SVKM'S NARSEE MONJEE MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT & ENGINEERING, NMIMS



PROJECT REPORT

FOR

**DATA ANALYSIS**

**for**

**MovieLens Dataset**

**(Predictive Modelling)**

**By,**

**Aditi Anup (B001), Samreen Reyaz (B005), Dishant Bhatt (B013) and Gabriel Francis (B027)**

**Case Title: Analysis of 10,000 samples of MovieLens Data**

**College: Mukesh Patel School of Technology Management & Engineering, NMIMS**

# PROBLEM STATEMENT

The goal of this report is to FIND THE VARIOUS CORRELATIONS BETWEEN A PARTICULAR USER AND THE RATING THEY GIVE A MOVIE and to PREDICT THE RATING A PARTICULAR USER WOULD GIVE A MOVIE BASED ON ITS GENRE.

# UNDERSTANDING THE TRAINING DATA AT HAND

The data consists of three main tables. One stores all the movies that are involved in the data set, the second stores all the users involved in the dataset and the last stores the rating the users have given to a particular movie.

There are 100,000 ratings (1-5) from 943 users on 1682 movies and Each user has rated at least 20 movies.

The **movies table** has the following columns:

**movie id:** The primary key of the table

**movie title:** The title of the movie along with the year it was released

**release date:** The date the movie was released into theatres

**video release date:** The date the movie was released into home video formats

**IMDb URL:** A URL that points to the IMDb page of the movie

**Unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western:** Uses 1-Hot encoding the represent the genre of the movie

The **user table** has the following columns:

**user id:** The primary key

**Age:** The exact age of the reviewer

**Gender:** The gender of the reviewer

**Occupation:** The domain of occupation of the reviewer

**zip code:** The zip code of the location in which the reviewer decides

The **ratings table**:

**user id:** The foreign key from the users table

**item id:** The foreign key from the movie table

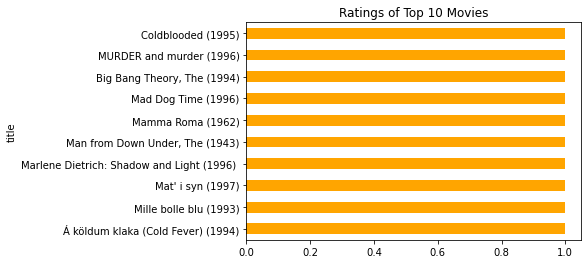
**Rating:** The rating given by a user for a movie

**Timestamp:** The timestamp of the time the rating was given

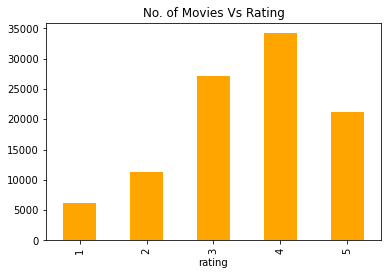
# EXPLORATORY DATA ANALYSIS AND VISUALIZATION

