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**

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**
* Plot of the curves
* Conclusion

**For a user with given demographic information we can recommend him a genre type movie with 37%.**

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- **We have used SMOTE technique to handle the imbalance class.  
  Used box plot to identify outliers**

1. Write a Data Science Proposal for achieving the objective mentioned.

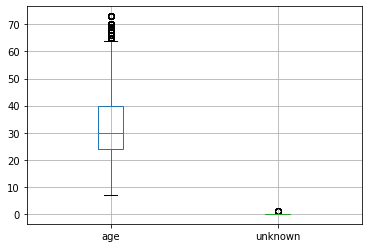
**Answer-1 Business problem - Based on users’ demography we want to recommend what genre type movies user may like. This is a generic problem.  
 Target audience – Recommendation engine.  
 Evaluation method- F- score is used to evaluate  
 Acceptable criteria- Accuracy of the solution should be greater than 50%.**

2. Perform exploratory analysis on the data.

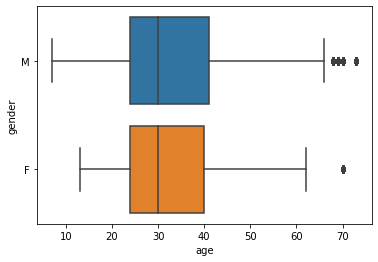
3. Perform data wrangling / pre-processing.

**Answer 2 and 3:**

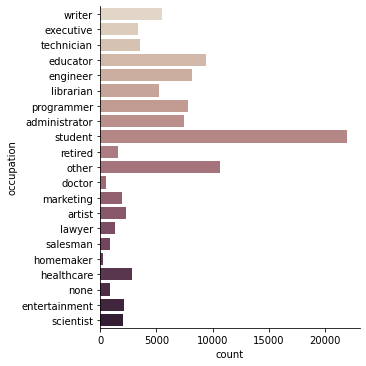
1. **Observation is that Unknowns are zero and this column can be dropped.**

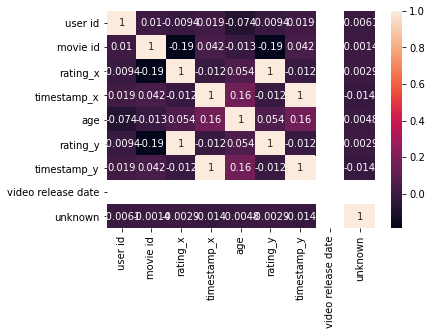


1. Observation: We have found some outlier here. lower age limit of Male is below Female. Median age seems same. Spread in Male age is more than Female.

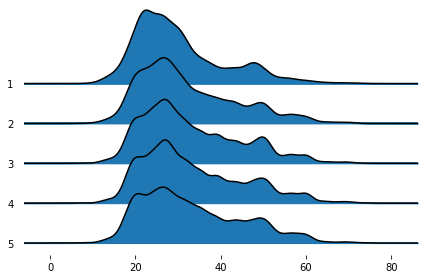


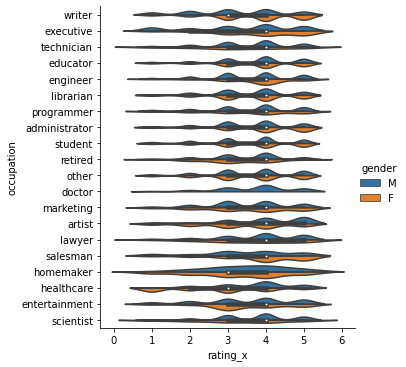
1. Observation : Student occupation is maximum, some records are none, ‘None’ can be merged with 'other'.

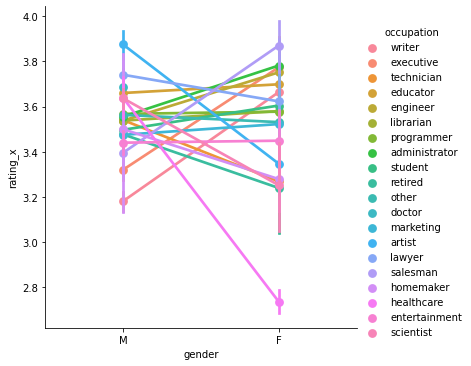


1. Observation : drop high correlated columns timestamp\_y, rating\_x,   
   video release date has no data , so this column can be dropped
2. 

