Subscription Fee Reference for Streaming Platforms Launching in New Markets

Group 11

***Li Ruirui** (A0231941L) **Wu Jitong** (A0218855U) **Xu Han** (A0218930E)*

***Zhou Lingyi** (A0218871X) **Zhou Yin** (A0218923B)*

Abstract

Streaming platforms are gradually taking over the traditional TV industry those days. Netflix, as the 1st streaming video platform, reaching 165 millions paying customers globally and launching in most of countries. The problem statement has been focused on other streaming service platforms who has only launched in a few countries, what the subscription fee should be set when entering the new market. By searching multiple features and pre-processing datasets, features could be better implemented into models (Random Forest, Light GBM, XGBoost, Neural Network). Depends on overall MAE, RMSE, MAPE value, the best model has been decided to predict the subscription reference fee. Business insights and future improvement also been discussed.

1. Introduction

With the evolution of communication and network technology, there is an increasing number of people giving up traditional television channel subscription and turning to watching movies and drama series through online streaming platforms, such as Netflix, Amazon Prime Video, Disney+ and so on. These companies have developed the convenient tools for people to access the online entertainment at any time with any devices. Users can navigate the resources in the organized manners and resources are plenty, including movies, documentaries, and television series. To enrich the content, some of them even invest in the programs hence reserving the exclusive authority and copyright. For a paid streaming platform, it only charges a flat subscription fee. For only a flat monthly fee you paid, you could consume unlimited shows at any time on whatever you prefer [1].

Netflix is the very first one that started the streaming platform in 2007 and now already become a mature platform and owned a huge user community that has about 167 million paying customers globally [2]. As one of the most successful online streaming companies, the strategy to invest in resources and price for different services in different areas is critical to improve the service quality and attract more users. The project team will utilize the subscription data from Netflix and macro-economy data for various countries to explore factors that potentially impact Netflix's pricing strategy. The pricing model developed will be able to predict the optimal price that should be set with different resource pools for different service levels in different regions. This could act as subscription fee guideline for other streaming platforms who is planning to launch their service in countries which Netflix has already operated in so that the price could be competitive as Netflix.

1.1 Motivation

There are some interesting facts that motivated the team to research further on this topic:

- Nowadays, streaming platforms are gradually taking over the traditional TV industry. Within the project team, all the team members have subscribed to at least one online streaming service and none of them subscribe to the traditional TV service.

- Taking a glance at Netflix statistics [2] at below figures, it is obvious that Netflix has increased its revenues significantly, especially in 2020 where Covid 19 hit human society and people had to stay longer time at home. However, the average revenue earned per user has remained steady since 2016 which implies a huge increase of user base.

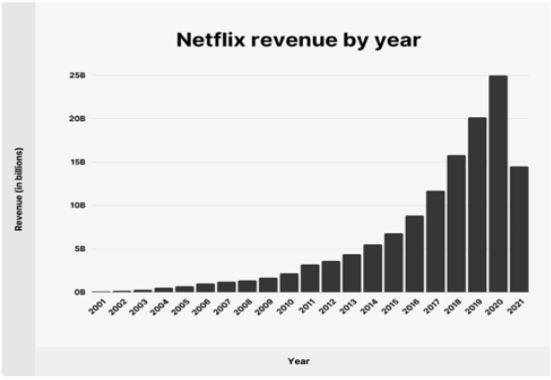


Figure1.1.1 Netflix Revenue by Year

Year	Average revenue per user
2016	$8.61
2017	$9.43
2018	$10.31
2019	$10.82
2020	$10.91
2021 (Q1 & Q2)	$10.86

Figure1.1.2 Netflix Average Revenue per User

- The profitability of the online streaming market has attracted followers to participate in the race. However, it cannot be denied that Netflix has set a benchmark and industry standard for its competitors to follow. Netflix has already grown to be a global platform which has been launched in 65 countries in the world which the rest still have rooms to pick up in terms of user coverage.

Based on those, problem statement of the project has been come up. For other streaming service platforms who has only launched in a few countries, what the subscription fee should be set when entering the new market where Netflix has launched given the resources they owned and how to ensure the competitiveness of the pricing strategy. One of the practical examples is that, if Disney+ would like to expand the service network which currently is only available in 36 countries and launch in Turkey [3], what the target subscription price should be.

1.2 Assumption

It is complicated to define whether a subscription fee is set successfully as there are other factors may affect business performance including customer loyalty, viewing experience, movie/drama series popularity, etc. These factors are unpredictable and difficult to measure based on the given dataset hence will not be considered in modelling. Project team assumes all the streaming platforms/services are at the same competency level and all the resources are identical. Hence the subscription fee reference will be based on the features listed down below only.

