Methodology

* The 3 classifiers used- **Decision tree, logistic regression**
* Ensemble pipeline
* Other models considered**- only** **Decision tree and logistic regression are considered**
* Hyper-parameter tuning- **We have tuned the parameters using grid search cross validation**.

**criterion="entropy", max\_depth=12**

Before starting with the assignment, we read and understood the business problem behind the assignment and converted that business problem into the analytical problem. We have identified features relevant to solve the objective. Once we have fixed our data variables and the problem-solving approach, we have filtered out irrelevant features and created some new features from the existing variables using feature engineering techniques. We performed in-depth data exploratory data analysis using various visualization methods for example boxplot, heat map etc., refer graph in following pages.

After completing pre-processing, we moved our focus towards predictive analytics. In predictive analysis we used two algorithms namely decision tree and logistic regression and calculated their accuracy. We tried to improve the accuracy of decision tree by using more optimal features for example we changed criteria from default value to entropy and maximum depth to 12 in decision tree.

Overview

* Objective- **Based on users’ demography we want to recommend what genre type movies user may like.**
* Methodology- **CRISP** – **Created business statement and converted it to analytical problem. Collect and understand all the relevant features to solve the problem. Followed by data engineering, EDA, modeling.**

Dataset

* How many features- **33 features**
* Size of the dataset- **100 000**
* Multiple files- **3**
* What kind of data – numerical or character**- numerical, categorical, DateTime**
* Balanced or imbalanced – what is the distribution- **imbalanced**
* Distribution of Training set, validation set, testing set- **60-20-20 percentage**
* Missing data and Preprocessing challenges- **Release date has missing data, we removed null values and converted it to years. Also had release date DateTime format needs preprocessing. Occupation None seems incorrect data. Using exploratory data analysis, we tried to summarize the important characteristics. Age feature has some outliers.**

Results

* Table for the evaluation metric for each ML technique used:   
  **Confusion matrix, given in below pages.**
* Plot of the curves

**All the plots are in page below**

* Conclusion

**For a user with given demographic information we can recommend the user a movie of genre with 37% accuracy.**

Feature Engineering Techniques

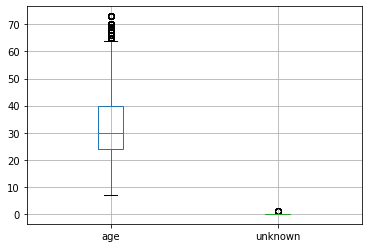
* Features removed**- "movie id","timestamp", "IMDb URL","video release date", "user id", “movie title”**
* Feature creation- **converted release date to years**
* Feature ranking- **computational limitation**
* Class imbalance treatment- **created synthetical data**
* Any other- **Using SMOTE technique to handle the imbalance class would have improved our solution.  
  Used box plot to identify outliers**

**Answering questions given with the assignment.**

Q-1. Write a Data Science Proposal for achieving the objective mentioned.   
**Answer 1:  
 Business problem - Based on users’ demographic information we want to recommend a genre of movies to user. This is a generic problem.  
 Target audience – Recommendation engine will be using our algorithm.  
 Evaluation method- F- score is used to evaluate.  
 Acceptable criteria- Accuracy of the solution should be greater than 50%.**

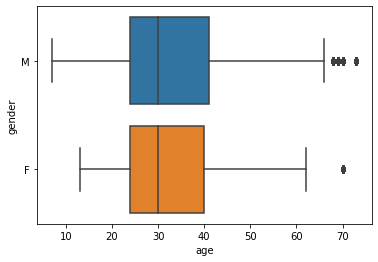
Q-2. Perform exploratory analysis on the data.   
Q-3. Perform data wrangling / pre-processing.   
**Answer 2 and 3:**

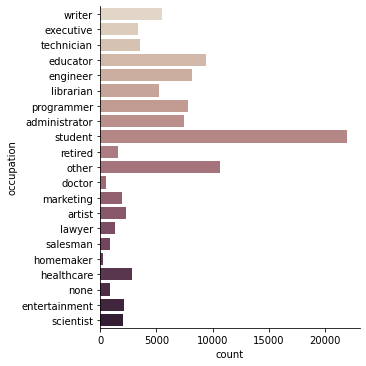
1. **Observation from below graph is that “Unknown” are zero and this column can be dropped.**

