MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT AND ENGINEERING

Project Documentation
For

MOVIE RATING PREDICTION AND RECOMMENDER SYSTEM

April 2020

By-
B003- Shilpika Agarwal
B004- Harsh Agarwal
B011- Raj Bhagwani
B014- Amit Birajdar
B015- Manan Bolia
B019- Vedang Gupte

GOAL/PROBLEM STATEMENT

The goal of this project is to predict the rating that a user will give to a movie by analyzing various factors that affect the rating given. User demographics like gender, age, and occupation are studied to understand the likelihood of a particular rating being given. The genre of the movie is also taken into account to analyze the correlation of these factors with the final ratings. This analysis has two applications-

- The user can be recommended to watch certain movies that are more likely to match his/her interests by analyzing the ratings given by other users and the user's ratings for other movies as well.
- Production houses and marketers can get an insight into the target demographic which is likely to react favorably to their movie and towards which their movies can be marketed.

UNDERSTANDING THE DATASET

Dataset used: MovieLens
Contents: 100,000 ratings from 943 users on 1683 movies.

MovieLens datasets were collected by the GroupLens Research Project at the University of Minnesota.

This data set consists of:

- 100,000 ratings (1-5) from 943 users on 1682 movies.
- Each user has rated at least 20 movies.
- Simple demographic info for the users (age, gender, occupation, zip)

DETAILED DESCRIPTIONS OF DATA FILES

Here is a brief description of the data-

u.data - The full user data set, 100000 ratings by 943 users on 1682 items.
Each user has rated at least 20 movies.
Users and items are numbered consecutively from 1.
The data is randomly ordered.
This is a tab-separated list of user id | item id | rating | timestamp.
The timestamps are Unix seconds since 1/1/1970 UTC

u.info - The number of users, items, and ratings in the user data set.

u.item - Information about the items (movies);
This is a tab-separated list of

movie id	Thriller
movie title	Horror
release date	Mystery
IMDb URL	Musical
unknown	Film-Noir
Action	Fantasy
Adventure	Drama
Animation	Documentary
Children's	Romance
Comedy	Sci-Fi
Crime	Western
War	

- The last 19 fields are the genres where ‘yes’ indicates the movie is of that genre and no’ indicates it is not.
- Movies can be in several genres at once.
- The movie ids are the ones used in the u.data data set.

u.genre - A list of the genres.

u.user - Demographic information about the users; this is a tab-separated list of
user id | age | gender | occupation | zip code
The user ids are the ones used in the u.data data set.

u.occupation - A list of the occupations.

EXPLORATORY DATA ANALYSIS

1. FREQUENCY OF RATING BY GENDER

We initially decided to analyze whether the gender of the user played any role in the rating that was given for the movie. It was necessary to understand if the likelihood of a movie getting a higher or lower rating depended on the gender of the user. If there is indeed any correlation between the gender and the rating, then it would be easier to predict the rating based on this factor.

Thus we split our data into two based on the gender of the user. The frequency of each rating was calculated for both the genders. A pie chart was then derived from this data which is shown below.