1.3 Objectives

Areas of focus in this project are:

- Analyze Netflix subscription fee dataset for different countries.
- Explore potential correlations between subscription fee and macro-economic and statistical figures (e.g., Gini Index, population, GDP) over different countries and determine whether or how those measures can be utilized to improve business profit.
- Discover Netflix pricing strategies for different users across different countries and its practicality for streaming platforms to replicate when entering a new market.
- Provide the subscription fee reference guide for streaming platforms or services which planned to enter a new market (in countries never been launched this platform before) based on most successful on-demand Video streaming platform - Netflix's dataset.

2. Data Pre-processing & Feature Generation

Below is the detail table of the data after pre-processing, which will be used in the following model training.

Feature	Description
Country Code	Short alphabetic code to represent countries, ISO 3166-1 alpha-2 standard is used.
Country	Country name
Total Library Size	Total Movie and TV shows available on Netflix in this country
No. of TV Shows	No. of TV shows available on Netflix
No. of Movies	No. of movies available on Netflix
Cost per Month (Standard)	Standard subscription fee per month (USD)
Movie_aveRating	IMDb weighted average of all the individual user ratings for Movies in Netflix
TV_aveRating	IMDb weighted average of all the individual user ratings for TV shows in Netflix
GDP	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.
Population	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
Gini_Index	Also named as Gini coefficient, is a measure of statistical dispersion intended to represent the income inequality or wealth inequality within a nation.
Happiness_Share	Data from World Happiness Report evaluation question. Represent share of people (in percentage) who say they are 'very happy' or 'rather happy' from 5

	categories they were asked for each country.
Life_Satisfaction	Data from self-reported life satisfaction. Average numbers are calculated for different countries. The best possible life for them being a 10, and the worst possible life being a 0.
Broadband_subscriptions	Broadband subscriptions refer to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s.
Cellular_subscriptions	Mobile phone subscriptions, measured as the number per 100 people.
averageRating	IMDB rating average of top 10 movies and TV Shows in each country.
numVotes	IMDB votes number average of top 10 movies and TV Shows in each country.

Table2.1 Features used for Model Training

There are 4 data sources of the dataset here, and the pre-processing detail for each data source will be introduced in each part later.

2.1 Netflix subscription fee and library size in different countries

The data source is Comparitech [4], and it consists of following features.

Feature	Description
Country Code	Short alphabetic code to represent countries, ISO 3166-1 alpha-2 standard is used.
Country	Country name
Total Library Size	Total Movie and TV shows available on Netflix in this country
No. of TV Shows	No. of TV shows available on Netflix
No. of Movies	No. of movies available on Netflix
Cost per Month (Basic)	Basic subscription fee per month (USD)
Cost per Month (Standard)	Standard subscription fee per month (USD)
Cost per Month (Premium)	Premium subscription fee per month (USD)

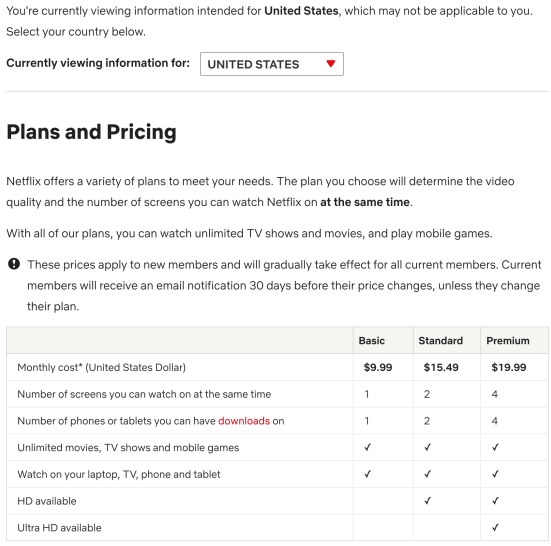
Table2.1.1 Netflix features used for Model Training

Below is a sample look of this data.