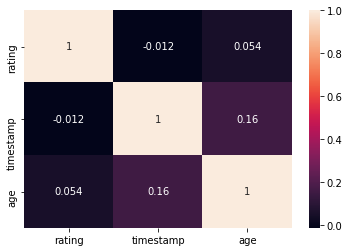


1. **Observation**: **We have found some outlier here. Lower age limit of Male is below Female.**

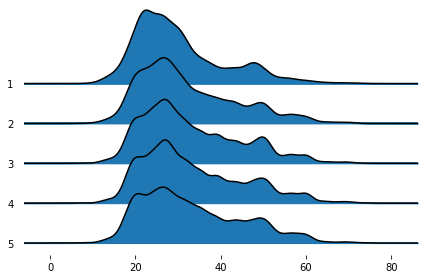
**Median age seems same. Spread in Male age is more than Female.**

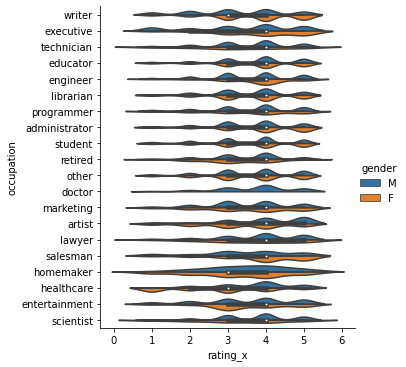


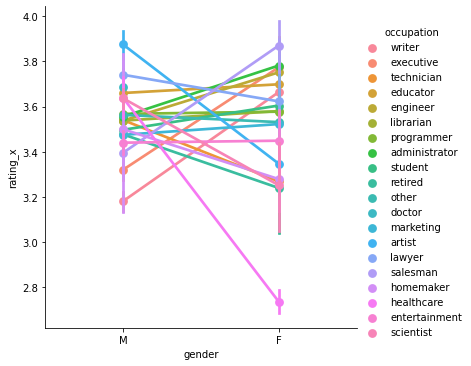
1. **Observation: Student occupation is maximum, and some records are none. ‘None’ can be merged with 'other'.**
2. **Observation: There is weak corelation between age, time-stamp and rating.**



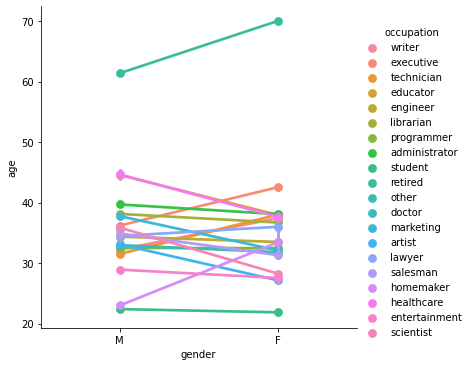
1. **Observation: All ratings seems to have similar age distribution pattern.**

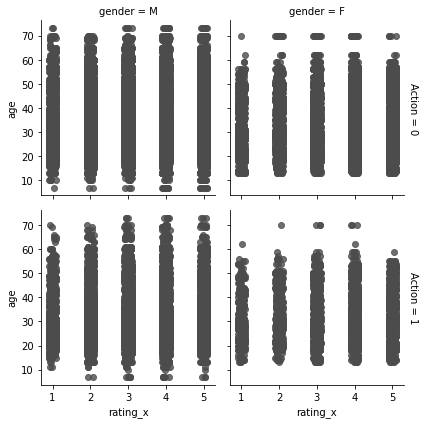


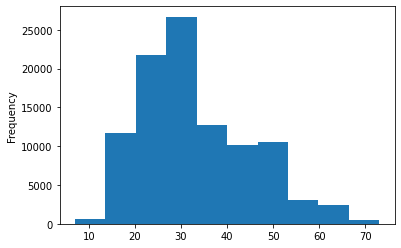
1. **Observation: Occupation doctor has missing gender female, this might me incorrect data.**
2. **Observation: Number of Female healthcare data looks skewed.**