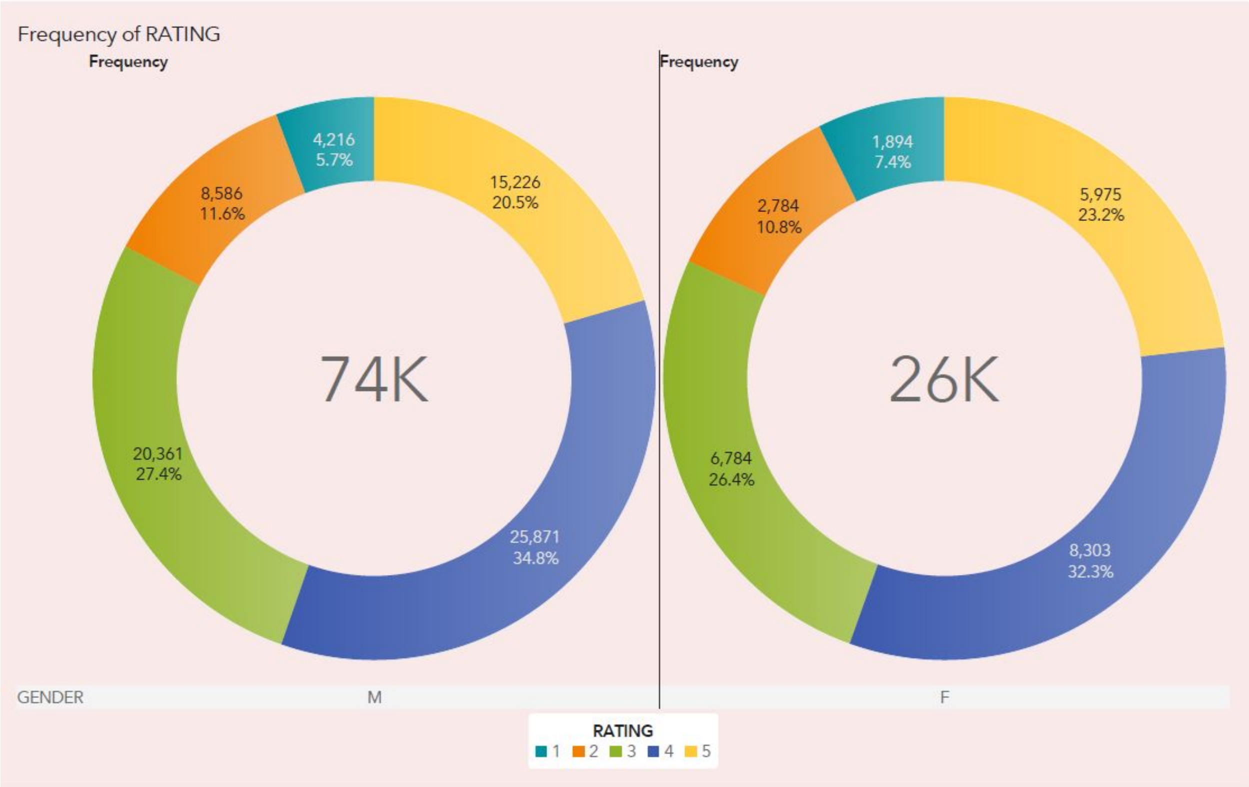


Fig 1: FREQUENCY OF RATING BY GENDER

The results of the analysis are-

- The relative order of the frequency of rating is identical for both genders. 4 is the most frequent rating followed by 3, 5, 2 and 1 being the least selected rating.

- However, there is a minute difference between the frequency percentage for the same rating. Males are more likely to give a rating of 4 or a 5 compared to females. The percentage of females rating a 1, 2, or 3 is higher than that given by men.

Thus males seem to give a higher rating more frequently than females. However, this difference in the distribution is less than 2% and thus not significant in most cases. The only case where the difference in rating is 2.7% is that for a rating of 5, implying men give significantly more 5-star ratings.

However, it is important to note that the total number of males is 670 while females were 273. This could imply that the survey was biased, otherwise the number of females should have been close to the number of males. If the sample was random then either-

- The random sample should be taken again for a better split.
- Females do not watch as many movies as males.
- Females were less inclined to participate in the study due to which the conclusions might be inaccurate.

2. AGE DISTRIBUTION AND OCCUPATION OF USERS

To understand the user demographic, we plot a histogram showing the age of users. We then analyzed the role that the occupation of a user plays in our study. These two factors could shed some light on the demographics that need to be targeted while making a movie recommendation system or

First, we look at the age of users, for which we plot a histogram. Based on the histogram we can draw some conclusions-

- The histogram is left-skewed, thus most of the population surveyed is young.
- The bulk or more than half of the sample age lies between the age of 15 and 35. Thus this demographic should be targeted for marketing since it is likely that they watch more movies.
- The frequency of people below 15 and those above 65 is marginal (less than 5 % combined), thus they do not play an important role in the calculated average rating of a movie.

.

