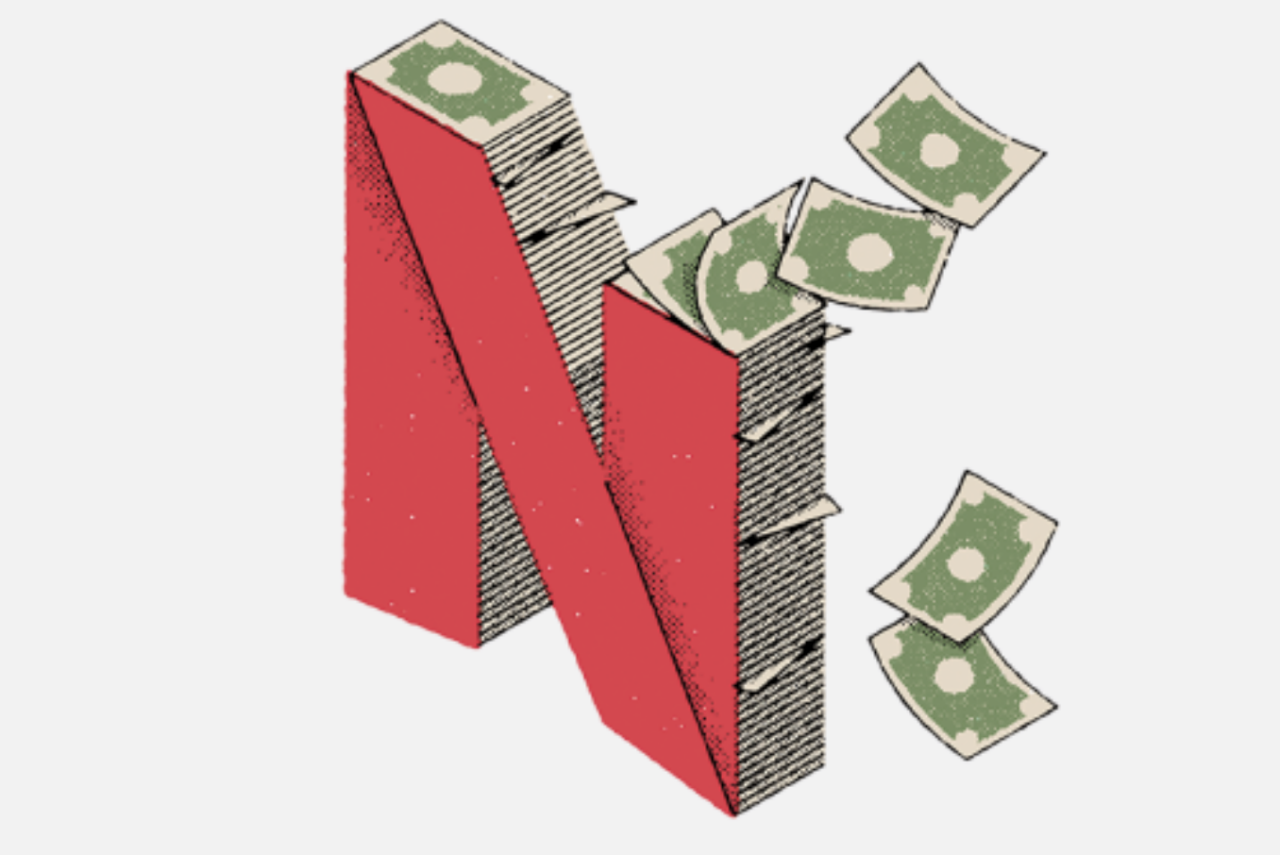
**Project Report: Recommendation System**



**The George Washington University**

Department of Information System and Technology Management

**——————————————————————————————————————**

ISTM 6290\_80: IoT Management

May 10th, 2020

Table of Contents

[1. EXECUTIVE SUMMARY: 4](#_Toc39954007)

[2. ABOUT NETFLIX: 5](#_Toc39954008)

[2.1 Overview 5](#_Toc39954009)

[2.2 Netflix Users and Revenue 5](#_Toc39954010)

