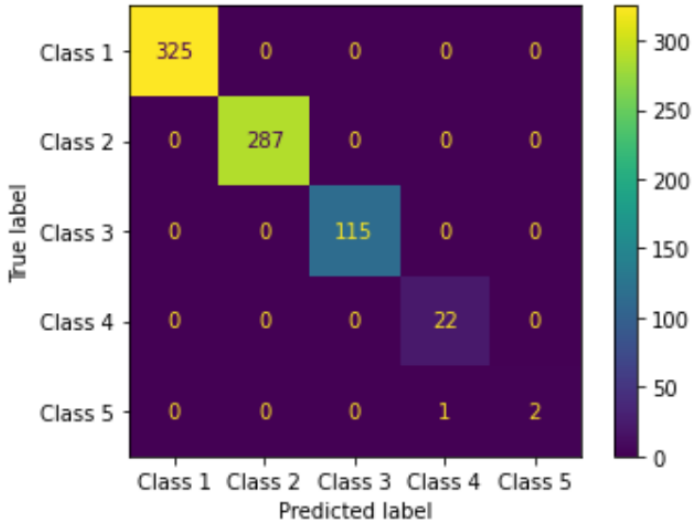
Cross- Validated: 0.99868

Using an 80/20 train-test split, the decision tree model performed at 99.9% success rate.

80/20 Split 0.99867

Mon-03Apr2023

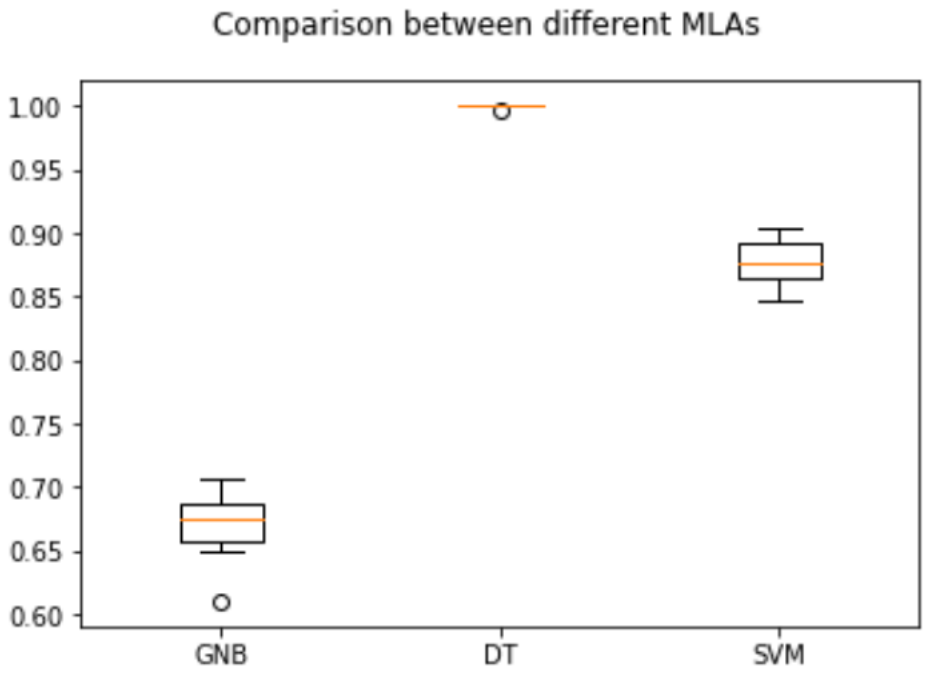
In the confusion matrix below, the predictions generated by decision tree with traditional 80/20 split are evaluated against the class variable.



Experiment #4- Evaluate all models simultaneously:

Using a looping function combined with a 10- fold cross validation, Naïve Bayes, Support Vector Machine and Decision Tree were all executed simultaneously and captured into one object. Accuracy scores and standard deviation of the accuracy scores for each algorithm are displayed below:

```
GNB: 0.669769(0.025730)
 DT: 0.999667(0.001000)
SVM: 0.875826(0.018746)
```



Model	Train Time *(seconds)*	Running Time (*seconds)*
Gaussian Naïve Bayes	0.017716407775878906	0.0037841796875
Support Vector Machine	0.5163547992706299	0.15779733657836914
Decision Tree	0.023783206939697266	0.0017905235290527344

The methodology and approach section presents an overview of the data transformation steps taken to the original IMDB dataset to prepare it for use in our classification- based machine learning model. Four experiments were conducted utilizing three different machine learning models.

Relative to Past and Future Works:

This is an area where extensive research has been conducted to utilize supervised machine learning techniques to predict movie successes. Some research focuses on predicting movie user ratings, while others focused on financial outcomes. In all existing works, the researchers utilized and compared multiple machine learning algorithm and methods to explore relative efficacy, which is what our work did as well.

B. Çizmeci and Ş. G. Ögüdücü, "Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media" built a model that predicted IMDB ratings of pre-release films based on available social media data. Specific accuracy scores were not reported although factorization machines had a higher accuracy compared to logistic regression and naïve bayes. The report did not provide evaluation metrics. The main difference in this existing work and our work is their model is designed to predict financial outcomes based on social media engagement activity from users, in advance of the release date. The social media data that we use does not distinguish between user data that was entered pre-release date vs. post- release date.