Flag	Country	Total Library Size	Library Size Change (Aug 18 to Apr 22)	No. of TV Shows	No. of Movies	Cost Per Month - Basic ($)	Cost Per Month - Standard ($)	Cost Per Month - Premium ($)
	Algeria	4,8...		1,830	3,...	7.99	9.99	11.99
	Argentina	5,049		1,900	3,1...	3.38	5.70	8.37
	Australia	5,880		1,969	3,911	8.18	12.65	17.12
	Austria	5,794		1,916	3,878	8.69	14.13	19.57
	Bahrain	4,9...		1,850	3,0...	7.99	10.49	15.79
	Belgium	5,095		1,6...	3,4...	9.78	14.67	19.57
	Bermuda	6,273		2,103	4,170	8.99	12.99	15.99
	Bolivia	5,0...		1,895	3,1...	7.99	10.99	13.99
	Brazil	5,0...		1,865	3,1...	5.51	8.50	11.90
	Bulgaria	7,162		2,024	5,138	8.69	10.87	13.04
	Canada	6,299		2,025	4,274	7.95	13.12	16.70
	Chile	5,0...		1,897	3,1...	7.31	10.23	13.16
	Colombia	5,044		1,895	3,1...	4.48	7.13	10.31
	Costa Rica	5,0...		1,895	3,1...	8.99	12.99	15.99
	Croatia	2,9...		718	2,2...	8.69	10.87	13.04
	Czechia	4,7...		1,...	3,5...	8.86	11.53	14.20
	Denmark	4,8...		1,5...	3,3...	11.55	16.67	21.79
	Ecuador	5,0...		1,896	3,1...	7.99	10.99	13.99
	Egypt	4,9...		1,849	3,0...	6.54	9.00	10.90
	Estonia	6,932		2,015	4,917	8.69	10.87	13.04

Figure2.1.1 Sample of Netflix Datasets

In this data, the prices included in are the based prices advertised by Netflix as the following. The prices here do not include the taxes and other charges.


You're currently viewing information intended for **United States**, which may not be applicable to you. Select your country below.

Currently viewing information for: UNITED STATES

Plans and Pricing

Netflix offers a variety of plans to meet your needs. The plan you choose will determine the video quality and the number of screens you can watch Netflix on **at the same time**.

With all of our plans, you can watch unlimited TV shows and movies, and play mobile games.

These prices apply to new members and will gradually take effect for all current members. Current members will receive an email notification 30 days before their price changes, unless they change their plan.

	Basic	Standard	Premium
Monthly cost* (United States Dollar)	$9.99	$15.49	$19.99
Number of screens you can watch on at the same time	1	2	4
Number of phones or tablets you can have downloads on	1	2	4
Unlimited movies, TV shows and mobile games	✓	✓	✓
Watch on your laptop, TV, phone and tablet	✓	✓	✓
HD available		✓	✓
Ultra HD available			✓

Figure2.1.2 Plans & Pricing

And after comparing the ratio between Basic, Standard and Premium Cost per Month as below graph

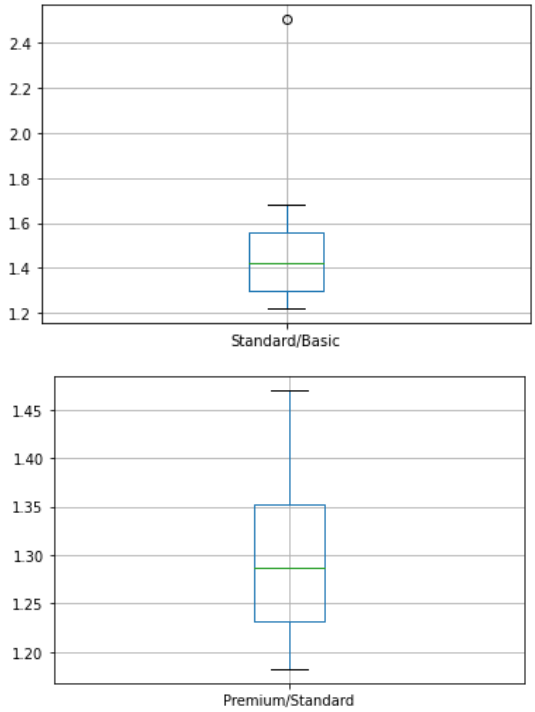


Figure2.1.3 Standard/Basic & Premium/Standard Price Comparison

The ratio between Standard and Basic is around 1.4 +- 0.1, and it is 1.3 +- 0.05 for the majority, so only the Cost per Month (Standard) will be predicted in the following model training.
The outlier in Standard versus Basic Cost per Month is India, whose Cost per Month – Basic is $2.64, however, Standard is $6.61.

2.2 IMDb rating data for Movies and TV shows

There 2 subsets of IMDb[5] data used,
title.ratings.tsv.gz, which contains the IMDb rating and votes information for titles, consists of the following features.

Feature	Description
tconst (string)	alphanumeric unique identifier of the title
averageRating	weighted average of all the individual user ratings
numVotes	number of votes the title has received

Table2.2.1 Features used for Model Training

title.episode.tsv.gz, which contains the tv episode information, it is used to differentiate the rating data of Movie and TV Shows. It consists of the following features.

