Netflix Rooms: MMAI 5040 Group Project Proposal

Agnel Mamachan (218596346), Bintu Alkassoum (219116128), Krystaleah Ramkissoon (213663604), Pooja Ramakrishna Ganiga (219118678), Somiya Jena (215213051), Yiman (Catherine) Zhou (218525485), Zhiyi Wang (218718726)

Schulich School of Business at York University

MMAI 5040 Business Applications of Artificial Intelligence

Professor Divinus Oppong-Tawiah, Ph.D

February 28, 2022

Organization Overview: Netflix Inc.

For this project, we will be focusing on Netflix Inc. Netflix has become one of the most successful global streaming services (See Exhibit 4; Rivera, 2019). As of 2020, Netflix was the largest subscription-based video streaming service globally with 195 million subscribers and $70 billion in content spending (Business of Apps, n.d.).

Problem Statement

In a survey conducted by Statistics Canada (2020), nearly 24% of 46,000 Canadians surveyed reported their mental health as "fair or poor" compared to 8% in 2018 (Rahman & Arif, 2021). During a time plagued by loneliness, uncertainty, and isolation, Netflix has the opportunity to address the well-being of its users, while also providing stakeholders with added company value. Due to the competitive video streaming landscape, Netflix must continue to enhance the customer experience to maintain its competitive advantage **(See Exhibit 4)**. To do so, we propose *Netflix Rooms*. *Netflix Rooms* is a machine learning feature that will allow subscribers to stream movie premieres, television shows, and live events through a virtual movie theater experience. This will allow Netflix to:

1) Attract new subscribers by allowing current subscribers to create rooms and pay to add non-subscribers; and
2) Increase profits by charging to add non-subscribers to the personal rooms of subscribers, attracting new subscribers, and strengthening its retention on current subscribers.

As mentioned, Netflix is known for providing users with a highly personalized streaming experience (Rivera, 2019). *Netflix Rooms* will do the same through the use of exploratory analysis, followed by cluster analysis to create meaningful groups of users based on common interests and demographics including age, location, and occupation (Kaggle, 2017). We will also leverage user-based filtering with the use of Nearest Neighbor algorithm in the following ways (Kaggle, 2021):

1) To find users with similar rating patterns; and
2) Use these ratings to predict movie rooms to the user.

Once we complete the beta version of *Netflix Rooms*, we will run a cost-benefit analysis—with consideration to economic, environmental, social, and ethical impacts. This will ensure that our model is *worth* testing and eventually deploying.

Netflix Rooms' objective is to connect similar users together through this proposed virtual movie theater experience to provide an element of human-connection to streaming. The aim is to improve the well-being of Netflix users, while also providing the company with a clear return on investment and a competitive edge.

Description of Data

Netflix Movies and TV shows Dataset: Netflix has more than 8,000 movies and TV shows on its platform (Kaggle, 2021). This data set contains all the titles on Netflix, along with details about their actors, directors, audiences, release years, duration, and more (Kaggle, 2021).

MovieLens 100k Dataset: This dataset provides us with 100,000 ratings (on a scale of 1 to 5) for 1682 movies by 943 users and demographic data between the periods of September 1997 and April 1998. This dataset contains titles that are not exclusive to Netflix, so we intend to filter out the titles to only include Netflix titles. We will then combine these two datasets, forming the "movietitle" as their common link **(See Exhibit 7)**.

Data Overview

Netflix Movies and TV Shows

Quality: There are 8807 rows of data available, primarily complete with a few missing data in the fields like "director, cast, country, date_added, rating, duration" **(See Exhibit 5)**. The dataset contains mainly qualitative data, which requires encoding during the project for accurate analysis. We can easily handle this in the project's first stage, which focuses on data preparation and exploration.

Relevancy: Fields like country, rating, listed_in give information relevant to the project's scope to derive user preferences. Rating will tell us how the user receives each movie/show, and listed_in provides information about the genre of the show. The genre will help us understand what kind of show the user perceives by this dataset.

MovieLens 100K Dataset

Quality: We will retrieve that data using the website "movielens.umn.edu." The dataset is already partially cleaned, as user rows with less than twenty ratings and missing demographic data have been removed. However, in the first stage of our project, we will further analyze the dataset and eliminate unnecessary data.

Relevancy: We are mainly focusing on deriving valuable user insights using this dataset. Attributes related to users such as age, gender, location, occupation and the corresponding ratings enables us to map users into relevant groups. Using the two datasets, we will focus on clustering of users based on their respective similarities.

Data Integration

For our project, we will need to combine our two datasets to align movies and shows with ratings and user profiles. This process will include the integration itself, data cleaning, and feature engineering. We will present the data in a denormalized format that best represents the rating and user data needed.