[2.3 Competitors Analysis 5](#_Toc39954011)

[3. DATA DESCRIPTION: 6](#_Toc39954012)

[3.1 Data Source: 6](#_Toc39954013)

[3.2 Data Size: 6](#_Toc39954014)

[3.3 Data Format and Explanation: 6](#_Toc39954015)

[4. RESEARCH QUESTIONS: 7](#_Toc39954016)

[5. METHODOLGY: 8](#_Toc39954017)

[5.1 Data Preprocessing: 8](#_Toc39954018)

[5.2 Data Exploration: 9](#_Toc39954019)

[5.3 The Model 11](#_Toc39954020)

[6. TESTING AND EVALUATION: 12](#_Toc39954021)

[7. RESULTS AND FINDINGS: 13](#_Toc39954022)

[7.1 Confusion Matrix/ Violin Plot 13](#_Toc39954023)

[7.2 Root Mean Square Error (RMSE) and Accuracy 14](#_Toc39954024)

[7.3 Importance of Recommendation System: 15](#_Toc39954025)

[8. CHALLENGES AND LIMITATIONS: 15](#_Toc39954026)

[8.1 Gigantic Dataset 15](#_Toc39954027)

[8.2 Data Sparsity 16](#_Toc39954028)

[9. CONCLUSION: 16](#_Toc39954029)

[10. APPENDIXES: 16](#_Toc39954030)

[10.1 Appendix (A) - Project Code 16](#_Toc39954031)

[10.2 Appendix (B) – Results and Findings 16](#_Toc39954032)

[11. REFERENCES: 16](#_Toc39954033)

Table of Figures

[Figure 1: Customer Ratings Dataframe 7](#_Toc39937368)

[Figure 2: Combined Dataset with Customers and Movies 8](#_Toc39937369)

[Figure 3: Ratings per Customer and Movies 9](https://gwu0.sharepoint.com/sites/NetflixProjectIoT-GRP/Shared%20Documents/General/ProjectReport.docx#_Toc39937370)

[Figure 4: Ratings per Weekday (Left) and Ratings per Month (Right) 9](#_Toc39937371)

[Figure 5: Word Clouds showing the Top Rated and Low Rated Movies 10](#_Toc39937372)

[Figure 6: 10 Most Frequently Rated Movies and their Ratings Distribution 11](#_Toc39937373)

[Figure 8: Building the Recommendation Model on the Training Dataset 12](#_Toc39937374)

[Figure 9: Recommender Function 12](#_Toc39937375)

[Figure 10: Results of Recommender Function 13](#_Toc39937376)

[Figure 11: Confusion Matrix for First Dataset 13](#_Toc39937377)

[Figure 12: Violin Plot on First Dataset 14](#_Toc39937378)

[Figure 13: RMSE and Accuracy on Different Datasets 15](#_Toc39937379)

# EXECUTIVE SUMMARY:A screenshot of a cell phone Description automatically generated

With every day developing online businesses and stores, the usage of recommendation engines has become a significant part of the e-commerce industry.

What exactly is a Recommendation Engine or System in the world of Artificial Intelligence/Machine Learning?

We have always used products or did things that our friends and colleagues recommended to us based on our tastes and preferences in life, and to solve the same purpose, instead of friends and family, there is an automated engine that tells us what will interest us, based on our history of purchases or activities.

A Recommendation system is an extensive class of web applications that involves predicting the user responses to the options. These are simple algorithms which aim to provide the most relevant and accurate items to the user by filtering useful stuff from a huge pool of information base. Recommendation engines discovers data patterns in the data set by learning consumers choices and produces the outcomes that corelates to their needs and interests.

Recommendation systems are quickly becoming the primary way for users to expose to the whole digital world through the lens of their experiences, behaviors, preferences, and interests. And in world of information density and product overload, a recommendation engine provides an efficient way for companies to provide consumers with personalized information and solutions.

This project aims at building an effective Recommendation Engine for the streaming giant Netflix, so that, instead of having to browse through thousands of box sets and movie titles, the consumer can get a much narrower selection of items that he is likely to enjoy. This will save customer’s time and delivers a better user experience, hence there will be low cancellations, and a lot of money will be saved by the company.