Feature	Description
tconst (string)	alphanumeric unique identifier of the title
parentTconst	season number the episode belongs to
seasonNumber	episode number of the tconst in the TV series
episodeNumber	episode number of the tconst in the TV series

Table2.2.2 Features used for Model Training

However, there is no Netflix library data in part 1 - Netflix subscription fee in different countries data, so the Movie_aveRating and TV_aveRating are calculated in following method, random selection applied here according to each country`s Movie and TV Show library size, then the average rating score was calculated.

2.3 Country-level Metrics

Feature	Description
GDP	GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.
Population	Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.
Gini_Index	Also named as Gini coefficient, is a measure of statistical dispersion intended to represent the income inequality or wealth inequality within a nation.
Happiness_Share	Data from World Happiness Report evaluation question. Represent share of people (in percentage) who say they are 'very happy' or 'rather happy' from 5 categories they were asked for each country.
Life_Satisfaction	Data from self-reported life satisfaction. Average numbers are calculated for different countries. The best possible life for them being a 10, and the worst possible life being a 0.
Broadband_subscriptions	Broadband subscriptions refer to fixed subscriptions to high-speed access to the public Internet (a TCP/IP connection), at downstream speeds equal to, or greater than, 256 kbit/s.
Cellular_subscriptions	Mobile phone subscriptions, measured as the number per 100 people.

Table2.3.1 Features used for Model Training

Data are extracted from World Bank[6] and Our World in Data[7]. Year 2020 data is selected as the latest representation. There are some countries that do not have 2020 data, therefore we choose the latest available data (before 2020). We found there are still missing values in the dataset. By leveraging other similar information (e.g. Country ranking by happiness index[8] and broadband subscription information[9]), a portion of missing values are filled by other countries' metrics that have a similar ranking. For remaining missing values that can't be found, with the understanding of feature values distribution via boxplot in Figure2.3.1, data especially GDP and population are quite skewed, in the end, the median value is chosen to fill in missing data.

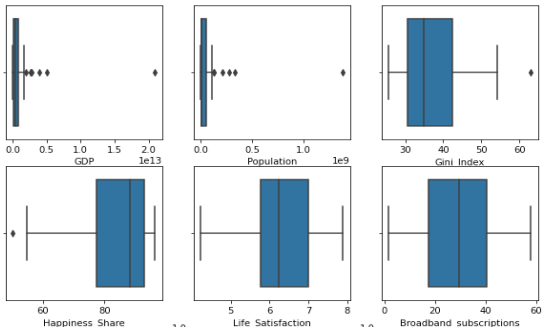


Figure2.3.1 Feature Values Distribution via Boxplot

2.4 Weekly top 10 watches on Netflix

The following two features were built to associate users' preferences on content offered by Netflix

using top 10 movies and TV Shows data from Netflix's Top 10 list[10].

Feature	Description
averageRating	IMDB rating average of top 10 movies and TV Shows in each country.
numVotes	IMDB votes number average of top 10 movies and TV Shows in each country.

Table2.4.1 Netflix used for Model Training

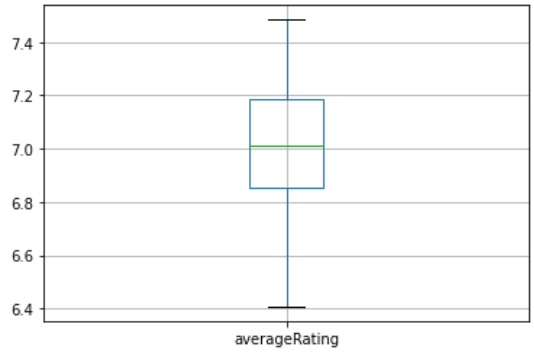


Figure2.4.1 Average Rating over Country Ranges

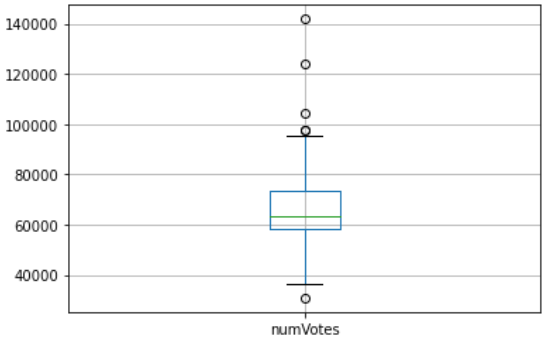


Figure2.4.2 Number of Votes Lookup

The weekly top 10 dataset is available on Kaggle[11]. We matched the title of contents from top-10 dataset with the title in IMDB dataset for rating and number of votes lookup. The average rating over countries ranges from 6.4 to 7.48 with a standard deviation of 0.24 in Figure 2.4.1. There are also some areas that can be improved in terms of feature engineering for these two features: 1) Only the top-10 movie data in week 2021-11-14 was used to compute average ratings and number of votes, it could be more accurate and robust if all data over a long-time span can be used for the feature extraction. 2) Some title of movie or TV shows is associated with multiple instances in the IMDB dataset. To simplify the feature engineering process, the latest released movie or tv show is eventually matched to the title from top-10 dataset to obtain ratings and the number of votes.

