Netflix Project Report

Group Members
Marcos Diaz
Olajide Fiyinfoluwa

Table of Contents

Chapter One

1.1 Introduction

In the ever-evolving world of digital entertainment, streaming services like Netflix have become integral to our daily lives, offering a vast sea of cinematic experiences at our fingertips. However, this abundance of choices often leads to a paradox of choice, where users find themselves overwhelmed, unable to decide what to watch next. Our project aims to address this dilemma by developing a movie recommendation system tailored for Netflix users.

The core challenge in this domain is navigating the expansive and diverse catalogue of movies available on Netflix and presenting users with options that resonate with their tastes and viewing history. The motivation behind this project is twofold: firstly, to enhance the user experience by making movie discovery effortless and personalized, and secondly, to assist Netflix in increasing viewer engagement and satisfaction, ultimately driving their platform's success.

1.2 Motivations

As we know that this kind of project has several related works, We would like to have our approach to this problem as an exercise to grasp and apply all the contents from this course.

Enhanced User Experience**:** By providing personalized movie recommendations, we can enhance the user experience on the streaming platform, making it easier for users to find content they will enjoy.

Increase User Engagement: Users are more likely to stay engaged with a platform that understands their preferences and provides relevant content.

Business benefits: Streaming platforms can benefit from user retention.

Chapter Two

2.1 Related Work

Part of Netflix's success came from its sophisticated recommendation system called Cinematch. Cinematch suggested to customers what they might like to order next, based on viewer choices and ratings. To grow the business further, Netflix decided it needed to improve recommendations further, so it announced a public competition, the Netflix Grand Prize. The goal was to create a recommendation system that was 10% better than Cinematch. It offered a $1M prize for the winner [1].

On June 26, 2009, a team called BellKor's Pragmatic Chaos which was composed of Chris Volinsky and his AT&T colleagues Robert Bell and Yehuda Koren, along with four other engineers from the United States, Austria, Canada and Israel was a combination of three teams: Pragmatic Theory, BellKor and BigChaos won the prize with a better algorithm than the current [2]. Despite that, Netflix did not end up using the final winning version. The version is a combination of different algorithms into a very complex solution where one of the teams, Big Chaos, documented their contribution to the Grand Prize Solution [3].

Our idea with this project is to solve this challenge using the knowledge acquired in this course as an exercise of knowledge discovery to offer recommendations.

Chapter Three

3.1 Methods

We used Google Colab and a dataset from https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data. Our pipeline consists of the following

1. Data collection
2. Data preprocessing,
3. Data cleaning and visualization,
4. Cross-validation,
5. Training and testing.

3.2 Data Collection

Our dataset, which is the Netflix prize dataset, consists of two types of files. The training dataset contains the following fields:

1. MovieIDs ranging from 1 to 17770 sequentially.
2. CustomerIDs ranging from 1 to 2649429, with gaps. There are 480189 users.
3. Ratings are on a five-star (integral) scale from 1 to 5.
4. Dates have the format YYYY-MM-DD.

The second data contained in our dataset is the movie titles, which consist of the following:

1. MovieID do not correspond to actual Netflix movie IDs or IMDB movie IDs.
2. YearOfRelease can range from 1890 to 2005 and may correspond to the release of the corresponding DVD, not necessarily its theatrical release.
3. The title is the Netflix movie title and may not correspond to titles used on other sites. Titles are in English.

3.2.1 Considerations

1. We chose two million rows of the dataset due to memory and processing limitations.
2. We ignored the date columns while carrying out the project.

3.2.2 Data preprocessing

We transformed the dataset into a standardized Pandas data frame to initiate the cleaning process. Given the large amount of data and the compute resource limitation of Google Colab, we focused on one of the four training data files. The preprocessed data was transferred into another output CSV file.

3.2.3 Data cleaning and visualization

We carried out the data cleaning by checking for missing values, duplicates, and null values. We dropped empty rows if found in the training dataset outputted from the data preprocessing task.
After cleaning the data, we wanted to view the ratings of the users; therefore, we created two histograms using the matplotlib library, which showed how ratings were distributed.

3.2.4 Cross Validation

We discovered a library called Surprise, which is very focused on recommendation systems. We used it to perform cross-validation on our given dataset. RMSE (Root Mean Square Error) and MAE (mean absolute error) were the measures used in the cross-validation process, and SVD (Singular Value Decomposition), NMF (non-negative matrix factorization) and SVDPP (an

extension of SVD) were candidates. After the cross-validation, we discovered that SVD was the best performer.

Chapter Four

4.1 Training and Testing

After the cross-validation evaluation, we discovered that the SVD model was the best fit for the project since it had the best Root Mean Square Error (RMSE). We chose SVD and started the training process. After completion of the training, we tested the model using a randomly selected user. Finally, we created a function to return the top ten recommendations given a user ID.

4.2 Results and Discussion

More accuracy can be achieved if we work with all the files. However, due to the limitation of the Google Colab, we were restricted to two million rows of the dataset. Since these types of algorithms are very focused on recommendation systems, they give you a pretty good result.

4.3 Conclusions

1. The surprise library used in this project is accurate and optimized for the recommendation system.
2. We recognize the pivotal role of prediction systems in shaping the future. By harnessing the power of data, analytics, and cutting-edge technology, we aim to contribute to a world where foresight is not just a luxury but an essential tool for progress and resilience.
3. We handled many data and noticed that a modification to a parameter significantly increased the time taken to train the model.

4.4 Future Work

We can process all the data in the dataset to train the model more accurately using a paid Google Colab tool that gives more memory and processor or an alternative application. We can also increase the number of epochs in training the model.

References

[1] Was Rahman, "The Netflix Prize - How Even AI Leaders Can Trip Up," https://towardsdatascience.com/the-netflix-prize-how-even-ai-leaders-can-trip-up-5c1f38e95c9f.
[2] Dan Jackson, "The Netflix Prize: How a $1 Million Contest Changed Binge-Watching Forever," https://www.thrillist.com/entertainment/nation/the-netflix-prize.
[3] A. Töscher and M. Jahrer, “The BigChaos Solution to the Netflix Grand Prize,” Oct. 2009.
[4] Verma Yugesh, "Beginners Guide To Truncated SVD For Dimensionality Reduction," https://analyticsindiamag.com/beginners-guide-to-truncated-svd-for-dimensionality-reduction/.