A Recommendation Engine can significantly boost Click-Through Raters (CTRs), conversions, and other essential metrics. It can have positive effects on user experience, thus translating to higher customer satisfaction and retention.

# ABOUT NETFLIX:

## 2.1 Overview

Netflix is a media service provider and production company, founded in 1997. The company business model is based on subscription for online streaming services in US and international market and is currently available in more than 200 countries. Netflix has over 182 million subscribers around the world, as of April 2020.

## 2.2 Netflix Users and Revenue

The top three countries subscribers and revenue generator in 2019:

1. United States has solely generated more than 8.5 Billion revenue by 62.7 million subscribers, which is almost half of the Netflix overall revenue
2. Australia accounts for 1.5 Billion revenue and more than 14.4 million subscribers
3. United Kingdom generated around 1.3 Billion revenue by 12.5 million subscribers

## 2.3 Competitors Analysis

Netflix has been facing increasing competition in online streaming and content generating. Along with Amazon and Hulu there are growing number of new streaming services, such as, Disney+, HBO, CBS and Apple. The competitors are either competing in the online streaming, such as Hulu, or mainly in content generating or acquisitions like Disney and Warner Media. However, Disney lately released its new product, Disney+, to compete not only with the content generating, but also with the online streaming services. In the same fashion, Amazon started the battle with only online streaming, and it moved toward content generating, which urge Netflix to keep its foot pressed strongly on the online streaming battle. As a result of this rigorous competition, Netflix has been investing giant amount of money not only in content, but also in complex algorithms to enhance its services and products.

# DATA DESCRIPTION:

## Data Source:

The data that we have used to build the Recommendation Engine is from the famous Netflix Prize Open Competition, that the company held in 2006 to 2009 for the best algorithm to predict user ratings for films.

Link: <https://www.kaggle.com/netflix-inc/netflix-prize-data>

## Data Size:

5 GB (Four Customer Data Files and One Movie Data File)

## Data Format and Explanation:

**Customer Data File Description:**

**Training Dataset consists of 4 text files in the following format:**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Treatment** |
| MovieID | These are the ID given to specific movies ranging from 1 to 17770 sequentially | Integers |
| CustomerID | These are the ID given to specific customers ranging from 1 to 2649429, with gaps. There are 480189 users. | Integers |
| Ratings | These are on a five-star (integral) scale from 1 to 5 | Integers |
| TimeStamp | These are Dates in the format YYYY-MM-DD. | Date Format |

**Movies File Description:**

Movie information in "movie\_titles.txt" is in the following format:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Treatment** |
| MovieID | These are the ID given to specific movies ranging from 1 to 17770 sequentially | Integers |
| Movie\_Title | These are the Netflix movie title and may not correspond to titles used on other sites. Titles are in English. | String |
| YearOfRelease | These can range from 1890 to 2005 and may correspond to the release of corresponding DVD, not necessarily its theatrical release. | Integers |

Combining the 4 text files gave us 100 million rows of data. Below is a sample:

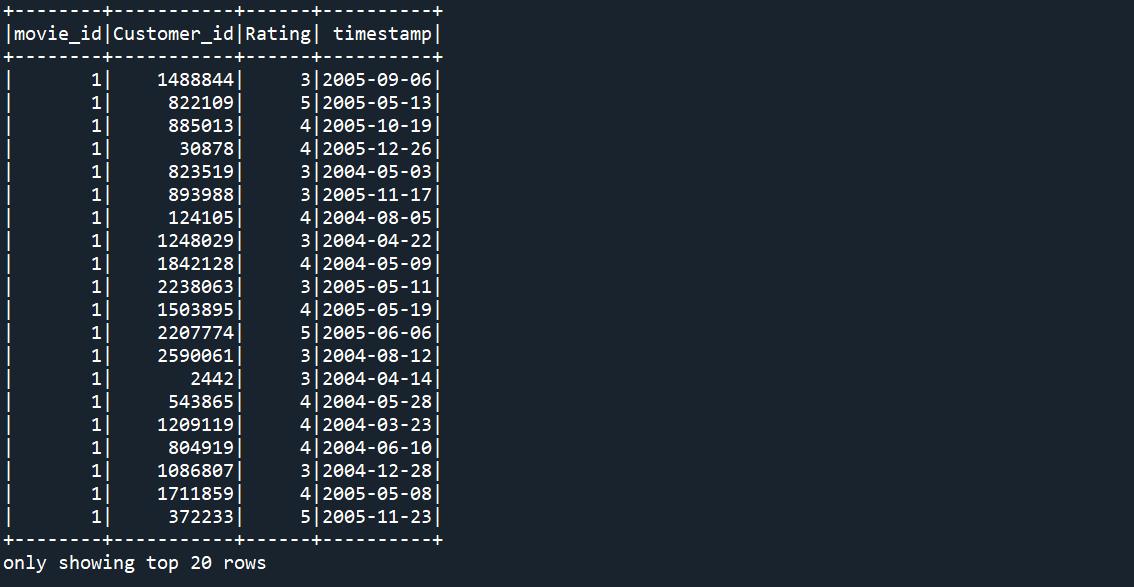


Figure 1: Customer Ratings Dataframe

# RESEARCH QUESTIONS:

Our main objective for this project is to build a Movie Recommendation system for Netflix, so that the users are recommended the movies they haven’t rated and are most likely going to like based on their viewing history and data from other similar users.

In addition to this, the following questions play an important part in our study:

1. Why is a Recommendation System important for a streaming company like Netflix?
2. How can a Recommendation system help in improving business?
3. How Collaborative Filtering can be a useful method to build a Recommendation system?

# METHODOLGY:

To achieve our objective of building a recommendation model, we followed:

1. Data Preprocessing and Cleaning
2. Data Exploration
3. The Model

## Data Preprocessing:

We have four files of Customer Data, with the following format:

**MovieID:**

**CustomerID, Rating, TimeStamp**

So, we had this challenge of creating a spark dataframe with a separate MovieID column with all the other columns of CustomerID, Rating and Timestamp.

Also, there were 7 missing YearOfRelease in the Movie\_Titles file, which were fixed manually.

Then all the files were combined to form a single spark dataframe which looked like:

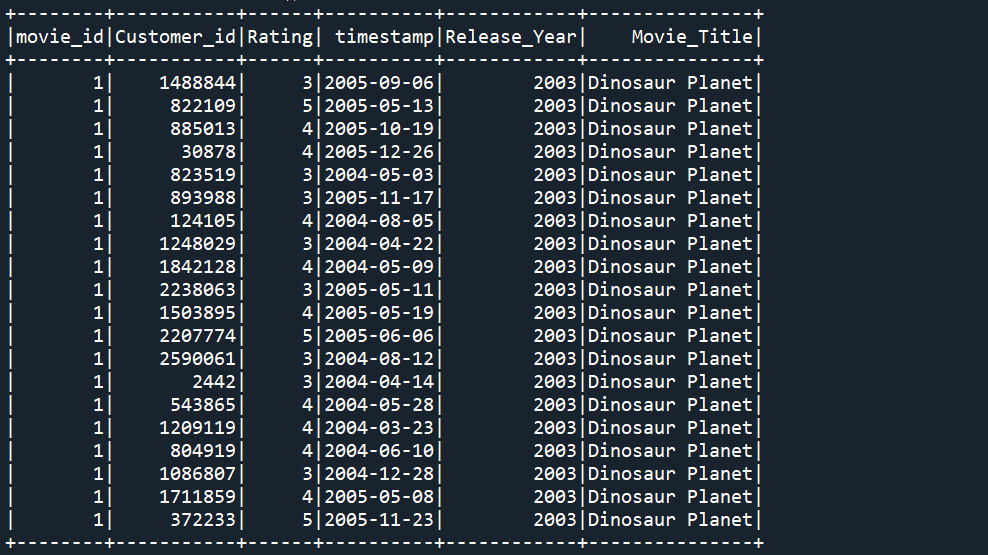


Figure 2: Combined Dataset with Customers and Movies

## Data Exploration:

Though we had a limited dataset, not in terms of size, but in terms of value. Still, we tried to make the most out of it.