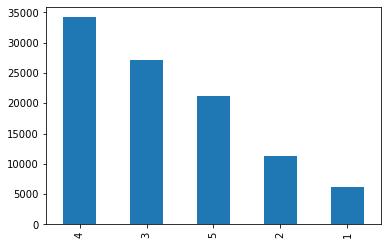
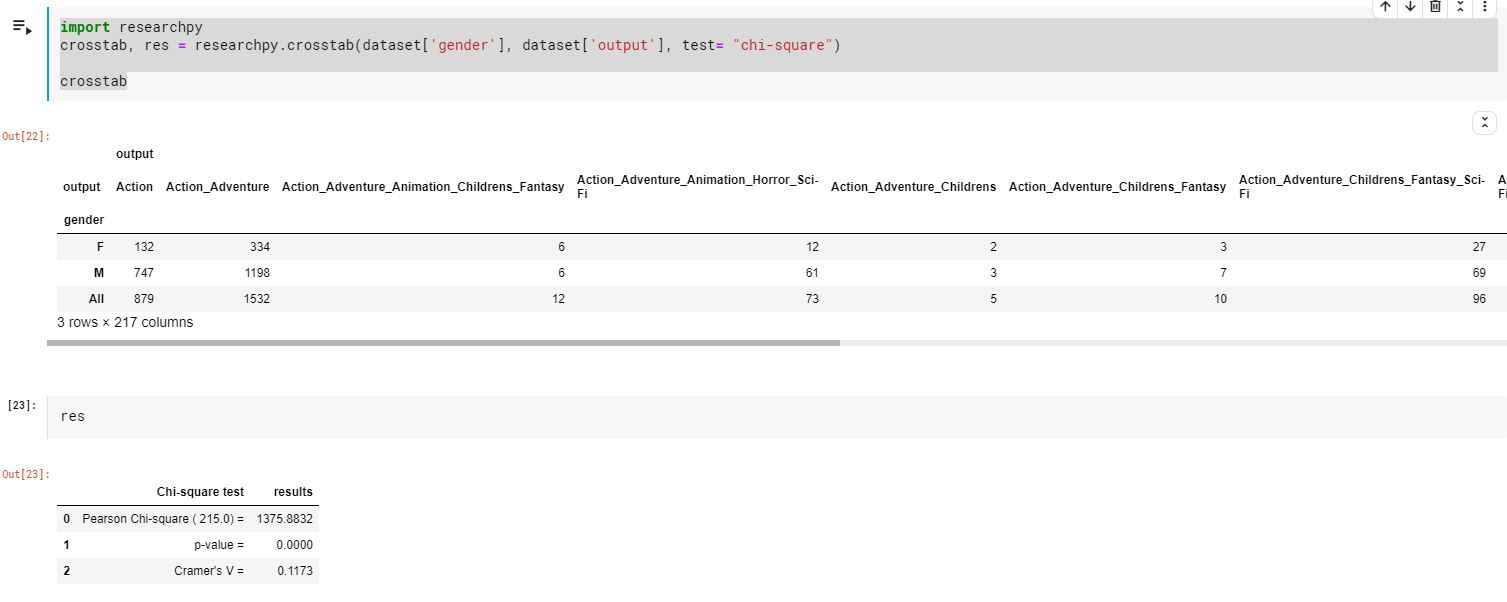


1. **Observation: Age of all retired occupation for male and female is above 60, which implies data is correct. Similarly, students age is near 20 which implies ages of students are correct.**



1. **Observation: Rating, age seems to have no relation for Action movies**  
   
2. **Observation: DataSet has maximum number of users from age group 25-30 and least for 5-10/65-70.**



1. **Observation: Rating of 4 has highest occurrence however rating 1 has least occurrence in data set**
2. **Observation: We wanted to test the hypothesis that is there any relation between gender and movie type.  
   From chi-square test we have found that gender and genre move types are related.**  
   