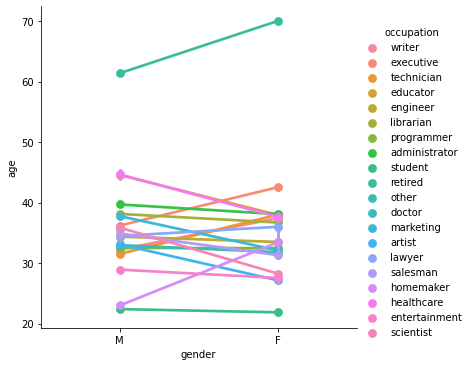
1. Observation : All ratings seems to have similar age distribution pattern

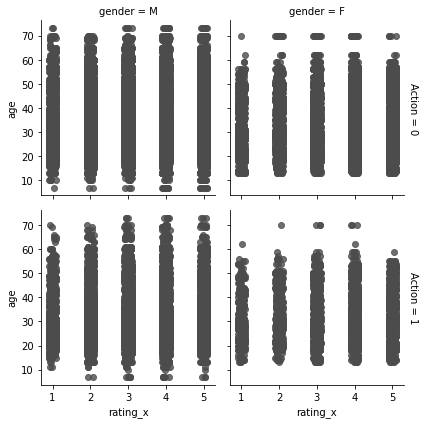
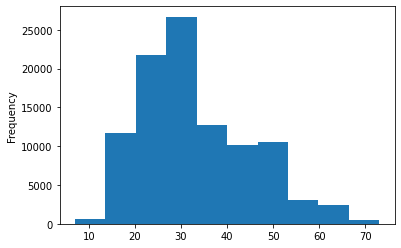
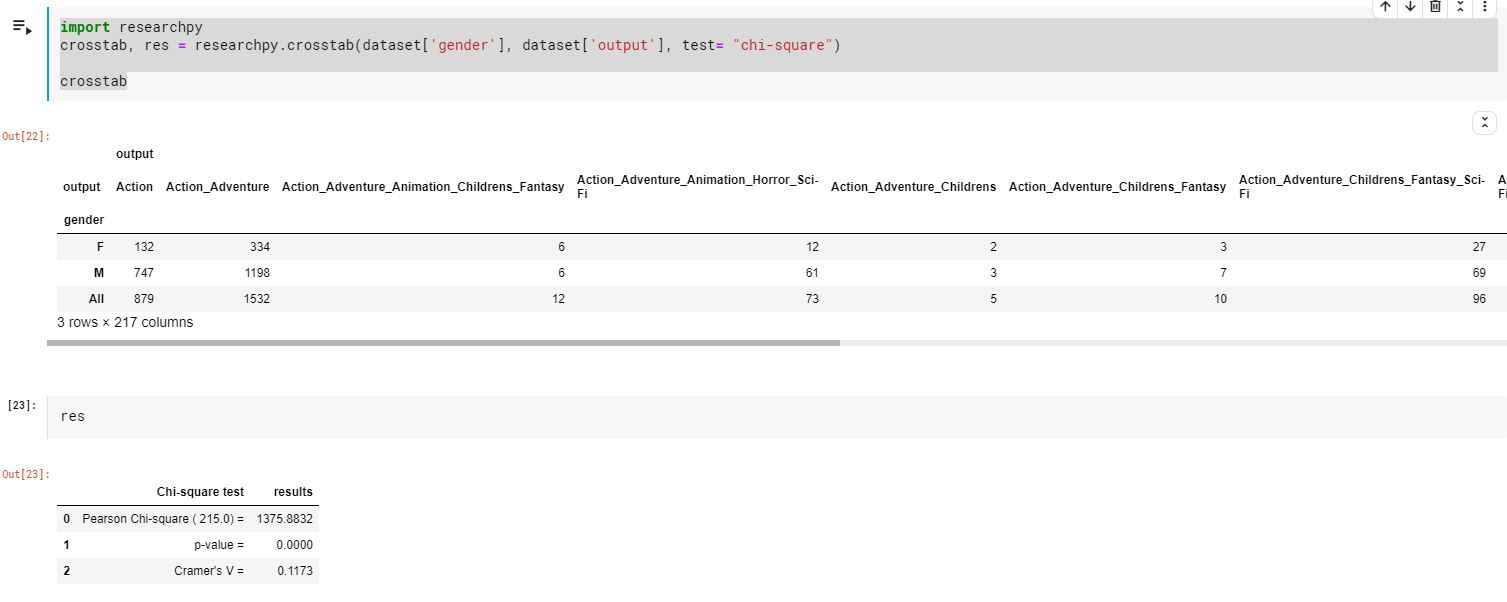


1. Observation: Occupation doctor has missing gender female, this might me incorrect data
2. 
3. Number of Female healthcare data looks skewed



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 also 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.  
   
3. 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.  
   

4. Apply any 2 features engineering technique.

Answer-4) a) We have applied Release date from DateTime format to years.  
 b) Converted all the categorical variables into vectors using One-hot-encoding.

5. Plot top 10 features.  
 Refer Answer 2 and 3.

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

Answer: F-score to be improved.

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

8. Compare the performance of selected feature engineering techniques.

Answer: We can do this by

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

Answer: We can do this by

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

Refer page-1 of this document.