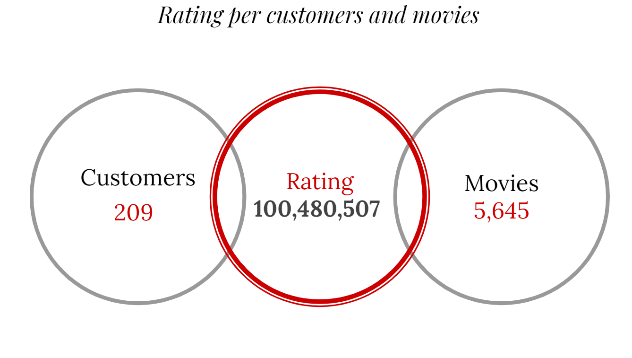
We tried to get deeper insight into the data.

Figure 3: Ratings per Customer and Movies

There is a total of 100,480,507 ratings for 17,770 movies by 480,189 Customers. So, on an average a customer rates 209 movies and each movie is rated by 5645 customers.

Since we were working on spark dataframes, we used SQL Queries to get more information about the data. We were able to discover a few patterns in the data and extract information about the movies and their rating patterns.

By using the timestamp column, which is the date when the rating was given, we extracted the days of week and months of the year to see which days and which months are the ones when most customers rate the movies, that can help the company plan when should they be uploading more content to get better viewership.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 4: Ratings per Weekday (Left) and Ratings per Month (Right)

We also got an insight into which were the movies that were mostly given a Rating of 5 and which were the movies that were mostly rated 1, as per our dataset. So, we thought of plotting word clouds for the top and the bottom 50 movies in accordance to their rating pattern.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 5: Word Clouds showing the Top Rated and Low Rated Movies

From these word clouds it is abundantly clear, that “The Godfather” is an all-time favorite and is mostly rated 5 by the customers. On the other hand, “The Royal Tenanbums” is the movie that mostly gets a rating of 1.

Another question that came to our mind was to know which are the most frequently rated movies in our dataset and what ratings are mostly given to them.

Here is the graphical representation of the 10 Most frequently rated movies and distribution of their ratings:

A picture containing game

Description automatically generated

Figure 6: 10 Most Frequently Rated Movies and their Ratings Distribution

Clearly from the graph above, “The Godfather”, “Forest Gump” and “Pirates of Caribbean” are the most loved movies of all times.

## The Model

To achieve our objective of creating a Recommendation system, we resorted to the Collaborative Filtering Technique of Pyspark and tried building the **Alternating Least Squares (ALS) Model.**

There were 100 million rows in the Combined Data file, ALS was not able to handle that much data altogether. Hence, we had to build ALS with single customer data file (with 25 million rows) and get an idea of the performance of our model.

We first divided the dataset into 80% and 20% for the Training and Validation Datasets respectively, setting the random seed value to “12345”, for result replication.

ALS is a preset model for Collaborative filtering in Spark, where we must train the model on the user, item and the ratings, as per our dataset.

The parameter **ColdStartStrategy** was set to “**drop**” to avoid any discrepancies in our results due to some non-rated or less rated movies.

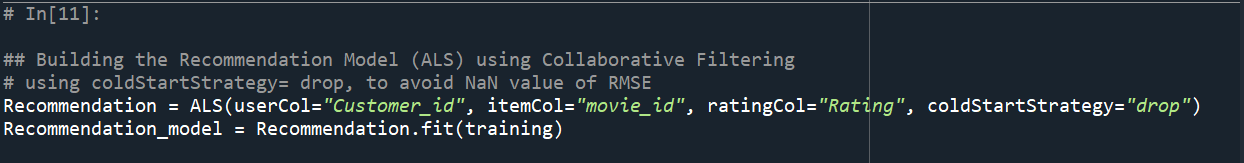


Figure 7: Building the Recommendation Model on the Training Dataset

# TESTING AND EVALUATION:

We trained our model separately on all the four customer data files and plotted confusion matrices, Violin Plots for our Actual and Predicted data, and calculated Root Mean Squared error for all the iterations, in order to get a clear picture of model’s performance.