3. **Confusion Matrix: Average F-score for decision tree is 0.39.**

precision recall f1-score support

Action 0.27 0.23 0.25 166

Action\_Adventure 0.44 0.42 0.43 321

Action\_Adventure\_Animation\_Childrens\_Fantasy 0.00 0.00 0.00 1

Action\_Adventure\_Animation\_Horror\_Sci-Fi 1.00 1.00 1.00 10

Action\_Adventure\_Childrens 0.00 0.00 0.00 0

Action\_Adventure\_Childrens\_Fantasy 0.00 0.00 0.00 2

Action\_Adventure\_Childrens\_Fantasy\_Sci-Fi 0.00 0.00 0.00 15

Action\_Adventure\_Childrens\_Sci-Fi 0.00 0.00 0.00 2

Action\_Adventure\_Comedy 0.00 0.00 0.00 13

Action\_Adventure\_Comedy\_Crime 0.14 0.32 0.20 59

Action\_Adventure\_Comedy\_Horror 0.00 0.00 0.00 13

Action\_Adventure\_Comedy\_Horror\_Sci-Fi 0.00 0.00 0.00 19

Action\_Adventure\_Comedy\_Musical\_Thriller 0.00 0.00 0.00 18

Action\_Adventure\_Comedy\_Romance 0.35 0.43 0.39 90

Action\_Adventure\_Comedy\_Sci-Fi 0.81 0.95 0.88 60

Action\_Adventure\_Comedy\_War 0.00 0.00 0.00 10

Action\_Adventure\_Crime 0.43 0.54 0.48 56

Action\_Adventure\_Crime\_Drama 0.00 0.00 0.00 34

Action\_Adventure\_Crime\_Thriller 0.00 0.00 0.00 3

Action\_Adventure\_Drama 0.31 0.17 0.22 24

Action\_Adventure\_Drama\_Romance 0.00 0.00 0.00 19

Action\_Adventure\_Drama\_Romance\_Sci-Fi\_War 0.39 0.66 0.49 73

Action\_Adventure\_Fantasy 0.52 1.00 0.68 26

Action\_Adventure\_Mystery 0.99 1.00 0.99 66

Action\_Adventure\_Mystery\_Sci-Fi 0.00 0.00 0.00 3

Action\_Adventure\_Romance\_Sci-Fi\_War 0.75 0.84 0.79 204

Action\_Adventure\_Romance\_War 0.26 0.83 0.40 29

Action\_Adventure\_Sci-Fi 0.37 0.30 0.33 349

Action\_Adventure\_Sci-Fi\_Thriller 0.59 0.55 0.57 130

Action\_Adventure\_Sci-Fi\_War 0.00 0.00 0.00 45

Action\_Adventure\_Thriller 0.78 0.67 0.72 275

Action\_Adventure\_Western 0.00 0.00 0.00 9

Action\_Animation\_Childrens\_Sci-Fi\_Thriller\_War 0.00 0.00 0.00 4

Action\_Childrens 0.00 0.00 0.00 4

Action\_Comedy 0.24 0.22 0.23 49

Action\_Comedy\_Crime\_Drama 1.00 1.00 1.00 1

Action\_Comedy\_Crime\_Horror\_Thriller 1.00 1.00 1.00 20

Action\_Comedy\_Drama 0.00 0.00 0.00 39

Action\_Comedy\_Musical 0.26 0.35 0.30 52

Action\_Comedy\_Musical\_Sci-Fi 0.00 0.00 0.00 9

Action\_Comedy\_Sci-Fi\_War 0.44 0.24 0.31 45

Action\_Comedy\_War 0.00 0.00 0.00 10

Action\_Comedy\_Western 0.54 0.33 0.41 88

Action\_Crime 0.00 0.00 0.00 7

Action\_Crime\_Drama 0.82 0.60 0.69 126

Action\_Crime\_Mystery 0.00 0.00 0.00 11

Action\_Crime\_Romance 0.00 0.00 0.00 15

Action\_Crime\_Sci-Fi 0.00 0.00 0.00 18

Action\_Crime\_Thriller 0.00 0.00 0.00 63

Action\_Drama 0.00 0.00 0.00 41

Action\_Drama\_Mystery 0.00 0.00 0.00 8

