Case Study

Netflix Movie Recommendation System

A screenshot of a computer

Description automatically generated with low confidence

DSC630-T301 Predictive Analytics

Bellevue University

Sowmya Chavali

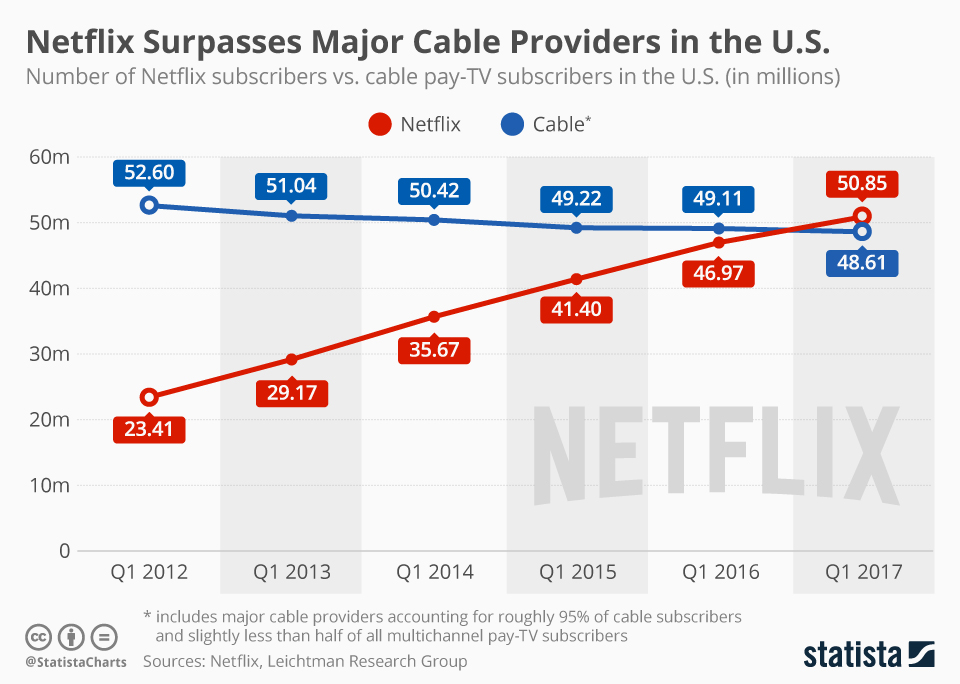
**Overview**

Different people have different taste in movies, and this is not reflected in a single score that we see when we google a movie. We use machine learning to build a personalized movie recommendation system based on user’s previous movie ratings. Our movie scoring system helps users instantly discover movies to their liking, regardless of how distinct their tastes may be. Current recommender systems generally fall into two categories: content-based filtering and collaborative filtering.[1]

I would want to experiment with collaborative filtering approach in this project(user and Item based). For collaborative filtering, the input to our algorithm will be the observed users’ movie rating. To achieve this, I plan on using KNN and matrix factorization to predict user’s movie ratings. A random recommender will also be built to prevent recommending already seen movies.

**Introduction**

TV and video streaming services have been around for a while with early services like Netflix, Amazon video, Hulu etc., popping up between 2006 and 2008. Let’s be honest though, the last decade is when streaming truly became mainstream. It’s been an amazing decade for streaming services. Nowadays, many people want to watch TV shows or series anytime and anywhere they want. In recent years, online TV has experienced exponential growth. To be exact, regarding the Digital Democracy Survey by Deloitte [2], which is an annual survey about changes in the digital environment, 49% of the United States households are subscribed to one or more streaming video services in 2016, compared to 31% in 2012.



Netflix vs Cable subscriptions in US[3]

Recently, the number of Netflix subscriptions within the United States exceeded the number of subscriptions for regular paid cable TV. Every day, over 125 million hours of video is watched on Netflix, and the number of titles keeps increasing. From these numbers, one can conclude that Netflix collects a lot of data, which can be used in many ways. For example, they can analyze data to increase the revenue, for marketing purposes, and to improve their customer satisfaction.

**Data Understanding**

The dataset I plan to use in this research comes from an open machine learning competition, called the Netflix Grand Prize. This competition started in October 2006 and lasted till 2009. The main goal of this competition is to replace their existing system, called Cinematch. The dataset contains a total of 100,480,507 ratings, based on 17,700 movies which come from a total of 480,189 users from the United States. Of each movie, titles and corresponding year of release are available. Besides, every movie had a unique movie ID, which was a sequence from 1 to 17,700. Next, of each movie a text file was provided, consisting of the rating of a user for the specific movie, the date of the rating, and a user ID.

Source - https://www.kaggle.com/netflix-inc/netflix-prize-data?select=README

**Data Preparation**

As part of data preparation, I plan on removing all movies that contained any NA values in the data. Unknown movies that have not many ratings, and therefore has almost no impact on the overall structure and distribution of the dataset. After removing the NAs, I may take a closer look at the reliability of ratings and movies. The methodology behind this approach is that a movie must have a certain number of ratings before the overall rating becomes reliable. If a movie is rated just 2 times, it might be biased and not give a good representation. In the last step of the data selection, I may take a look at the quality of the remaining ratings of users. Since the purpose of this research is to recommend movies, each user should have rated a movie with at least one good rating (a rating of 4 or 5 stars). If this is not the case, thus when a user has only rated movies with 3 or fewer stars, it is not usable for a recommender system, since recommender systems are built on good ratings.

**Data analysis**

Before one is able to make an initial recommendation based on historic ratings, one must get more insight of the data. Therefore, a data analysis is done to get more acquainted and familiar with the reduced and selected data. To start, the distribution of ratings is important for the recommender system: it is essential for such a system that there is some kind of diversity in this distribution.

**Model**

In the next phase of this research, I will build the model. First, I would be creating a random recommendation to create a benchmark. Next, I plan on using more advanced techniques to achieve more accountable recommendations.

Random recommendations -

The first recommendation system I plan to build is a random recommendation. In this system, random movies will be determined from the movie dataset to the users in the test set. In other words, this system does not consider the historic rating behavior of the user. Note that the set from which movies will be picked, excludes the already seen movies. This is done to prevent recommending already seen movies.

Item-based Collaborative Filtering -

After building a basic random recommender, I would create more explainable and accountable recommendations by more advanced techniques. Item-based collaborative filtering technique is a well-known and widely used recommender technique. It is a technique that produces recommendations based on the relationship between movies inferred from the rating matrix.

Timeline

Description automatically generatedTimeline

Description automatically generated

[4]

User based collaborative filtering –

With the movie IDs in our input, we can now get the subset of users that have watched and reviewed the movies in our input. Following by create subgroups, that is, grouping all the found users based on their user ID to find the topmost similar users to the input user. Highly similar movie rating behavior between users is detected by Netflix recommendation engine. If one person in the group has watched/given high rating to the movie Inception, Netflix recommends the movie to others in the subgroup.

**Expected Results**

This project will build a Netflix movie recommendation system project helps users instantly discover movies to their liking, regardless of how distinct/similar their tastes may be.

**References:**

1. <https://towardsdatascience.com/introduction-to-recommender-systems-1-971bd274f421>
2. <https://www2.deloitte.com/content/dam/Deloitte/za/Documents/technology-media-telecommunications/ZA_Deloitte_Digital_Democracy_Survey_Final.pdf>
3. <https://www.statista.com/chart/9799/netflix-vs-cable-pay-tv-subscribers/>
4. <https://medium.com/@toprak.mhmt/collaborative-filtering-3ceb89080ade>