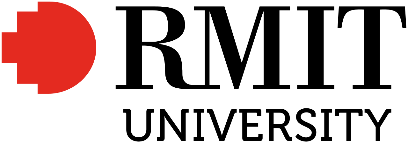
****

**Capstone Project** **Documentation**



**TOPIC: CREATE AND ANALYSE 4 TYPES OF VIDEO RECOMMENDATION ENGINES**

Victoria Loo

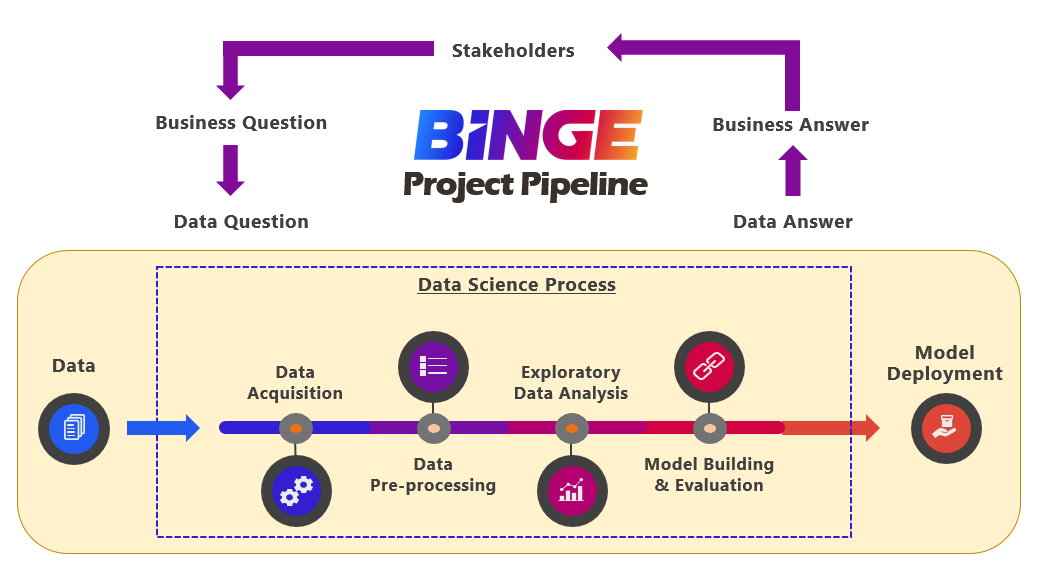
09 November 2021

# Introduction

Binge (owned by Foxtel Australia) is a video streaming subscription service that was recently launched in May 2020. In order to keep up with the growing on-demand culture, the need to build a robust video recommendation engine is paramount, especially for a video subscription-based company. A video recommendation engine is essentially a system that uses algorithms and data analysis techniques to recommend the most relevant videos to a particular customer.

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



# Problem Statement

Customer churn has a significant impact on the business revenue of a video subscription-based company. As the demand for instant gratification by customers increases every year, customer churn has become a growing concern in this highly competitive market. High customer churn rate can have a negative impact on the business revenue.

In 2016, Netflix has estimated that their recommendation engine is worth $1 billion because they believe it could lose $1 billion or more every year from subscribers cancelling their subscriptions.

By offering personalised recommendations that are tailored to customer’s unique preferences, customers are less likely to cancel their subscriptions when they like what they see. In this capstone project, I will introduce four different types of recommendation models that may improve the current recommendation engine in Binge.

# Industry

Binge is the latest streaming service that was recently launched in the entertainment and media industry of Australia. As a new Australian streaming service, the biggest challenge for the business is subscription loyalty or avoiding customers from cancelling their subscriptions.

# Stakeholders

The key stakeholders are:

1. CEO: Julian Ogrin
2. Executive Director: Alison Hurbert-Burns
3. Executive Management Committee

# Business Question

How can we offer personalised recommendations that are tailored to every customer unique preference and indirectly increase business revenue?

# Data Question

What machine learning models can we develop and use to offer personalised recommendations to customers?

# Data

The data was originally sourced from Kaggle. It is a dataset of a movie recommendation service, MovieLens. It contains 1,048,575 ratings of 27,278 movies, that were rated by 7,120 users between year 1995 and 2015.