Action\_Drama\_Mystery\_Romance\_Thriller 0.00 0.00 0.00 14

Action\_Drama\_Romance 0.11 0.08 0.09 91

Action\_Drama\_Romance\_War 0.00 0.00 0.00 22

Action\_Drama\_Thriller 0.00 0.00 0.00 87

Action\_Drama\_Thriller\_War 0.91 0.96 0.93 52

Action\_Drama\_War 0.38 0.44 0.41 177

Action\_Drama\_Western 0.17 0.07 0.10 29

Action\_Horror 0.46 0.73 0.56 67

Action\_Horror\_Sci-Fi 0.00 0.00 0.00 28

Action\_Horror\_Sci-Fi\_Thriller 0.24 0.29 0.27 78

Action\_Mystery\_Romance\_Thriller 0.84 0.98 0.90 52

Action\_Mystery\_Sci-Fi\_Thriller 0.00 0.00 0.00 32

Action\_Mystery\_Thriller 0.27 0.50 0.35 8

Action\_Romance 0.13 0.16 0.14 51

Action\_Romance\_Thriller 0.38 0.20 0.26 178

Action\_Romance\_War 0.00 0.00 0.00 32

Action\_Sci-Fi 0.81 0.55 0.65 62

Action\_Sci-Fi\_Thriller 0.43 0.22 0.29 192

Action\_Sci-Fi\_Thriller\_War 0.23 0.46 0.31 59

Action\_Sci-Fi\_War 0.86 0.99 0.92 85

Action\_Thriller 0.40 0.41 0.40 660

Action\_Thriller\_War 0.00 0.00 0.00 20

Action\_Western 0.93 0.96 0.95 27

Adventure 0.67 0.28 0.40 57

Adventure\_Animation\_Childrens\_Comedy\_Fantasy 0.00 0.00 0.00 20

Adventure\_Animation\_Childrens\_Comedy\_Musical 0.00 0.00 0.00 11

Adventure\_Animation\_Childrens\_Musical 0.18 0.29 0.22 7

Adventure\_Animation\_Sci-Fi\_Thriller 0.00 0.00 0.00 6

Adventure\_Childrens 0.66 0.30 0.42 102

Adventure\_Childrens\_Comedy 0.49 0.79 0.60 82

Adventure\_Childrens\_Comedy\_Fantasy\_Romance\_Sci-Fi 0.00 0.00 0.00 3

Adventure\_Childrens\_Drama 0.00 0.00 0.00 10

Adventure\_Childrens\_Drama\_Musical 0.43 0.92 0.58 49

Adventure\_Childrens\_Fantasy 0.00 0.00 0.00 12

Adventure\_Childrens\_Fantasy\_Sci-Fi 0.33 0.14 0.20 14

Adventure\_Childrens\_Musical 0.00 0.00 0.00 11

Adventure\_Childrens\_Romance 0.00 0.00 0.00 19

Adventure\_Comedy 0.20 0.10 0.13 10

Adventure\_Comedy\_Drama 0.08 0.04 0.05 51

Adventure\_Drama 0.00 0.00 0.00 30

Adventure\_Drama\_Western 0.11 0.09 0.09 47

Adventure\_Romance 0.00 0.00 0.00 24

Adventure\_Sci-Fi 0.00 0.00 0.00 17

Adventure\_Sci-Fi\_Thriller 0.38 0.71 0.50 7

Adventure\_Thriller 1.00 1.00 1.00 22

Adventure\_War 0.22 0.14 0.17 58

Animation 0.64 0.44 0.52 16

Animation\_Childrens 0.24 0.07 0.11 86

Animation\_Childrens\_Comedy 0.00 0.00 0.00 92

Animation\_Childrens\_Comedy\_Musical 0.15 0.21 0.18 42

Animation\_Childrens\_Comedy\_Romance 0.00 0.00 0.00 8

Animation\_Childrens\_Musical 0.37 0.32 0.34 285

Animation\_Childrens\_Musical\_Romance 0.00 0.00 0.00 12

Animation\_Comedy 0.25 0.07 0.11 74

Animation\_Comedy\_Thriller 1.00 1.00 1.00 30

Animation\_Sci-Fi 0.50 0.12 0.20 8

Childrens 0.00 0.00 0.00 2