3. Models & Feature Importance

Using following machine learning and deep learning techniques on Netflix subscription fee and external data at region level, to build a prediction model to achieve relatively high accuracy.

3.1 Models

3.1.1 Random Forest

Random Forest is the commonly used model for resolving regression problems. It takes average numbers based on the decision trees built on different samples. Applying bootstrap aggregation, bagging ensemble method constructs individual decision tree based on the sample selected and each will generate an output. The final output is based on the average figures of all the decision trees. Random forest model helps to avoid the curse of dimensionality as each tress does not consider all the features. At the same time, as each tree is created separately and run in parallel, the computing efficiency could be optimized.

Hyperparameters like n_estimators, max-features and mini_sample_leaf will help to increase the predictive accuracy while n_jobs and oob_score will increase the speed.

In our project, prediction result of three service categories (Basis, Standard, Premium) are populated first, where prediction for Premium subscription fee generated the lowest MAPE of 16.57%.

```
GridSearchCV(cv=5, estimator=RandomForestRegressor(random_state=0),
             param_grid={'bootstrap': [True, False], 'max_depth': [5, 10, 100],
                         'max_features': ['auto', 'sqrt'],
                         'min_samples_leaf': [1, 2, 4],
                         'min_samples_split': [2, 5, 10],
                         'n_estimators': [10, 50, 100]})
```

```
{'bootstrap': False,
 'max_depth': 10,
 'max_features': 'sqrt',
 'min_samples_leaf': 2,
 'min_samples_split': 5,
 'n_estimators': 50}
```

Figure3.1.1.1 Random Forest GridSearch & Results

By applying the grid search to adjust the key hyperparameters, below are the best set with variance among six hyperparameters. The result has improved significantly as compared to the default setting which is shown in the table below.

Parameters	Original	Fine-tuned	Improvement
MAE	1.75	1.35	22.7%
MAPE	18.02%	14.40%	20.1%
RMSE	2.18	1.64	25.0%

Table3.1.1.1 RF Parameters Improvements

Subscription Type	MAPE
Basis	23.90%
Standard	18.02%
Premium	16.57%

Table3.1.1.2 RF Subscription Type and MAPE

3.1.2 LightGBM

Light Gradiant Boosting Machine is also a decision tree-based algorithm with distributed high-performance framework. The advantage of the LightGBM is the fast training data hence gains the popularity for the large dataset problems. It also supports both parallel learning and GPU learning which makes the algorithm more attractive. Though the dataset for this project is relatively small, it is worthy to develop the prediction model based on LightGBM. For the LGBM Regressor, larger num_leaves leads to better accuracy though it increases the probability for over-fitting. max_depth and lambda will help to deal with over-fitting.

```
GridSearchCV(cv=5, estimator=LGBMRegressor(random_state=0),
             param_grid={'max_depth': [1, 5, 10, 20],
                         'min_child_samples': [5, 10, 20],
                         'n_estimators': [10, 50, 100, 200],
                         'num_leaves': [5, 10, 20, 30],
                         'reg_alpha': [1, 50, 100],
                         'reg_lambda': [1, 50, 100]})
```

```
{'max_depth': 5,
 'min_child_samples': 10,
 'n_estimators': 100,
 'num_leaves': 5,
 'reg_alpha': 1,
 'reg_lambda': 50}
```

Figure3.1.2.1 Light GBM GridSearch & Results

Similarly, after obtaining the prediction on test data for three different categories. The grid search is applied to find out the best hyperparameters for the LightGBM model. Unlike Random Forest, the best hyperparameters do not give much better results.