# Data Science Process

## Data Analysis

Below are the steps taken to wrangle raw data:

1. Data Acquisition

* Three .csv files were available but only acquired two files (movies.csv and ratings.csv) because only those two files were relevant to project.

1. Data Cleaning & Transformation

* Merged .csv files via Microsoft Excel and saved as data\_final.csv
* Deleted all N/A values

1. Data Analysis

* All data analysis was done using Pivot table function on Microsoft Excel

## Highlights of Exploratory Data Analysis

**Most movies have at least 3.0 rating.**

**Thriller is the most popular genre rated by user.**

**Number of movies released increases every year.**

## Modelling

Four types of personalised recommendation models were created:

### Content Based Filtering

* Movie Genres are used as the main feature for this model.
* Content Based Recommender works by finding the genre similarities between movies.
* Example: User A likes Comedy/Romance movies, model will find more movies with similar genres.
* Due to low computational power, only 5000 movies (out of 27,278 movies) were selected to create this model.
* Natural Language Processing technique was first applied, which involves tokenising genres to create a movie-genre matrix (5000 rows × 19 columns).
* This was then followed by applying T-SNE technique to reduce dimension of data, plotting all movies onto two-dimensional coordinates and lastly calculate the Euclidean distance between two points.
* The Euclidean distance between two points will indicate the genre similarities between two movies.
* Training time:

### Collaborative Based Filtering

* UserID and Ratings are used as the main features for this model.
* Collaborative Based Recommender works by finding correlation between users who gave similar ratings on similar movies.
* Example: if User 1 watches movie A, B, C, and User 2 watches movie A, C, the model will recommend movie A and C to a similar User 3.
* Model was able to train on all 1048575 ratings given by all users.
* Singular Value Decomposition (SVD) method was used to create Collaborative Based Recommender. The final model will provide the top 10 highest rating predictions of a particular user on selected movies based on previous users ratings.
* Choice of Evaluation Metric is Root Mean Square Error (RMSE).
* Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data.
* Average RMSE for this SVD model is 0.8623.
* Training time:

### Hybrid Content-Collaborative Based Filtering

* Hybrid Recommender relies on the algorithms from both Content-Based and Collaborative Based Filtering.
* This was carried out by first creating content-based and collaborative-based models separately and then combining them.
* Firstly, the system uses Content-Based Filtering method to figure out the most similar movies to a selected movie. And in the next stage, it uses Collaborative Filtering to assign an estimated rating to those movies. And then, the model will filter out the top 10 movies and recommend them to the user.
* Training time:

### Deep Learning

* Movie\_ID, User\_ID and Ratings are used as main features for this model.
* Deep Neural Network with Keras library was applied to create the recommendation model.
* By utilising embedding and dense layers, the network is able to find better combinations between users and movies, which result in better rating predictions of movies.
* The final model will make top 10 movie recommendations with the highest rating predictions for a particular user.
* To reduce model training time, I only built a ONE layer deep neural network.
* Photo below is the overall architecture of the deep neural network recommendation model.