We also tried evaluating our model by building a recommender function, which accepts CustomerID and number of movies (N), returns predicted ratings for N movies for that specific customer.

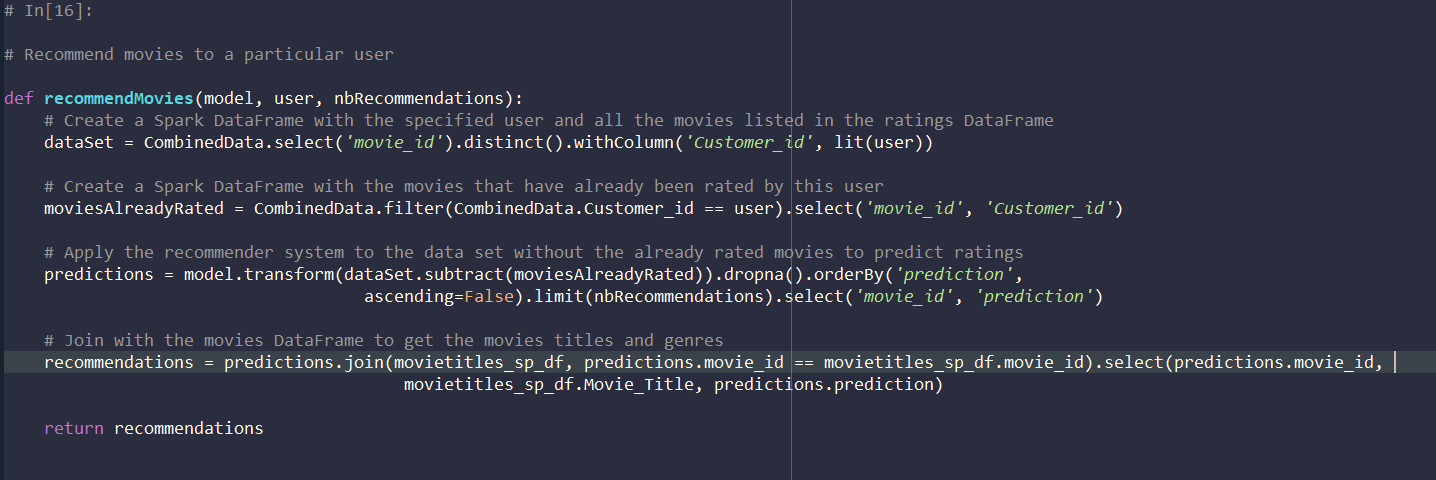


Figure 8: Recommender Function

We tested our Recommender Function with CustomerID 2442, and number of recommendations is 10. Here is the result:

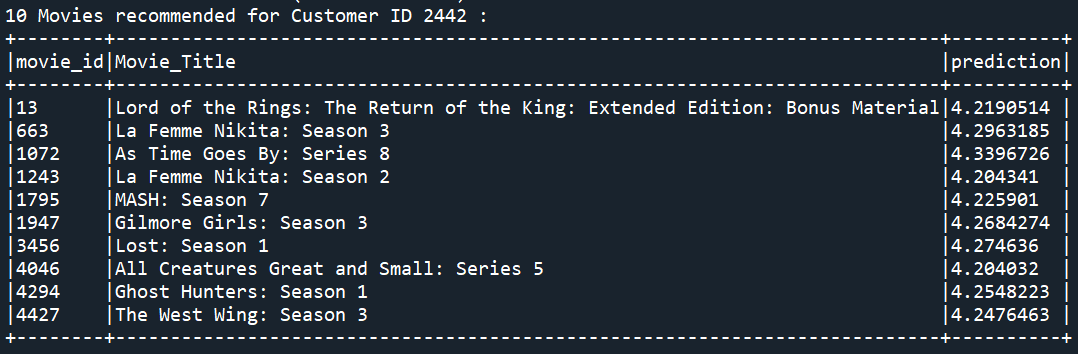


Figure 9: Results of Recommender Function

# RESULTS AND FINDINGS:

## Confusion Matrix/ Violin Plot

Since we trained our model on separate datasets, there were four plots /matrices (please refer Appendix B), below is the one for the first Customer ratings file:

A screenshot of text

Description automatically generated

Figure 10: Confusion Matrix for First Dataset

A close up of a logo

Description automatically generated

Figure 11: Violin Plot on First Dataset

The Confusion Matrix and the Violin Plot above tells us that our predictions for the ratings 3 and 4 are more accurately predicted, if we look at the medians for these values, they are pretty much there. Whereas the predictions for the ratings 1, 2 and 5 are all over the place.

This can be **a limitation of Collaborative Filtering**, that we only get to know about the ratings of users to particular movies, it doesn’t take into account all the other factors such as watching hours, total time of streaming, the genres, and all the other details of a customer’s watching history. All this information can play an important role in delivering better recommendations and predicting more accurate ratings.

## Root Mean Square Error (RMSE) and Accuracy

While training separate models for separate Customer Data Files, the RMSE and Accuracy was almost constant.

**RMSE: 0.88**

**ACCURACY: 32%**

Figure 12: RMSE and Accuracy on Different Datasets

As seen above, the general trend with each dataset separately gave us the idea how the recommendation system will behave with the complete dataset.

## Importance of Recommendation System:

Recommendation System or Recommendation Engine plays an important role for an online streaming company like Netflix as it helps in retaining the customers by engaging them and producing content that interests them. Having one will provide personalized menus to the customers, hence improving interactions and relationship with them.

It also helps to make Strategic Business Plans, as to what content should they invest on, what should be dropped off the list, when should they upload new content and what are the popular genres they might include.

# CHALLENGES AND LIMITATIONS:

## Gigantic Dataset

The Dataset in question when combined consists of over 100 million rows of data, which led to errors (Java Heap Space Errors) at various occasions while processing.

This is a problem associated with recommendation algorithms (Refer 7) because computation normally grows linearly with the number of users and items. Thus, it is crucial to apply recommendation techniques which can scale up in a successful manner as the number of datasets in a database increase.

## Data Sparsity

The Dataset is limited to CustomerID, MovieID, Ratings and Timestamps. Whereas, there are a lot of other factors that need to be considered while building a recommendation system, like but not limited to, the movie genres, the viewing pattern, location, and age of the customer, which play an important role in building a recommendation system with higher accuracy.

# CONCLUSION:

From this study, we understood the benefits of a Recommendation System, which is a powerful tool in assessing customer’s needs and choices. Using this system can help us generate a more personalized way of interacting with the customer hence leading to better relationships.

But, at the same time, since the accuracy of the model is not too good, it can be treated as one out of many parameters in the corporate decision making process. The company should not solely rely on this model, it needs supplemental information and data to lead to better results.

# REFERENCES:

1. “A simple way to explain the Recommendation Engine in AI” by Roger Chua (<https://medium.com/voice-tech-podcast/a-simple-way-to-explain-the-recommendation-engine-in-ai-d1a609f59d97>)
2. “Introduction to recommendation systems and How to design Recommendation system, that resembling the Amazon” by Madasamy M
3. (<https://medium.com/@madasamy/introduction-to-recommendation-systems-and-how-to-design-recommendation-system-that-resembling-the-9ac167e30e95>)
4. <https://www.smarthint.co/en/ai-product-recommendation-engine/>
5. https://variety.com/2020/digital/news/netflix-2020-content-spending-17-billion-1203469237/
6. <https://www.comparitech.com/tv-streaming/netflix-subscribers/>
7. “Recommendation systems: Principles, methods and evaluation” by F.O.Isinkaye, Y.O.Folajimi, B.A.Ojokoh

(<https://www.sciencedirect.com/science/article/pii/S1110866515000341>)

# APPENDICES:

## Attached Appendix A: The Project Code

## Attached Appendix B: Results and Graphs