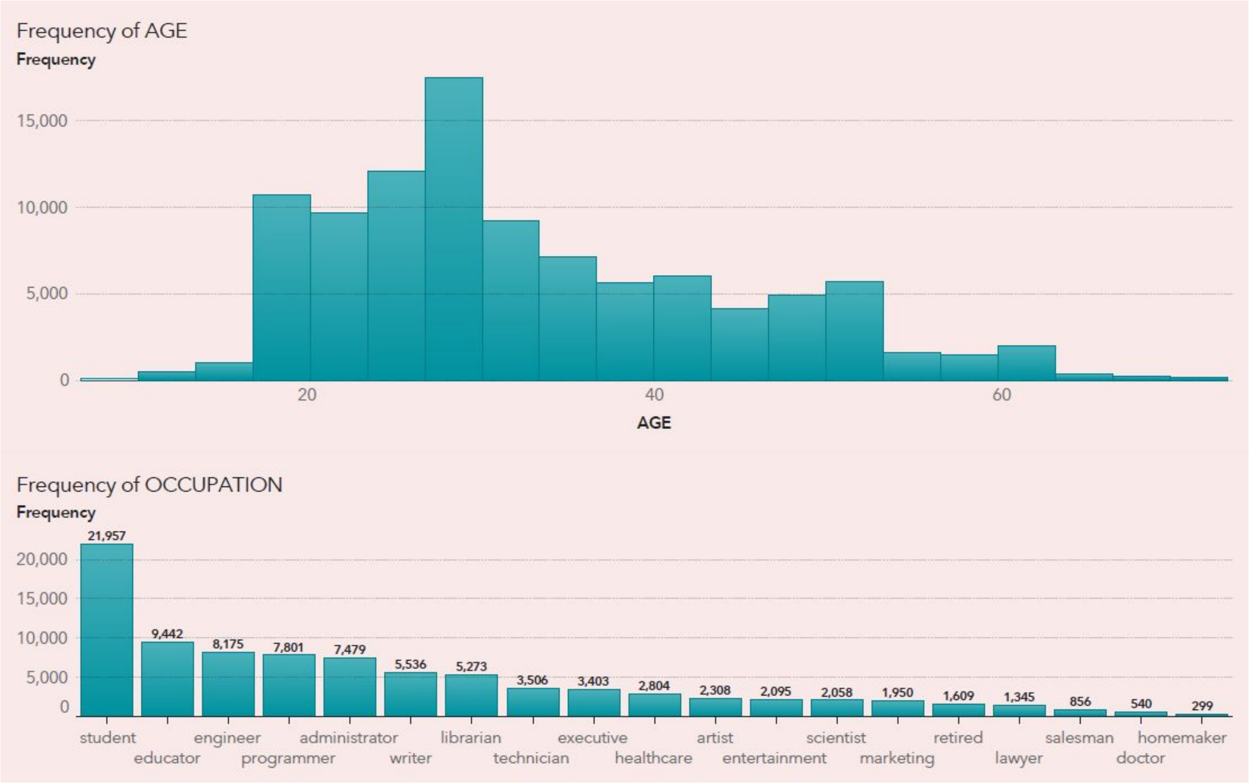


Fig 2: Frequency of age and occupation

Next, we plot a bar graph that divides the users based on their occupation. The observations are as follows-

- The most prominent group is that of students. Almost 22 % of the ratings were given by students.
- Educators, engineers, programmers, and administrators, when combined with students form over half of the tuples. Recommendations should be targeted towards these professionals.
- Marketers, retired, lawyers, salesmen, doctors, and homemakers are all less than 2% of the ratings. Some of these professions can be ignored for targeted marketing but not all since together they form a significant percentage of the sample.

Thus, overall the largest demographic seems to be young people and students, which often overlap.

3. FREQUENCY OF MOVIES BY YEAR

It is important to understand the spread of the movies included in the survey vis-a-vis the year in which it was launched. This gives a better idea about the important time period on which the data is focused.

