Mon- 13Mar2023

Report # 2

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 at the following url:

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

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 displays the quantity of 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
Langauge 12	Langauge 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 Extraction:

To utilize the information that is most relevant to the problem at hand, feature preliminary 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:

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.

```
                                     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
```

Preparation Continued:

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  \
0       1            620            723.00    178.00              0.00
1       1            538            302.00    169.00            563.00
2       1           1395            602.00    148.00              0.00
3       1            251            813.00    164.00          22000.00
5       1             62            462.00    132.00            475.00
6       1           1398            392.00    156.00              0.00
```

```
7       1           1125             324.00    100.00             15.00
8       1            839             635.00    141.00              0.00
9       1            364             375.00    153.00            282.00
10      1           1654             673.00    183.00              0.00
11      1            185             434.00    169.00              0.00
12      1            968             403.00    106.00            395.00
13      1            538             313.00    151.00            563.00
14      1            538             450.00    150.00            563.00
15      1           1654             733.00    143.00              0.00
16      1             52             258.00    150.00             80.00
17      1            839             703.00    173.00              0.00
18      1           1321             448.00    136.00            252.00
19      1            110             451.00    106.00            188.00
20      1           1229             422.00    164.00              0.00

    # Actor 1 Likes  Actor 2 Name  # Actor 1 Likes        Gross  Genres
0            855.00          1002          1000.00 760505847.00      91
1           1000.00          1592        e40000.00 309404152.00      85
2            161.00          1795         11000.00 200074175.00     107
3          23000.00           381         27000.00 448130642.00     243
5            530.00          1837           640.00  73058679.00     105
6           4000.00           880         24000.00 336530303.00     101
7            284.00           578           799.00 200807262.00     262
8          19000.00          1758         26000.00 458991599.00     105
9          10000.00           469         25000.00 301956980.00     371
10          2000.00          1222         15000.00 330249062.00     105
11           903.00          1366         18000.00 200069408.00     105
12           393.00          1391           451.00 168368427.00       1
13          1000.00          1592         40000.00 423032628.00      85
14          1000.00          1813         40000.00  89289910.00     109
15           748.00           394         15000.00 291021565.00      91
16           201.00          1671         22000.00 141614023.00      75
17         19000.00          1758         26000.00 623279547.00     105
18          1000.00          1831         40000.00 241063875.00      85
19           718.00          1470         10000.00 179020854.00      32
20           773.00            14          5000.00 255108370.00     376

    Langauge Country  Content Rating       Budget  Title Year  \
0          9      43               7 237000000.00          66
1          9      43               7 300000000.00          64
2          9      42               7 245000000.00          72
3          9      43               7 250000000.00          69
5          9      43               7 263700000.00          69
6          9      43               7 258000000.00          64
7          9      43               6 260000000.00          67
8          9      43               7 250000000.00          72
9          9      42               6 250000000.00          66
10         9      43               7 250000000.00          73
11         9      43               7 209000000.00          63
12         9      42               7 200000000.00          65
13         9      43               7 225000000.00          63
14         9      43               7 215000000.00          70
15         9      43               7 225000000.00          70
16         9      43               6 225000000.00          65
```

```
17          9      43               7 220000000.00          69
18          9      43               7 250000000.00          68
19          9      43               7 225000000.00          69
20          9      30               7 250000000.00          71

    # Actor 2 Likes IMDB Score Aspect Ratio  # Movie Likes      Classes
0             936.00       7.90          1.78          33000   Class Five
1            5000.00       7.10          2.35              0   Class Four
2             393.00       6.80          2.35          85000  Class Three
3           23000.00       8.50          2.35         164000   Class Four
5             632.00       6.60          2.35          24000    Class Two
6           11000.00       6.20          2.35              0   Class Four
7             553.00       7.80          1.85          29000  Class Three
8           21000.00       7.50          2.35         118000   Class Four
9           11000.00       7.50          2.35          10000   Class Four
10           4000.00       6.90          2.35         197000   Class Four
11          10000.00       6.10          2.35              0  Class Three
12            412.00       6.70          2.35              0  Class Three
13           5000.00       7.30          2.35           5000   Class Four
14           2000.00       6.50          2.35          48000    Class Two
15           3000.00       7.20          2.35         118000   Class Four
16            216.00       6.60          2.35              0  Class Three
17          21000.00       8.10          1.85         123000   Class Five
18          11000.00       6.70          2.35          58000  Class Three
19            816.00       6.80          1.85          40000  Class Three
20            972.00       7.50          2.35          65000   Class Four
```

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  \
0   1.00           0.37              0.89      0.48              0.00
1   1.00           0.32              0.37      0.45              0.02
2   1.00           0.84              0.74      0.38              0.00
3   1.00           0.15              1.00      0.43              0.96
4   1.00           0.04              0.57      0.32              0.02

   # Actor 1 Likes  Actor 2 Name  # Actor 1 Likes  Gross  Genres  ...  \
0             0.04          0.46             0.00   1.00    0.12  ...
1             0.04          0.73             0.06   0.41    0.11  ...
2             0.01          0.82             0.02   0.26    0.14  ...
3             1.00          0.17             0.04   0.59    0.33  ...
4             0.02          0.84             0.00   0.10    0.14  ...

   # Users for Reviews  Language  Country  Content Rating  Budget  Title
Year  \
0                 0.60      0.27     0.98            0.64    0.02
0.90
```

```
1                 0.24      0.27     0.98            0.64    0.02
0.88
2                 0.20      0.27     0.95            0.64    0.02
0.99
3                 0.53      0.27     0.98            0.64    0.02
0.95
4                 0.15      0.27     0.98            0.64    0.02
0.95

   # Actor 2 Likes  IMDB Score  Aspect Ratio  # Movie Likes
0             0.01        0.82          0.04           0.09
1             0.04        0.71          0.08           0.00
2             0.00        0.68          0.08           0.24
3             0.17        0.90          0.08           0.47
4             0.00        0.65          0.08           0.07
```

This data preparation report presents an overview of the steps taken to the original IMDB dataset to prepare it for use in our classification- based machine learning project. The dataset was obtained from Kaggle.com. Missing values were removed. Data transformation operations were also carried out, such as encoding categorical variables, and applying numerical scaling.
In the next report “Report #3- Testing and Evaluation”, the data will be split into training, and test sets, and various machine learning algorithms are applied and analyzed.