Parameters	Original	Fine-tuned	Improvement
MAE	1.47	1.47	-0.04%
MAPE	15.45%	15.03%	2.67%
RMSE	1.82	1.69	7.38%

Table3.1.2.1 LightGBM Parameters Improvements

Subscription Type	MAPE
Basis	17.42%
Standard	15.45%
Premium	14.44%

Table3.1.2.2 RF Subscription Type and MAPE

3.1.3 XGBoost

XGBoost, a popular decision-tree-based ensemble method, which boosts the performance of weak learners to attain the performance of stronger learners is applied to our dataset. We use Grid Search for hyperparameter tuning and mean absolute percentage error(MAPE) as the evaluation metric for best parameter selection. XGBoost has a lot of hyperparameters that can be adjusted which may easily lead to overfitting. Furthermore, the dataset size is quite small so when a complex algorithm like XGBoost is applied, the risk of overfitting increases drastically. After tuning, the model training MAPE is around 14.8% and the test MAPE is around 17.7%. Figure 3.1.3 visualizes the actual and prediction result and the prediction error is significant.

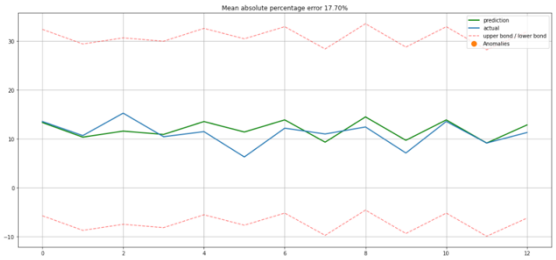


Figure3.1.3.1 XGBoost Actual vs Prediction Result

3.1.4 Neural Networks

A neural network is a computational system that can create predictions based on existing data. It is used Supervised learning in this case and the chosen features that form the input for this neural network are followings - 'Total Library Size', 'No. of TV Shows', 'No. of Movies', 'Movie_aveRating', 'TV_aveRating', 'GDP', 'Population', 'Gini_Index', 'Happiness_Share', 'Life_Satisfaction', 'Broadband_subscriptions', 'Cellular_subscriptions', 'averageRating', and 'numVotes'. 'Cost Per Month - Standard ($)' has been set as the y variable.

Coming to the neural network configuration, the most important considerations when training a neural network is choosing the number of neurons to include in the input and hidden layers. Given that the output layer is the result layer which will be the standard cost, this layer has 1 neuron present by default. Based on the dataset in this project, the number of neurons in each layer is configured as below:

- **Input layer:** Number of features in the training set + 1. In this case, as there were 13 features in the training set to begin with, 14 input neurons are defined accordingly.
- **Hidden layer:** The number of neurons in the hidden layer = Training Data Samples/Factor * (Input Neurons + Output Neurons) A factor of 1 is set in this case, the purpose of the factor being to prevent overfitting. With 14 neurons in the input layer, 1 neuron in the output layer and 52 observations in the training set, the hidden layer is assigned 4 neurons.
- **Output layer:** As this is the result layer, the output layer takes a value of 1 by default.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 13)                182
_________________________________________________________________
dense_1 (Dense)              (None, 4)                 56
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 5
=================================================================
Total params: 243
Trainable params: 243
Non-trainable params: 0
```

Figure3.1.4.1 Neural Network Model Configuration

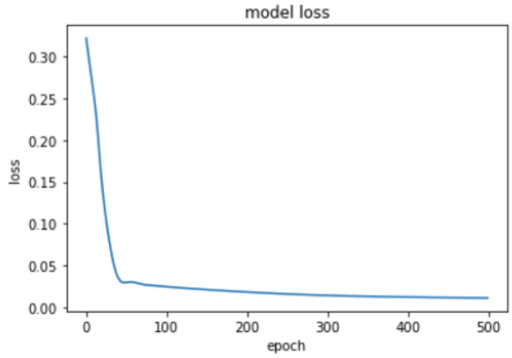


Figure3.1.4.2 Model Loss with Increasing Epoch

With the expectation that the loss will reduce while increasing each epoch to fit the model, which means the model is predicting standard cost more accurately when the model has been continued trained. The training loss has decreased as the number of epochs is increased, meaning that model gains a high degree of accuracy as number of forward and backward passes is increased. Eventually, by passing the test data to the same model to predict the standard cost and the results in MAPE 21.14% of testing data.

3.2 Feature importance

Figure 3.2.1 and Figure 3.2.2 show Random Forest and lightGBM feature importance ranking. Broadband_subscription, Happiness_Share, Life_Satisfaction are top three features that affect the model prediction. Netflix library features like No_of_Movies and Total_Library_Size seem to be less important.

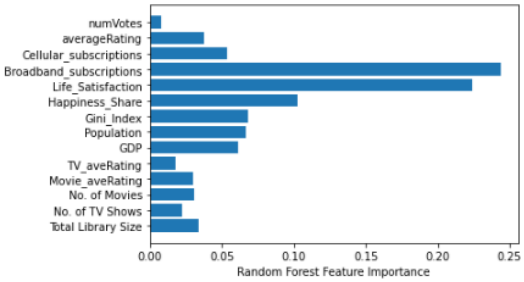


Figure3.2.1 Random Forest Feature Importance

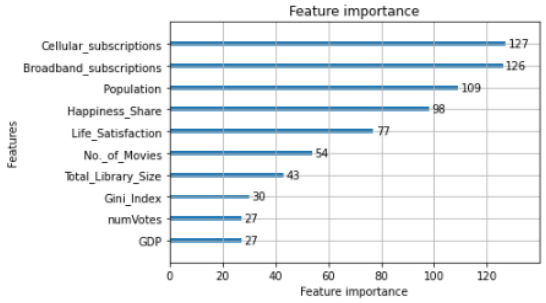


Figure3.2.2 LightGBM Feature Importance

A quite different feature importance ranking is obtained from XGBoost model as shown in Figure 3.2.3. The top 5 features influencing the model are averageRating, Happiness_Share, Gini_Index, GDP and No. of Movies for each country.

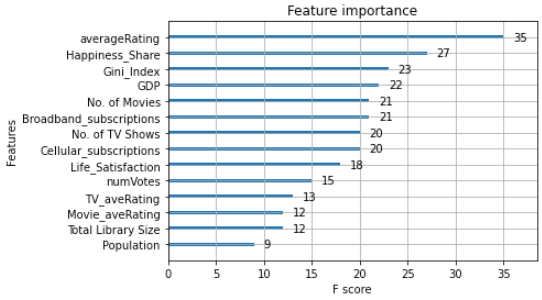


Figure3.2.3 XGBoost Feature Importance

4. Conclusion and Recommendations

4.1 Model Evaluation

	RandomForest	LGBM	XGBoost	Neural Network
RMSE	1.64	1.69	2.2	2.46
MAE	1.35	1.47	1.67	2.07
MAPE	14.40%	15.03%	17.7%	21.14%

Table 4.1.1 Model Performance Comparison

RandomForest has the lowest RMSE (1.64), MAE (1.35), and MAPE (14.40%) compared to other models. With the increase of model complexity, all three metrics of error are getting larger. For instance, the error of neural network is 50% higher than RandomForest in terms of RMSE. For complex model such as neural network, data size is crucial part of the training process due to substantial number of parameters needs to be tuned in hidden layer. Therefore, a complex model may not be an ideal option for a prediction problem if the data size is as small as the dataset used in this work. Most importantly, one of the best practices for machine learning modeling is to always start with a simple model followed by adding more complexities in machine learning algorithms we choose.

4.2 Business Insights

Netflix subscription fee is highly correlated with broadband which reveals users in countries with higher Internet popularity must pay more for monthly subscription of streaming service. Secondly,

customers are not getting a larger size of movie library if they pay more as the correlation between subscription fee and movie library size is –0.17 for standard subscription. This interestingly defies the common sense of users getting what they paid.

Also, subscription is cheaper in countries with a higher Gini index, which can be interpreted as subscription fee is to be set at a lower range so that the content provider is able to get a larger number of paid subscribers to reduce the marginal cost of upfront investment of entering this market where income and wealth inequality is significant. In another word, this is to ensure that the subscription service has a minimum number of users to cover the upfront investment cost to ensure the service is profitable.

4.3 Business Applications

With best performance model trained by Random Forest, quantitative evaluations could be analysed which will help on management decision to expand the service network. Two potential business scenarios could be applied using the established model and pricing strategy is only advised for standard subscription which aligned with assumptions made in the model evaluation part.