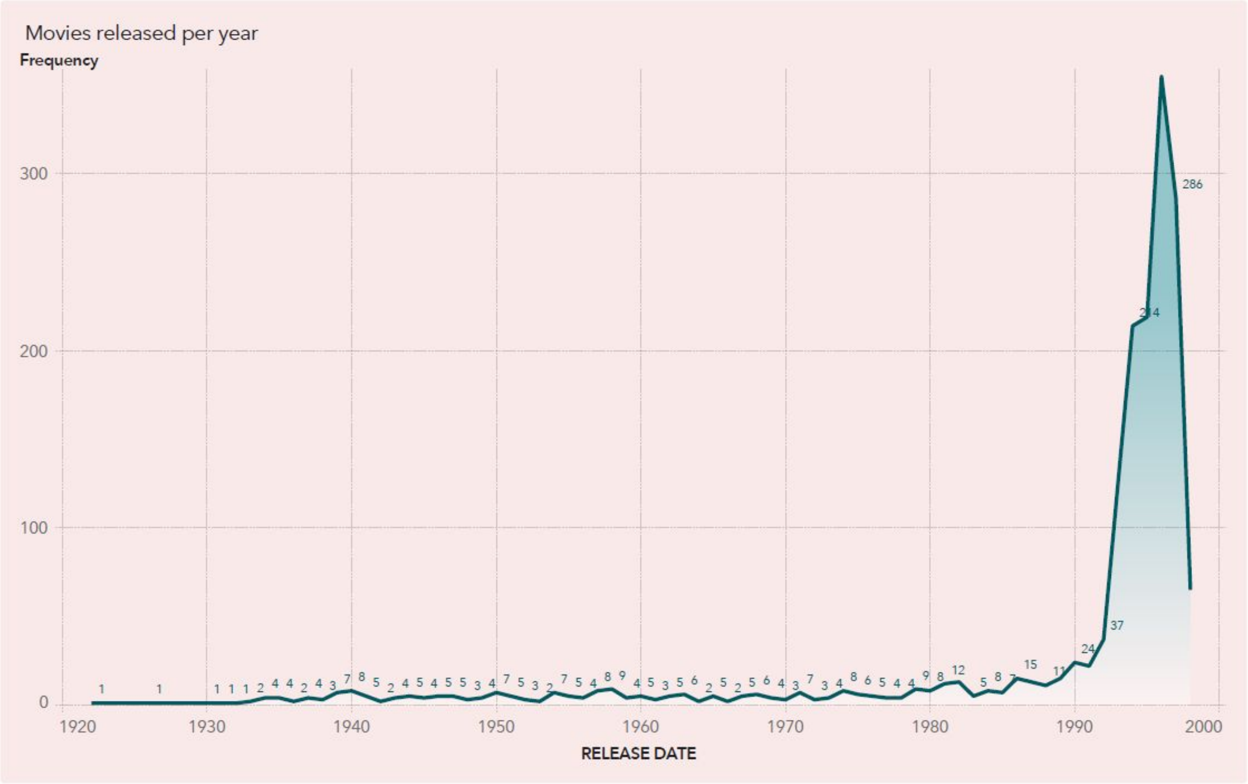


Fig 3: Movies released per year

From the graph shown above, some observations can be made-

- The frequency of movies released before 1930 is marginal and thus these can be ignored
- The percentage of movies released for each year before 1980 is around 0.5% or lower, thus no particular year has great significance.
- The cumulative percentage of ratings for movies between 1920-1980 is 13%. Therefore even though the frequency for each movie is lowest in this region, the cumulative effect cannot be ignored.
- After 1990, there is a major spike in the frequency of movies. More than half of the movies in the dataset were released. This is clearly the most important time period.
- The highest spike comes in the mid-1990s.

There is no data to suggest that the number of movies released has such a vast difference over the years. But there is a gradual increase in frequency observed after 1980 and two spikes after the 1990s. Post the mid 1990s the frequency declines. There is no data about the late 1990s and later. This suggests the data is relatively old and movies post that period are not included in the survey.

4. FREQUENCY OF OCCUPATION GROUPED BY RATING

The bar graph displayed below is bifurcated (by gender) between men and women and then is divided on the basis of the occupation of a particular user. The graph displays the frequency of a rating on its x-axis. An individual bar that is allocated to occupation then displays five different colors for both the genders. These colors depict different ratings given by different individuals of different occupations to a set of movies. Sea green depicts the rating 1 on 5. Orange depicts the rating 2 on 5. Green depicts the rating 3 on 5. Blue depicts the rating 4 on 5 and Yellow depicts the rating 5 on 5.

This graph helps us understand how people working in different sectors of an economy perceive and rate different movies. From this graph, we can analyze how different occupations end up liking a type of movie, then understand how many people are working in that particular field to target an audience and see which type of movie appeals to different sections of the society.

To read the graph is very easy, which is one of the reasons to choose this graph for analysis. You first have to pick which occupation you want to analyze. Then choose if you want to analyze the male section of that particular occupation or female section of it. Once you are done picking the section, you have to read the key which specifies which color indicates what rating. Now, check the x-axis of the graph which shows the frequency of a particular rating. For example, you want to check out of all the administrators that have rated the movies, how many male administrators have given a 5 on 5 ratings? So, you first go and fetch the occupation, that is, administrator then find the male section's bar. In that bar see how much of the part is covered with the color yellow, as the color yellow depicts the rating 5. In the graph given below, you can see that almost the last 20% of that bar is colored yellow, this tells us that out of all the male administrators that rated the movie, 20% of them rated it as a 5 on 5. So if there were 300 male administrators, 60 of them gave the rating 5 out of 5.

After looking at the graph you can draw several conclusions,

- Majority of the graph, both in the male and the female section is largely covered by the color green and blue, this shows that be it any occupation, most of the people tend to rate the movie either 3 or 4 on 5.

- The graph shows no reading for the female category, doctor section because of lack of data and people in that section refusing to give ratings.

- The graph for retired people in both male and female categories shows that retired people tend to easily get amused by any movie as the ratings given by this section tend to be on the higher end of the spectrum. Only about 3% retired males and about 1% of retired women have given the rating as 1 on 5. However, they are almost never really completely satisfied by a movie as they rarely tend to give a 5 on 5 ratings, in both male and female sections the yellow part which depicts 5 on 5 ratings is the least for retired people.