## GRAPHS

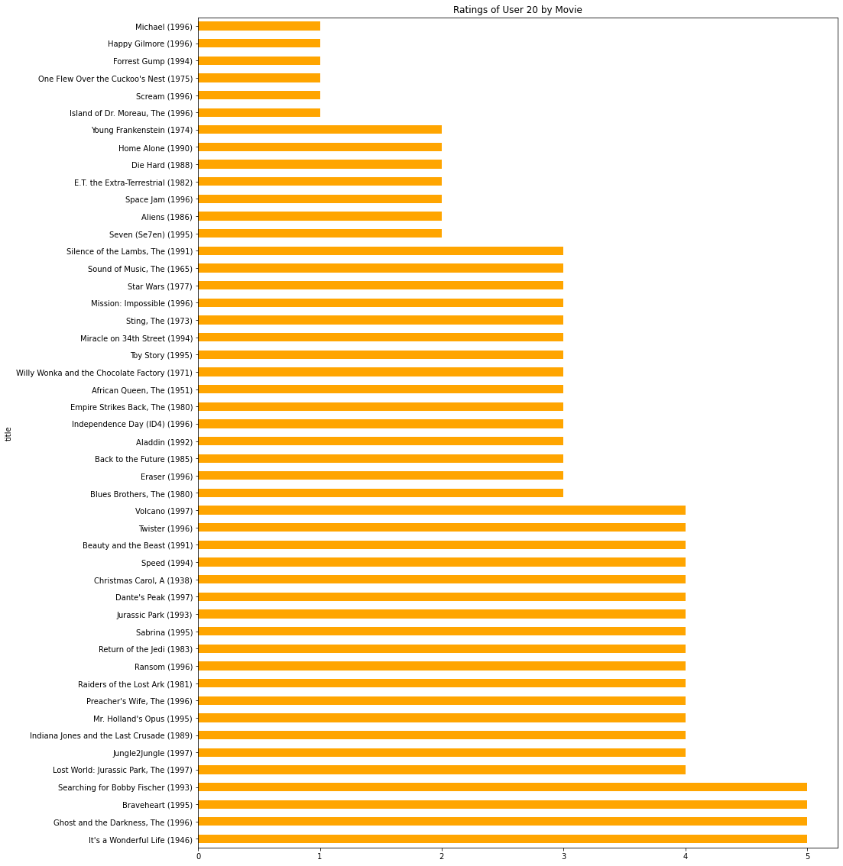
The below graph shows us the rating of the top 10 rated movies, here the ratings of all these movies are 1 or higher.



This graph shows us the no. of movies that have gotten a particular rating. E.g around 6000 movies got a rating of 1 of 5



A graph that shows the rating of a particular user (id: 20) to all rated movies:



Word Cloud of different genres of movies based on different ratings given to them, i.e., Highest rated genre (Film Noir) is shown the biggest.



# DATA CLEANING

The data we acquired from the source was mostly clean and required minimal cleaning work.

The release date column was found to have specific rows with incorrect values for month and incorrect and missing values for the year column. We had to remove these 9 rows and ended up with 99991 rows as opposed to the original 100k.

## COMMON DATA MANIPULATION

These are a set of steps we had to take to manipulate the data to make it fit for use in all the models:

1. Import all the datasets that we will be needed for further analysis and save them as DataFrames.
2. Frequency binning the ages into age groups as it will be easier for future analysis.
3. Merge all the datasets together based on a common column to make it into one DataFrame.
4. Drop the columns 'user id',  'item id', 'timestamp', 'movie id', 'movie title' and 'video release date'. This is because these columns are extremely specific and will not help us get generalized results when taking new values

# MODEL 1: LOGISTIC REGRESSION

As this can be considered to be a classification problem, we started with logistic regression.

## MOTIVATION

* Very easy to interpret.
* Gives us significant variables.  – which others don't

## ASSUMPTIONS

* Binary logistic regression requires the dependent variable to be binary, and ordinal logistic regression requires the dependent variable to be ordinal.
* Logistic regression requires observations to be independent of each other.  In other words, the observations should not come from repeated measurements or matched data.
* Logistic regression requires there to be little or no multicollinearity among the independent variables.  This means that the independent variables should not be too highly correlated with each other.
* Logistic regression assumes the linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

## RESULT

This method yields an accuracy of 0.3615361536153615.

# MODEL 2: DECISION TREES

As this can be considered to be a classification problem, we then tried to use decision trees.

## MOTIVATION

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 you to form a balanced picture of the risks and rewards associated with each possible course of action.

## ASSUMPTIONS

* No assumptions about underlying data
* It is a non-parametric model

## RESULT

 This method yields an accuracy of 0.36973697369736974

# MODEL 3: RANDOM FOREST

Finally, as our last classification method we tried to use random forest classification.

## MOTIVATION

* It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier.
* It runs efficiently on large databases.
* It can handle thousands of input variables without variable deletion.
* It gives estimates of what variables are important in the classification.