A. Bhave, H. Kulkarni, V. Biramane and P. Kosamkar's paper, "Role of different factors in predicting movie success," uses a data set of 220 Bollywood movies released in the last decade. The findings of this study suggest that genre, lead actor, budget and release data are critical factors in the success of a movie. The predictive accuracy scores of this study ranged from 0.61 to 0.73 for different models. The highest accuracy score of 0.73 was achieved for a model that combined genre, lead actor, budget, and release date as predictors. The main difference between this work and our work is their model explores which features were most impactful in successful predictions, and which features contribute less. This could help with informing real world application, model deployment on scale where operators would have to consider and balance model training and running time with predictive power.

David Opeoluwa Oyewola and Emmanuel Gbenga Dada (2022) explored the use of machine learning methods to predict the popularity of movies. The report utilized a range of techniques including support vector machines, linear regression, and decision trees. They used two datasets, one from Hollywood and one from Bangladesh. The predictive accuracy scores for the machine learning techniques utilized from 85-90%, with the highest being support vector machine at ~90%. Support vector machine logged an accuracy score of 95.4% for the Bollywood movie subset and 92.4% on the Hollywood movie subset.

W. R. Bristi, Z. Zaman and N. Sultana's paper, "Predicting IMDb Rating of Movies by Machine Learning Techniques," evaluated the performance of linear regression, decision tree and support vector machine models for predicting movie ratings of movie genres. The dataset used consisted of 4,814 movies and 12 attributes. Their results show that the decision tree random forest model outperformed the others, with an accuracy of 85% in successful predictions. This work is similar to our work with the main difference being our decision tree accuracy score was %99.99 with a 10-fold cross- validation, on a dataset of 3756 by 28 attributes.

N. Quader, M. O. Gani and D. Chaki's paper, "Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction," examines and compares the performance of support vector machines, Random Forests, Naïve Bayes, Gradient Boosting, Artificial Neural Networks, K-Nearest Neighbors and Decision Trees. Movies from 2012 to 2016 were used. The results show that Random Forests and Gradient Boosting produce the most accurate predictions, while K-Nearest Neighbors and Support Vector Machines had the lowest training and running times. This work reports that Random Forests had the highest accuracy score of 97.90%, while the lowest being K's Nearest Neighbours at 91.54%. This work compares the most algorithms. In addition, they use a more concise range of release dates whereas our work utilizes a range of release dates of 100 years. This may distort the data and using a smaller, more recent release date range would allow us to compare movies of the same era.

Limitations of the work:

Our work is limited by the data we have which is from the past. This means our model may not be able to accurately predict current trends or future outcomes. Additionally, our classes may be too wide ranged, leading to distorted data when comparing a wide range of release dates. Inflation and technology factors in different eras can distort the data. To address this, we suggest narrowing the range of release dates to get more accurate results.

Future Work Considerations:

Future work considerations include utilizing more precise gross revenue class variable buckets, incorporating text-based analysis on features such as movie titles and keywords, predicting IMDB movie ratings, collecting social media, critic and user activity prior to the movie being released, and using a more confined range of release dates. There is opportunity to include assessment of which features were relatively helpful in successful movie predictions.

Conclusion:

Mon-03Apr2023

This report shows that our decision tree machine learning model had the highest accuracy at 99.99% in predicting the financial success of movies from IMDB data. Support vector machine's performance ranged the most with a low of 28% when evaluated with a 10-fold cross- validation. However, when executed simultaneously alongside decision tree and support vector machine in Experiment #4, it performed at an accuracy rate of 87.5%. In addition, the train time for decision tree is second best while decision tree's run time being the best. The model displayed strengths and weaknesses, and opportunities for model enhancement were identified.

Mon-03Apr2023

Works Cited

[1] B. Çizmeci and Ş. G. Ögüdücü, "Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media," 2018 3rd International Conference on Computer Science and Engineering (UBMK), Sarajevo, Bosnia and Herzegovina, 2018, pp. 173-178, doi: 10.1109/UBMK.2018.8566661.

[2] A. Bhave, H. Kulkarni, V. Biramane and P. Kosamkar, "Role of different factors in predicting movie success," 2015 International Conference on Pervasive Computing (ICPC), Pune, India, 2015, pp. 1-4, doi: 10.1109/PERVASIVE.2015.7087152.

[3] David Opeoluwa Oyewola, Emmanuel Gbenga Dada (2022). Machine Learning Methods for Predicting the Popularity of Movies. Journal of Artificial Intelligence and Systems, 4, 65–82. https://doi.org/10.33969/AIS.2022040105.

[4] N. Quader, M. O. Gani, D. Chaki and M. H. Ali, "A machine learning approach to predict movie box-office success," 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2017, pp. 1-7, doi: 10.1109/ICCITECHN.2017. 8281839..

[5] W. R. Bristi, Z. Zaman and N. Sultana, "Predicting IMDb Rating of Movies by Machine Learning Techniques," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-5, doi: 10.1109/ICCCNT45670.2019.8944604.

[6] N. Quader, M. O. Gani and D. Chaki, "Performance evaluation of seven machine learning classification techniques for movie box office success prediction," 2017 3rd International Conference on Electrical Information and Communication Technology (EICT), Khulna, Bangladesh, 2017, pp. 1-6, doi: 10.1109/EICT.2017.8275242.