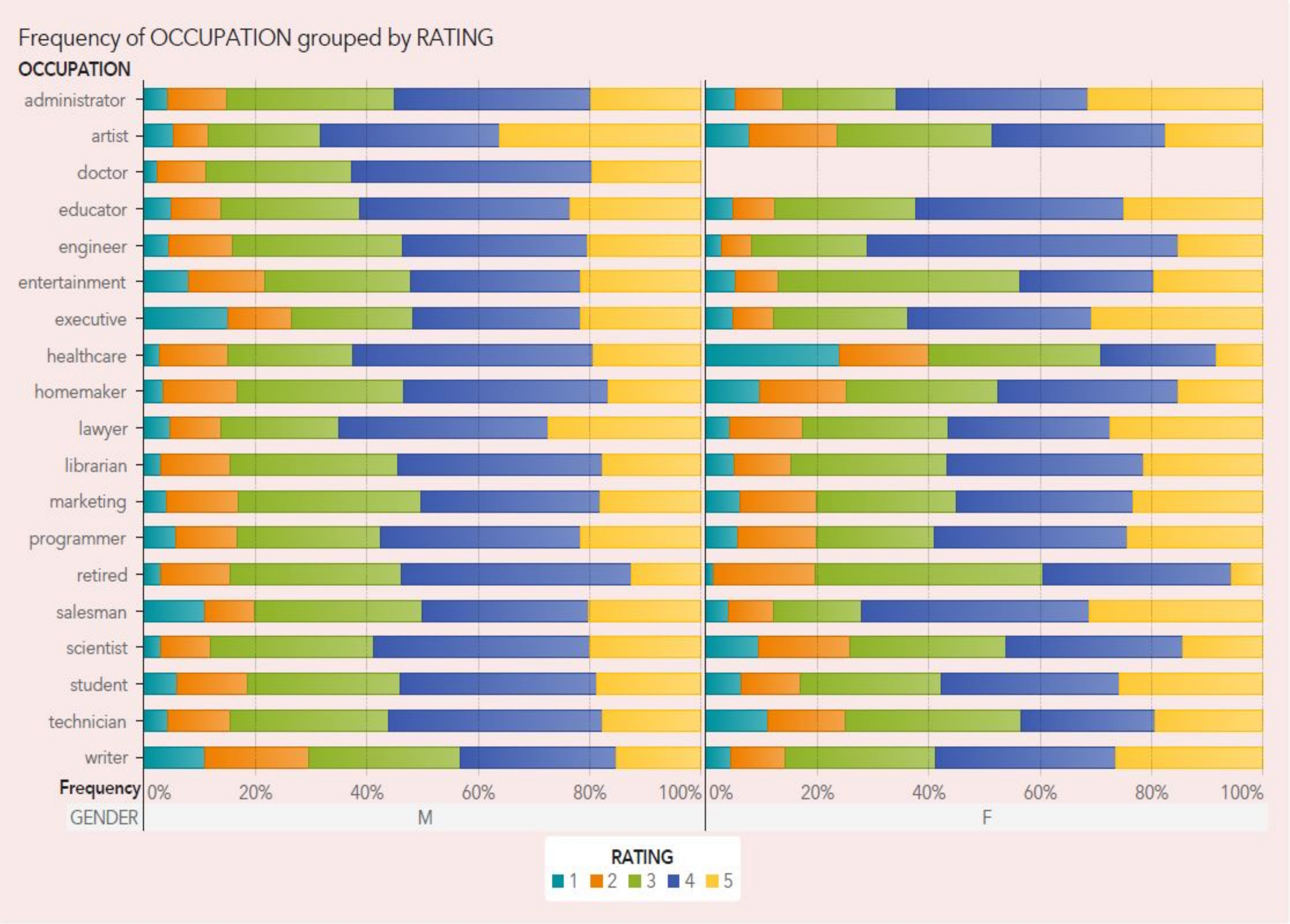
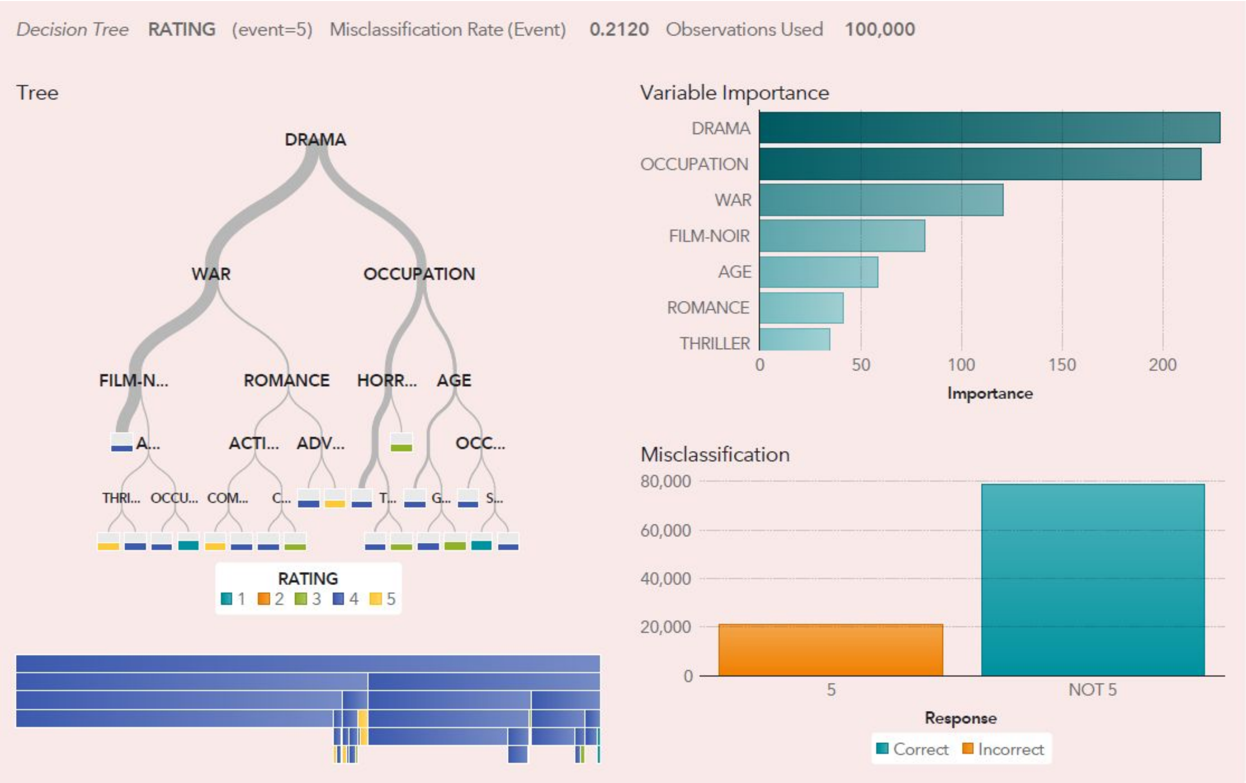