Movie\_ID **Input** Layer

User\_ID **Input** Layer

Movie **Embedding** Layer

User **Embedding** Layer

**Concatenate** reshaped embedding layers

**Add** one **Dense** layer

Model **Fit & Train**

Recommend Movies

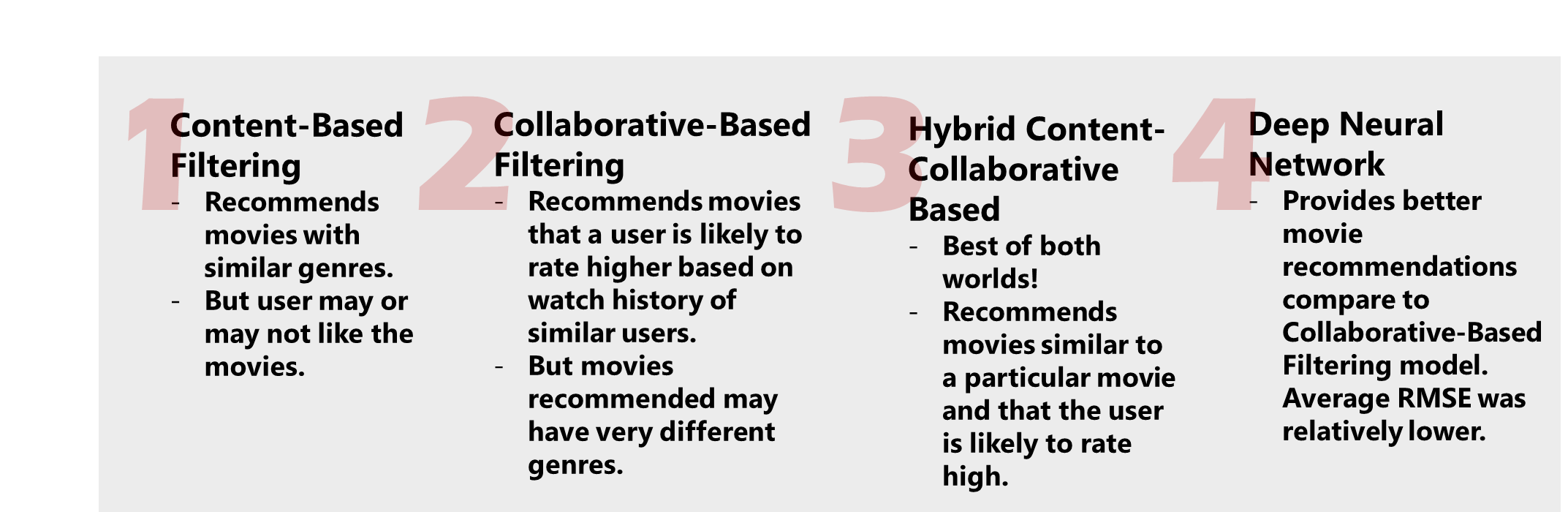
**Test** Model

**Reshape**

**Dropout** Regularisation

# Outcomes

All four models were successful at making video recommendations. The models that are most relevant to the business are Hybrid model and Deep Neural Network model. Both the models were able to provide relatively better recommendations to users.

Below is a summary comparison of all models:

# Implementation

Before implementing model in production, the stakeholders need to consider the following concern:

1. Cold start problem - This problem can arise when new users are added to the system. When User\_ID was absent during the model training process, we will not be able to recommend movies for that particular user. This problem can be eliminated by finding similar users based on the demographic or other features and recommend based on that for the new user. Therefore, it may be worth to consider including additional user features (eg. age, gender, location etc) into model training before implementing model.
2. Number of model types – depending on the budget and goals of the business, stakeholders should decide how many types of models they would like to implement into the business. As each of the four models created are different, it is possible to tailor and implement all four models into the business (provided that the business has the budget to implement all four of them).

# Data answer

Recommendation models were successfully created using machine learning and deep learning algorithms. The models were able to offer personalised recommendations to a particular user based on their taste and past watch history.

# Business answer

By applying machine learning and deep learning algorithms, we are able to confidently tailor recommendations according to customer’s unique preference.

# Response to stakeholders

With the use of machine learning and deep learning algorithm, it is possible to create a robust personalised recommendation engine that is paramount for a successful video streaming subscription business. High subscription loyalty rate and low customer churn rate will improve business revenue.

Depending on the current need of the business, the above models developed can be further enhanced by optimizing parameters or adding more data/features.

# End-to-end solution

Once RMSE of predictive models have reached desired value, models should be tested on a set of users to gain real-time feedbacks before deploying all models across all streaming platforms.

# References

* The list of Python libraries used:
  + Pandas
  + Numpy
  + Matplotlib
  + Wordcloud
  + Scikit Learn
  + Plotly
  + Ipywidgets
  + Surprise
  + Keras

- Data Source:

<https://www.kaggle.com/grouplens/movielens-20m-dataset>

* Articles and Study Resources:
* <https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe>
* <https://towardsdatascience.com/how-you-can-build-simple-recommender-systems-with-surprise-b0d32a8e4802>
* <https://www.johnwittenauer.net/deep-learning-with-keras-recommender-systems/>
* <https://app.datacamp.com/learn/courses/building-recommendation-engines-in-python>
* <https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-recommendation-engine-python/#h2_13>