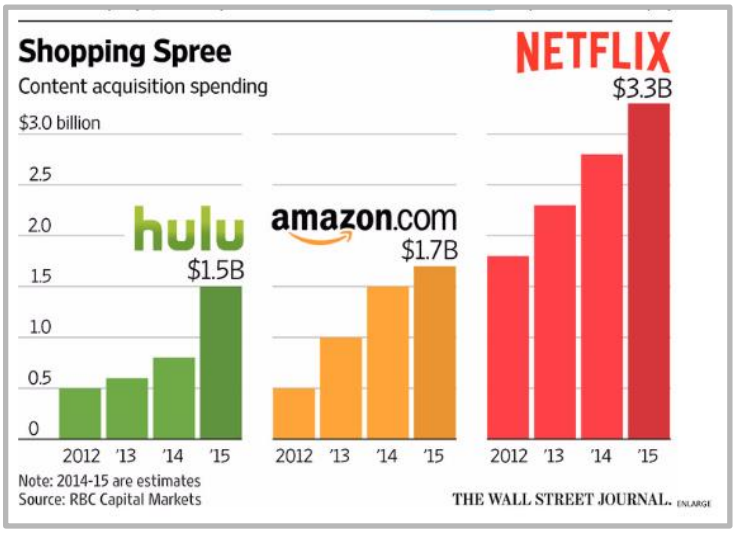
Exhibit 1: Netflix Inc. First Website

Exhibit 2: Netflix Inc. First Subscription Service

Exhibit 3: Netflix Inc. First Recommendation System

Sources (Exhibits 1-3): Netflix. (n.d.). About Netflix - Homepage

Exhibit 4: Netflix's Competitors



Source: Wall Street Journal

Exhibit 5: Count of Missing Data

Variable	Count
show_id	0
type	0
title	0
director	2634
cast	825
country	831
date_added	10
release_year	0
rating	4
duration	3
listed_in	0
description	0

Exhibit 6: Netflix TV Shows and Movies Dataset Variables

NETFLIX DATASET		
Feature	***Definition***	***Type of variable / method***
Show_id	Unique id for every Movie / TV show	• Numeric variable
Type	Identifier- Movie / TV show	• Categorical variable • Convert to dummy variable
Title	Title of Movie / TV show	• Categorical variable
Director	Director of Movie / TV show	• Categorical variable
Cast	Actor in Movie / TV show	• Categorical variable
Country	Country where Movie / TV Show produced	• Categorical variable • Convert to dummy variable
Date_added	Date it was added on Netflix	• Numeric variable
Release_year	Release year of Movie / Show	• Numeric variable
Rating	Age rating of Movie / Show	• Categorical variable • Convert to dummy variable
Duration	Duration of Movie / Show	• Numeric variable
Listed_in	Genre of Movie / Show	• Categorical variable • Convert to dummy variable
Description	Summary description	• Categorical variable

Exhibit 7: MovieLens 100k Dataset Variables

MOVIELENS 100K DATASET		
Datafile & Features	***Definition***	***Type of variable / method***
U.Data – userid / itemid / rating / timestamp	Contains information with ratings of 943 users on 1682 items, each user rated at least 20 items.	• Numeric variables
U.Info – movieid / movietitle/ releasedate / videoreleasedate / ImdB Url / Genre	Contains information about all the movies	• Numeric & Categorical variables
U.User – userid / age / gender / occupation / zip code	Demographic information about the users	• Numeric & Categorical variables

References

Business of Apps. (2022, January 11). *Video Streaming App Revenue and Usage Statistics (2022)*. Business of Apps. Retrieved February 26, 2022, from https://www.businessofapps.com/data/video-streaming-app-market/

Government of Canada. (2020, May 27). *The Daily — Canadians' mental health during the COVID-19 pandemic*. Statistique Canada. Retrieved February 26, 2022, from https://www150.statcan.gc.ca/n1/daily-quotidien/200527/dq200527b-eng.htm

Kaggle. (2017, January 5). *MovieLens 100K Dataset*. Kaggle. Retrieved February 26, 2022, from https://www.kaggle.com/prajitdatta/movielens-100k-dataset

Kaggle. (2021, September 26). *Netflix Movies and TV Shows*. Kaggle. Retrieved February 26, 2022, from https://www.kaggle.com/shivamb/netflix-shows

Netflix. (n.d.). About Netflix - Homepage. Retrieved February 25, 2022, from https://about.netflix.com/en

Rahman, K. T., & Arif, M. Z. U. (2021). Impacts of Binge-Watching on Netflix during the Covid-19 Pandemic. *South Asian Journal of Marketing*, *2*(1), 97-112.

Rivera, A. (2019, November 10). *Netflix's Mission Statement & Vision Statement: A Strategic Analysis*. Rancord Society. Retrieved February 25, 2022, from https://www.rancord.org/netflix-corporate-vision-statement-mission-statement-strategic-analysis