Fig 4: Frequency of occupation by rating and gender

MODEL 1: DECISION TREE

A decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter/differentiator in input variables. A decision is an effective model within which you can lay-out options and investigate the possibility of selecting those options.

They provide a highly effective structure within which you can lay-out options and investigate the possible outcomes of choosing those options. They also help to form a balanced picture of the risks and rewards associated with each possible course of action.



Decision Tree RATING (event=5) Misclassification Rate (Event) 0.2120 Observations Used 100,000

We use the decision tree model to predict the rating for a movie based on 19 genres, age, and occupation of the user.

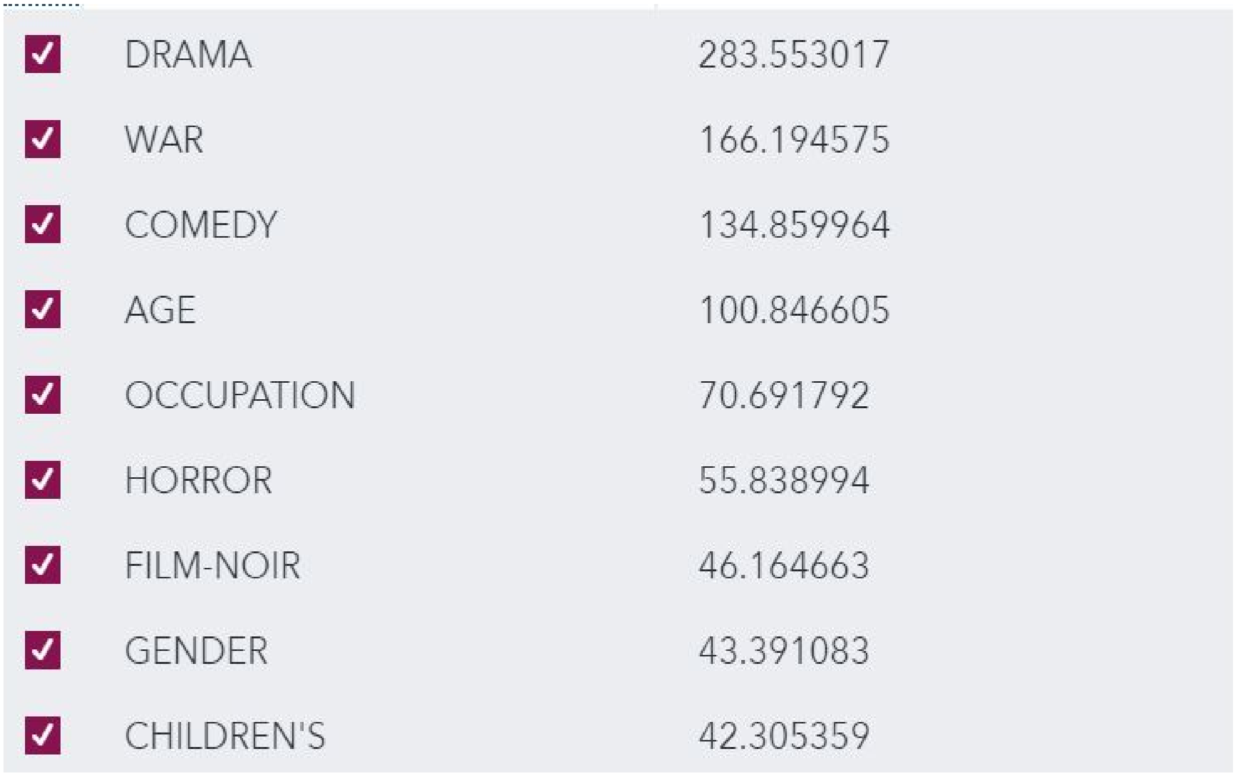
Input to the model:

- Genres
- Age of the user
- Occupation

Output:

- Prediction of rating for a movie

The model begins with the ‘DRAMA’ genre as the root node as it has the highest information gain value.