Childrens\_Comedy 0.29 0.21 0.24 234

Childrens\_Comedy\_Drama 0.15 0.04 0.06 51

Childrens\_Comedy\_Fantasy 1.00 0.20 0.33 15

Childrens\_Comedy\_Musical 0.48 0.42 0.45 52

Childrens\_Comedy\_Mystery 0.33 0.22 0.27 9

Childrens\_Comedy\_Western 0.00 0.00 0.00 3

Childrens\_Drama 0.35 0.15 0.21 54

Childrens\_Drama\_Fantasy 0.00 0.00 0.00 10

Childrens\_Drama\_Fantasy\_Sci-Fi 0.23 0.13 0.17 75

Childrens\_Fantasy 0.00 0.00 0.00 4

Comedy 0.34 0.46 0.39 1990

Comedy\_Crime 0.75 0.63 0.69 122

Comedy\_Crime\_Drama 0.00 0.00 0.00 1

Comedy\_Crime\_Drama\_Mystery 0.00 0.00 0.00 18

Comedy\_Crime\_Fantasy 0.00 0.00 0.00 30

Comedy\_Crime\_Horror 0.00 0.00 0.00 8

Comedy\_Crime\_Mystery\_Thriller 0.28 1.00 0.44 9

Comedy\_Drama 0.42 0.22 0.29 475

Comedy\_Drama\_Musical 0.13 0.09 0.11 33

Comedy\_Drama\_Romance 0.34 0.10 0.15 147

Comedy\_Drama\_Thriller 0.00 0.00 0.00 4

Comedy\_Drama\_War 0.00 0.00 0.00 18

Comedy\_Fantasy 0.00 0.00 0.00 5

Comedy\_Fantasy\_Romance\_Sci-Fi 0.68 1.00 0.81 32

Comedy\_Horror 0.00 0.00 0.00 100

Comedy\_Musical 0.27 0.12 0.17 50

Comedy\_Musical\_Romance 0.60 0.52 0.56 91

Comedy\_Mystery 0.00 0.00 0.00 9

Comedy\_Mystery\_Romance 0.00 0.00 0.00 2

Comedy\_Mystery\_Romance\_Thriller 0.00 0.00 0.00 4

Comedy\_Mystery\_Thriller 0.51 1.00 0.68 22

Comedy\_Romance 0.28 0.26 0.27 973

Comedy\_Romance\_Thriller 0.07 0.02 0.03 61

Comedy\_Romance\_War 0.00 0.00 0.00 66

Comedy\_Sci-Fi 0.42 0.43 0.43 136

Comedy\_Thriller 0.45 0.79 0.57 38

Comedy\_War 0.51 0.51 0.51 88

Comedy\_Western 0.00 0.00 0.00 9

Crime 0.39 0.42 0.41 59

Crime\_Drama 0.46 0.28 0.35 276

Crime\_Drama\_Film-Noir 0.00 0.00 0.00 40

Crime\_Drama\_Mystery 0.00 0.00 0.00 32

Crime\_Drama\_Mystery\_Thriller 0.57 1.00 0.73 4

Crime\_Drama\_Romance 0.36 0.56 0.43 18

Crime\_Drama\_Romance\_Thriller 0.29 0.15 0.20 47

Crime\_Drama\_Sci-Fi 0.00 0.00 0.00 41

Crime\_Drama\_Thriller 0.67 0.54 0.60 181

Crime\_Film-Noir 0.00 0.00 0.00 2

Crime\_Film-Noir\_Mystery 0.00 0.00 0.00 8

Crime\_Film-Noir\_Mystery\_Thriller 0.00 0.00 0.00 69

Crime\_Film-Noir\_Thriller 0.36 0.32 0.34 28

Crime\_Horror\_Mystery\_Thriller 0.00 0.00 0.00 44

Crime\_Thriller 0.46 0.36 0.40 172

Documentary 0.71 0.13 0.22 116

Documentary\_Drama 0.00 0.00 0.00 11

Documentary\_War 0.00 0.00 0.00 11

Drama 0.29 0.59 0.39 2668

Drama\_Fantasy\_Thriller 0.00 0.00 0.00 20

Drama\_Horror 0.00 0.00 0.00 69

Drama\_Musical 0.45 0.53 0.49 79

Drama\_Musical\_War 0.13 0.22 0.16 23

Drama\_Mystery 0.35 0.33 0.34 123

Drama\_Mystery\_Romance 0.00 0.00 0.00 9

Drama\_Mystery\_Sci-Fi\_Thriller 0.60 0.91 0.72 46

Drama\_Mystery\_Thriller 0.00 0.00 0.00 17