## ASSUMPTIONS

It has no model underneath, and the only assumption that it relies is that sampling is representative. But this is usually a common assumption. For example, if one class consists of two components and in our dataset one component is represented by 100 samples, and another component is represented by 1 sample - probably most individual decision trees will see only the first component and Random Forest will misclassify the second one.

## RESULT

This method yields and accuracy of 0.3844384438443844

# MODEL 4: MULTIPLE LINEAR REGRESSION

## MOTIVATION

* Has the ability to determine the relative influence of one or more predictor variables to the criterion value.
* Has the ability to identify outliers or anomalies.
* More comfortable to implement, interpret, and very efficient to train.
* It is prone to overfitting, but it can be easily avoided using some dimensionality reduction techniques, regularization techniques and cross-validation.

## ASSUMPTIONS

* The random x variables must be normally distributed and linearly related to y.
* The x values (independent variables) must be free of error.
* The variance of y (the dependent variable) as a function of the x variables must be constant. This is referred to as homoscedasticity.
* The x variables must be relatively uncorrelated.

## WORKING

* Step 1: Perform 1-hot encoding, which converts the categorical values into continuous values.
* Step 2: We separate the target variable from the dataset.
* Step 3:  Train and split the dataset, perform Linear Regression on the dataset, fit the dataset, and get the predicted values for the target variable.
* Step 4: Create the DataFrame, which contains the actual and the predicted value, and compares the accuracy of regression.
* Step 5: Calculate the r\_score, which shows the accuracy of Multiple Linear Regression.

## RESULTS

|  |  |
| --- | --- |
| The DataFrame containing the actual and predicted values for the target variable: | The r score value which shows how close the data are to the fitted regression line: |
|  |  |

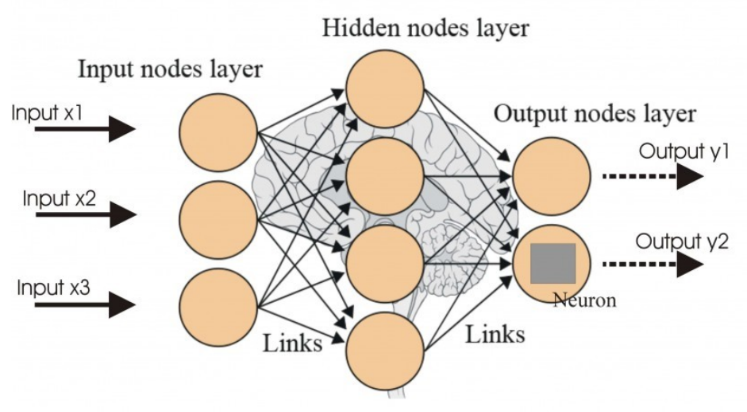
# MODEL 5: NEURAL NETWORK

## MOTIVATION

* Neural networks are complex models, which try to mimic the way the human brain develops classification rules.
* Parallel processing capability:  Artificial neural networks have numerical strength that can perform more than one job at the same time.
* After ANN training, the data may produce output even with incomplete information. The loss of performance here depends on the importance of the missing information.

## ASSUMPTIONS

* It relies upon the assumption that sampling is representative. However, this is usually a common assumption. If this is not the case, low accuracy may be obtained.
* It can only work with the numerical representation of data. Hence only numerical data can be fed to it.



## NEURAL NETWORK MODEL FOR DATASET

Shown in the figure above is a general design for an artificial neural network. It contains multiple layers, such as an input layer, an output layer, and zero or more hidden layers. Each of these layers contains nodes that are connected to nodes from other layers.

Every node receives inputs from the nodes in the layer before it. It then uses the input in some formula to generate an output which it passes to the nodes in the next layer after passing it through an activation function. The links connecting the nodes have individual weights attached to them, which are used in the calculation of output for the nodes.