☑	DRAMA	283.553017
☑	WAR	166.194575
☑	COMEDY	134.859964
☑	AGE	100.846605
☑	OCCUPATION	70.691792
☑	HORROR	55.838994
☑	FILM-NOIR	46.164663
☑	GENDER	43.391083
☑	CHILDREN'S	42.305359

The model calculates information gain value after every split and selects the node with the greatest value. The tree generated is the simplest form with the best validation.
The misclassification chart shows the total number of true negatives and false negatives predicted by the decision tree model.

The misclassification statistics for each level are:

Rating	True Positive	False Negative	True Negative	False Positive	Misclassification rate
1	344	5766	93571	319	0.0609
2	0	11370	88630	0	0.1137
3	0	27145	72855	0	0.2715
4	0	34714	65826	0	0.3417
5	0	21201	78799	0	0.2120

This table summarizes the accuracy of the decision tree model.

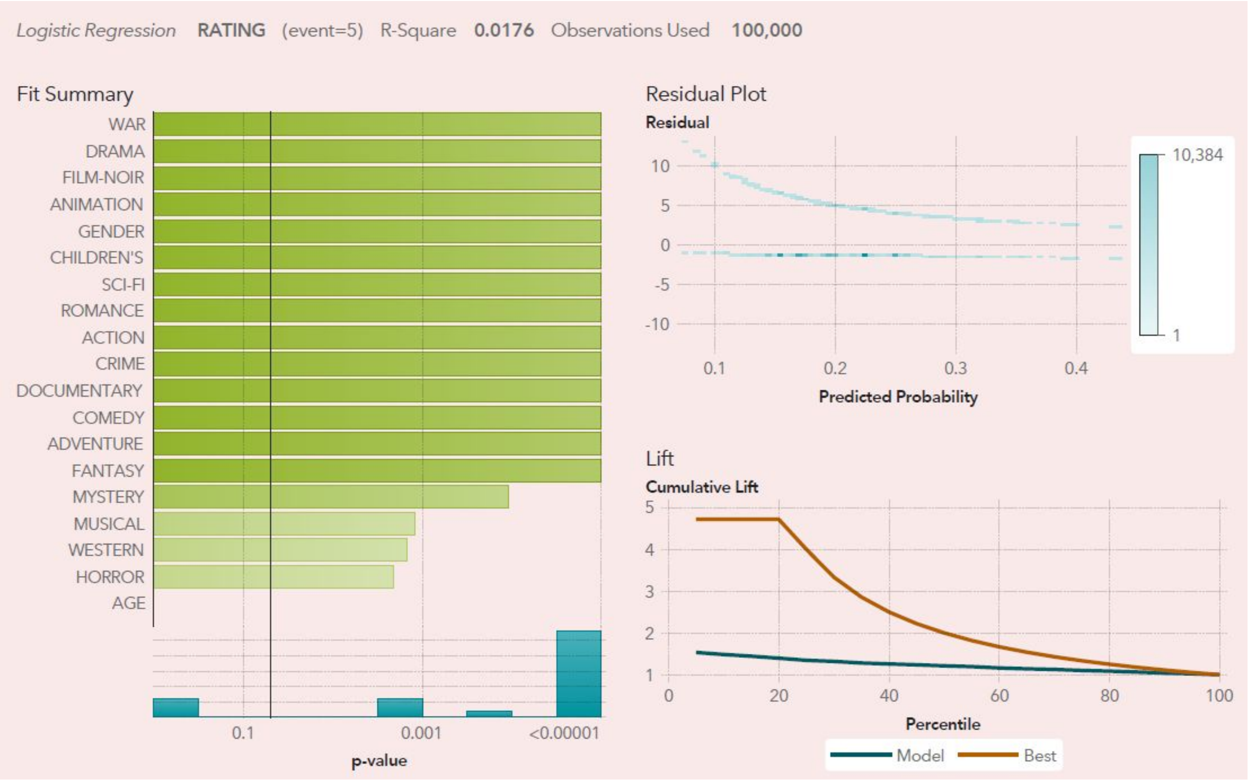
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

MODEL 2: LOGISTIC REGRESSION

We now begin with **Logistic Regression:** The standard algorithm to start with in order to build a supervised classification model. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Reasons to implement Logistic regression is that-

- It is very easy to interpret
- Gives us significant variables which others do not



Logistic Regression **RATING** (event=5) R-Square **0.0176** Observations Used **100,000**

We have used a logistic regression model to predict the probability or possibility of rating (1 through 5) for any movie based on certain variables.

