Capstone Project –

NETFLIX RECOMMENDATION SYSTEM

**Table of Contents :**

1. Problem Statement

2. Project Objective

3. Data Description

4. Data Pre-processing Steps and Inspiration

5. Choosing the Algorithm for the Project

6. Motivation and Reasons for Choosing the Algorithm

7. Assumptions

8. Model Evaluation and Techniques

9. Inferences from the Same

10. Future Possibilities of the Project

11. Conclusion

12. References

**Problem Statement:**

Customer Behaviour and its prediction lies at the core of every Business Model. From Stock Exchange, e-Commerce and Automobile to even Presidential Elections, predictions serve a great purpose. Most of these predictions are based on the data available about a person’s activity either online or in-person

Recommendation Engines are the much-needed manifestations of the desired Predictability of User Activity. Recommendation Engines move one step further and not only give information but put forth strategies to further increase users’ interaction with the platform.

In today’s world OTT platform and Streaming Services have taken up a big chunk in the Retail and Entertainment industry. Organizations like Netflix, Amazon etc. analyse User Activity Pattern’s and suggest products that better suit the user needs and choices.

**Project Objective:**

We will be creating one such Recommendation Engine from the ground-up, where every single user, based on their area of interest and ratings, would be recommended a list of movies that are best suited for them.

**Data Description:**

Project consist of 2 files

1. “combined.txt’ file which contains

* Customer ID : It consists of Unique ID of the Customer.
* Ratings : It tells what where the ratings given by customer to the movie. Ratings are on a five-star (integral) scale from 1 to 5.
* Movie ID : It consists of Unique ID of each movie.
* This file consists of 24058263 rows and 2 columns

2. “MOVIE TITLES “ file contains

* MOVIE ID: It consists of Unique ID of each movie.
* MOVIE NAME: IT consist of Names of The Movie.
* YEAR OF RELEASE: Consist of Year of release for each year.
* This file consists of 17770 rows and 3 columns

**Data Pre-processing Steps and Inspiration:**

* Data is encoded so we have to decode it first.
* Data needs to be pre-processed before going further.
* First of all, there are no names are given to the column so we assign them names
* Below movie ID there is customer id who have given rating for that movie .We could observe that wherever there is movie id ratings column in that row is null.
* We could create a custom loop which uses null values of rating column and find out the range of movie id for which customer have given ratings.
* We set the cutoff values for minimum rating for movie id (i.e. 1798) and minimum number of ratings (i.e. 52) that should be given by customer so that ,the data to be considered in dataset for model.
* Datatype of some columns where also changed from object to float

DATA INSIGHTS:

* Data consist of 4499 movies in the data set and total Customer count is 470758
* Movies with ratings 4 where the highest
* Total number ratings given by customer 24053764

**Choosing the Algorithm for the Project:**

* The Singular Value Decomposition algorithm was chosen because it is very effective in collaborative filtering.
* Collaborative filtering depends on user-item interactions to make recommendations and SVD specially is well-suited for this task.
* It factorises the user-item interaction matrix into three matrices, it also captures latent features of users and items.

**Motivation and Reasons for Choosing the Algorithm:**

* SVD is useful in handling missing data as recommendation system often contains missing data , which is a common problem in recommender systems.
* As it has ability to reduce dimensionality, it’s a best technique for large-scale recommendation systems.
* And also makes it great for preserving essential information and its complexity also get reduced
* As a result, it is successfully implemented in recommendation system or collaborative filtering applications in real world .

**ASSUMPTION:**

* The computation of the SVD decomposition can be efficiently handled but in real life computing the full SVD decomposition may become computationally expensive
* SVD-based models assumes that the latent features extracted are meaningful, even though they might not have clear meaning.
* Preference of users with is assumed to be same with time but in real life it may change with time

**Model Evaluation and Techniques:**

* The results of the model evaluation using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) for the Singular Value Decomposition (SVD) algorithm for 4 Folds.
* The RMSE values range from approximately 0.988 to 1.004, and the MAE values range from approximately 0.786 to 0.805. These metrics indicate that the SVD algorithm is making reasonably accurate predictions, as the errors are relatively low.

**INFERENCES FROM THE SAME:**

* Prediction by SVD model was great as seen by RMSE,MAE results.
* Time taken for training was good , so we can conclude the it can be suitable for large data sets
* There was variability in time during testing process so further strategies can be considered

**FUTURE POSSIBILITIES:**

* We can do Hyper-parameter tuning and could a possibly increase the accuracy and reduce variability seen during testing
* We could also explore more Deep learning model that could help us to obtain better results

**Conclusion:**

* Successful implementation of the SVD algorithm for building a recommendation system on the Netflix dataset
* The SVD algorithm was very effective in accurately predicting user preferences for movies .
* Model Performed very well this data as RMSE and MAE values very low
* Model can also perform great for large data set as training time was good
* There was variability in testing time which can be improved by various methods

**REFERENCES :**

# Dataset- Netflix Prize data (Kaggle)