During training, specific neural networks also use the difference between the expected output and the output generated by the network to retrain the network, i.e., update the weights connecting the nodes by a factor, known as learning rate, to get an output closer to the actual expected output. Training is done in iterations, known as EPOCHS, and continues until a certain number of iterations, or on receiving the desired output.

The working of our model is as follows:

1. Import the movie, user and rating datasets and give column names
2. Convert individual ages in user dataset to age groups
3. Join the three datasets based on primary and foreign keys
4. Convert required features to categorical representation
5. Drop columns such as primary key and unique features
6. Balance output classes using upscaling
7. Create a class for the neural network with desired activation function
8. Divide the dataset into batches to be given to the neural network at once
9. Train the network for the desired number of EPOCHS with desired learning rate
10. Run network with the testing dataset to get accuracy

## RESULTS

The neural network contains three hidden layers of 64 nodes each. It was trained for 10 EPOCHS. A learning rate of 0.001 was used. The following activation functions were used:

|  |  |  |
| --- | --- | --- |
| Activation  Function | Name | Formula |
| ReLU | Rectified Linear Unit | ReLU(*x*)=max(0,*x*) |
| ReLU6 | Rectified Linear Unit – v6 | ReLU6(*x*)=min(max(0,*x*),6) |
| SELU | Scaled Exponential Linear Unit | SELU(*x*)=*scale*∗(max(0,*x*)+min(0,*α*∗(exp(*x*)−1))) |
| CELU | Continuously Differentiable Exponential Linear Unit | CELU(*x*)=max(0,*x*)+min(0,*α*∗(exp(*x*/*α*)−1)) |
| GELU | Gaussian Error Linear Unit | GELU(*x*)=*x*∗Φ(*x*) |
| Sigmoid | - | Sigmoid(*x*)=1/(1+exp(−*x*)1) |
| Logsigmoid | - | LogSigmoid(*xi*​)=log(1/(1+exp(−*xi*​)) |

We obtained the following results:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Training and Testing Set | ReLU | ReLU6 | SELU | CELU | GELU | | Sigmoid | Logsigmoid |
| u1 | 0.321 | 0.3139 | 0.3137 | 0.3087 | | 0.3123 | 0.2952 | 0.3059 |
| u2 | 0.3122 | 0.3054 | 0.3096 | 0.3049 | | 0.3102 | 0.2947 | 0.3014 |
| u3 | 0.3073 | 0.3115 | 0.3082 | 0.3054 | | 0.3142 | 0.2833 | 0.3068 |
| u4 | 0.2994 | 0.2962 | 0.2917 | 0.3002 | | 0.2979 | 0.2768 | 0.3046 |
| u5 | 0.3129 | 0.2999 | 0.2881 | 0.3006 | | 0.3024 | 0.2808 | 0.2849 |
|  |  |  |  |  | |  |  |  |
| Average | **0.31056** | 0.30538 | 0.30226 | 0.30396 | | 0.3074 | 0.28616 | 0.30072 |

As we can see, the best accuracy obtained is **0.31056** when the ReLU activation function is used. To try and improve accuracy, we tried configuring the neural network with multiple hidden layers and different learning rates. But they did not give any better results.

The low accuracy could be due to the following reasons:

* A small sample size of only 100,000 entries
* Varied data where a good pattern cannot be found
* A low number of EPOCHS

# CONCLUSION

As we saw from the initial graphs, the data was widely varied in nature, as is usually expected from data based on personal opinions. Socially gathered data is tough to predict due to the large number of factors that affect the choice of an individual, and the data set usually does not provide all these parameters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Logistic Regression | Decision Tree | Random Forest | Linear Regression | Neural Network |
| Accuracy | 0.36153 | 0.36973 | 0.38443 | 0.26577 | 0.31056 |

From the above chart, we can see that Random Forest Classifier yields the best accuracy despite being more straightforward than a neural network and, in itself, not being a real classification algorithm. And while it is the best model, it still has a relatively low accuracy of <0.4, which can be attributed to the varied nature of users' interest in movies.