Response:

Rating

Continuous effect:

Age

Classification effects:

Genres

Occupation

This model of logistic regression can be categorized as an ordinal logistic regression model.
The residual graph indicates the spread of residual values. Residual values are calculated as observed value - predicted value pairs. So if the model predicts a higher value than the observed value, then we get a negative residual. In the above plot, we observe that there are no outliers and the residual values are spread uniformly.

R-squared is a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.
The R-square value for this model is 0.0246.
Misclassification for this model is 0.2122.

RESULTS AND CONCLUSION

We have analyzed the dataset to better understand how the movie ratings can be predicted to either recommend movies to individuals or identify the target demographics of various types of movies.
Based on exploratory analysis, we find that -

- Males seem to give a higher rating more frequently than females. There is a significant difference in the percentage of 5-star ratings by both genders. Males are more likely to give a full rating of 5 and thus contribute to higher average ratings. Movies are likely to get higher ratings if the sample is dominated by males. This strategy can be used by movie marketers to artificially higher ratings than the competition.

- Most users are between 15-65 years of age. The proportion of younger people between 15- 35 is the largest demographic and thus this portion of the population should be targeted intensively for better results.

- The largest demographic is that of students which should be the main focus for a recommendation system. Professionals like Educators, engineers, programmers, and administrators should be targeted as they seem to consume more content.

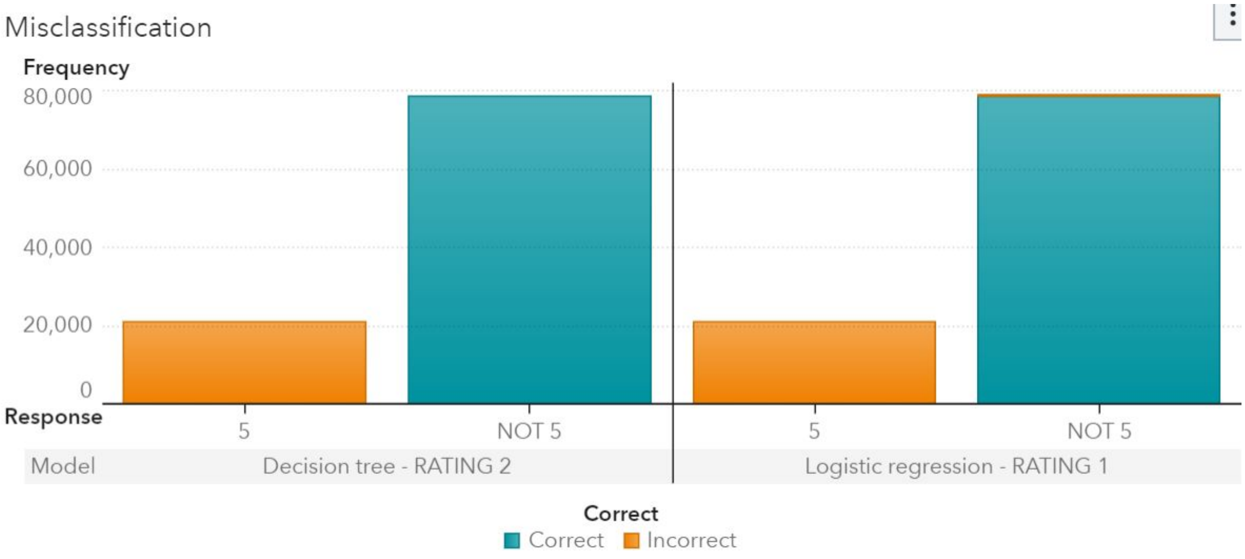
- Movies released after 1980 seem to be more popular. Post-1990 seems to be the most favored period. However, this could be attributed to the bias of surveyors.

- Out of all the occupations/ non-working, retired people, the retired people are the easiest to amuse, in both the categories and they barely dislike any kind of movie.

- Also, it is the hardest to fully satisfy retired personnel, in both the categories as studies show that this class of people rarely give a movie a 5 on 5 ratings.

Further based on the 2 models we can conclude-

- The misclassification rate for the decision tree is slightly lower than the logistic regression model. This indicates that the decision tree model predicts more values correctly than the logistic regression model. Hence, for our dataset, the decision tree model is better.


Misclassification
Frequency
80,000
60,000
40,000
20,000
0
Response
5
NOT 5
5
NOT 5
Model
Decision tree - RATING 2
Logistic regression - RATING 1
Correct
Correct
Incorrect