Drama\_Romance 0.39 0.34 0.36 937

Drama\_Romance\_Thriller 0.00 0.00 0.00 20

Drama\_Romance\_War 0.74 0.74 0.74 200

Drama\_Romance\_War\_Western 0.00 0.00 0.00 17

Drama\_Sci-Fi 0.86 0.50 0.63 206

Drama\_Sci-Fi\_Thriller 0.00 0.00 0.00 22

Drama\_Thriller 0.50 0.55 0.53 536

Drama\_Thriller\_War 0.00 0.00 0.00 27

Drama\_War 0.39 0.18 0.25 405

Drama\_Western 0.00 0.00 0.00 4

Fantasy 0.00 0.00 0.00 1

Film-Noir 0.33 0.25 0.29 12

Film-Noir\_Mystery 0.25 0.06 0.10 34

Film-Noir\_Mystery\_Thriller 0.25 0.19 0.21 27

Film-Noir\_Romance\_Thriller 0.00 0.00 0.00 8

Film-Noir\_Sci-Fi 0.26 0.41 0.32 64

Film-Noir\_Sci-Fi\_Thriller 0.00 0.00 0.00 1

Film-Noir\_Thriller 0.40 0.35 0.37 46

Horror 0.40 0.28 0.33 303

Horror\_Mystery\_Thriller 1.00 1.00 1.00 13

Horror\_Romance 0.10 0.15 0.12 20

Horror\_Romance\_Thriller 0.62 0.86 0.72 44

Horror\_Sci-Fi 0.50 0.15 0.23 20

Horror\_Sci-Fi\_Thriller 0.00 0.00 0.00 13

Horror\_Thriller 0.57 0.65 0.61 139

Musical 0.67 0.69 0.68 62

Musical\_Romance 0.42 0.32 0.36 90

Mystery 0.49 0.58 0.53 43

Mystery\_Romance\_Thriller 0.60 0.75 0.67 8

Mystery\_Sci-Fi 0.00 0.00 0.00 1

Mystery\_Thriller 0.60 0.50 0.55 285

Romance 0.62 0.28 0.38 58

Romance\_Thriller 0.00 0.00 0.00 23

Romance\_War 0.50 0.29 0.36 7

Sci-Fi 0.37 0.23 0.28 122

Sci-Fi\_Thriller 0.16 0.12 0.14 48

Sci-Fi\_War 0.48 0.40 0.44 40

Thriller 0.41 0.07 0.12 208

War 0.00 0.00 0.00 4

Western 0.09 0.10 0.10 116

accuracy 0.39 19676

macro avg 0.26 0.26 0.24 19676

weighted avg 0.37 0.39 0.36 19676

Q-4. Apply any 2 features engineering technique.

**Answer 4:**

1. **We have applied Release date from DateTime format to years.**
2. **Converted all the categorical variables into vectors using One-hot-encoding.**

Q-5. Plot top 10 features.  
 Answer 5: Refer Answer 2 and 3.

Q-6. Identification of the performance parameters to be improved, for the given problem statement.

**Answer: F-score to be improved.**

Q-7. Design Machine Learning models – Logistic regression and Decision tree to predict.   
 **Answer: It is done in code, please refer .ipynb file attached.**

Q-8. Compare the performance of selected feature engineering techniques.

**Answer: Accuracy without performing feature engineering is 0.14 and after feature engineering is 0.38.**

Q-9. Compare the performance of the 2 classifiers – Logistic regression and Decision tree to predict.

**Answer: F-score of logistic regression is 0.14 and F-score for decision tree is 0.39.**

Q-10. Present the conclusions/results in the format shared.

**Answer: Refer page-1 of this document.**