Firstly, when entering a new market, what is the recommended price to be set given fixed macro-data for a specific country and pre-defined library size. For the same country, how will the price change in response to the library size. It is critical to understand if increasing resource available in the platform will help on the pricing as each resource on-boarding to the platform occurs authorization expense. In the elementary exploration, we selected China, Egypt, and Vietnam as examples to test out considering the economic scale and representativeness. For each country, we set five different library sizes for prediction based on the distribution of the train dataset. It is interesting to find out that the higher library size does not always mean the higher pricing based on the model trained. For example, both Egypt and Vietnam return the highest pricing under (5500, 3600, 1900) while highest price for China market is under library size of (5000, 3200, 1800).

Library Size	Total Library Size	No. of TV Shows	No. of Movies
Library Size 1	4500	2900	1600
Library Size 2	5000	3200	1800
Library Size 3	5500	3600	1900
Library Size 4	6000	4000	2000
Library Size 5	6500	4500	2000

Table 4.3.1 Library Size Information

Country	Total Library Size	No. of TV Shows	No. of Movies	Cost Per Month - Standard ($)
China	4500	2900	1600	11.71
	5000	**3200**	**1800**	**11.80**
	5500	3600	1900	11.74
	6000	4000	2000	11.76
	6500	4500	2000	11.68
Egypt	4500	2900	1600	8.15
	5000	3200	1800	8.02
	5500	**3600**	**1900**	**8.17**
	6000	4000	2000	8.06
	6500	4500	2000	8.09
Vietnam	4500	2900	1600	10.31
	5000	3200	1800	10.20
	5500	**3600**	**1900**	**10.47**
	6000	4000	2000	10.32
	6500	4500	2000	10.32

Table 4.3.2 Library Size v.s Standard Cost

Secondly, for the existing markets, is there any opportunity to improve the revenue by adjusting the library size. The test countries chosen are North Europe countries, namely Denmark, Finland, Norway, and Sweden. We noticed that for these four countries, the average subscription fees are higher among others as they are well-developed countries with high-income level, but library size for counties are relatively small. Hence, it is worthy to examine if increasing library size will help on subscription fee. Applying the same sets of library sizes for these four countries, it is aspiring to find out that for Denmark, Finland, and Norway, slightly increasing the library size to (5000, 3200, 1800) will be able to generate the highest pricing. However, for Sweden, the original library size is still the best.

	Total Library Size	No. of TV Shows	No. of Movies	Cost Per Month - Standard ($)
Denmark	4558	2978	1580	15.04
	5000	**3200**	**1800**	**15.39**
	5500	3600	1900	15.17
	6000	4000	2000	14.72
	6500	4500	2000	14.72
Finland	4045	2638	1407	13.54
	5000	**3200**	**1800**	**14.70**
	5500	3600	1900	14.46
	6000	4000	2000	13.93
	6500	4500	2000	13.93
Norway	4528	2955	1573	12.17
	5000	**3200**	**1800**	**13.73**
	5500	3600	1900	13.56
	6000	4000	2000	13.39
	6500	4500	2000	13.41
Sweden	**4361**	**2973**	**1388**	**14.20**
	5000	3200	1800	14.03
	5500	3600	1900	13.78
	6000	4000	2000	13.38
	6500	4500	2000	13.35

Table 4.3.3 Library Size v.s Standard Cost

From above explorations, we could conclude that it is not necessary to put all the resources on expanding the library resource as the subscription fee will peak

at certain range. The conclusion seems to be a good reflection of current situation Netflix is facing. While investing plenty of money and resource on the original series, the new users did not grow as per expectation which resulted in a bad financial performance for Q1 2022 and the stock market has responded dramatically. In order to sustain the business growth, Netflix could spend more capital to acquire new users and enhance the market penetration. However, here the only variable considered is the library size and each movie or TV show is treated identically. The size of the subscribers is not included as well due to confidentiality hence the exact profit analysis is not able to be calculated.

4.4 Limitations and future improvements

- Netflix is available in over 190 countries as of now. However, the subscription dataset only reveals price insights of approximately one-third of countries. A larger dataset with more countries would be ideal for machine learning problems in terms of robustness of model and accuracy of predicted results.
- The title of top movies and TV shows was fuzzily matched with IMDB rating in feature engineering. This could be calibrated manually to reduce the mismatched ratings so that the average rating and number of votes would be more accurate.

5.Reference

[1] Investopedia. 2022. *How Netflix Is Changing the TV Industry*. [online] Available at: <https://www.investopedia.com/articles/investing/060815/how-netflix-changing-tv-industry.asp>

[2] Backlinko. 2022. *Netflix Subscriber and Growth Statistics: How Many People Watch Netflix in 2022?*. [online] Available at: <https://backlinko.com/netflix-users>

[3] whattowatch.com. 2022. *Disney Plus price: what it costs in all the countries where it's available*. [online] Available at: <https://www.whattowatch.com/watching-guides/disney-plus-price-what-it-costs-in-all-the-countries-in-which-its-available>

[4]Comparitech. 2022. *Which countries pay the most and least for Netflix? - Comparitech*. [online] Available at: <https://www.comparitech.com/blog/vpn-privacy/countries-netflix-cost/>

[5] IMDb. 2022. - *IMDb*. [online] Available at: <https://www.imdb.com/interfaces/>

[6] Data.worldbank.org. 2022. *World Bank Open Data | Data*. [online] Available at: <https://data.worldbank.org/>

[7]Our World in Data. 2022. *Our World in Data*. [online] Available at: <https://ourworldindata.org/>

[8] Goshwami, S., 2022. *World Happiness Index 2022 Country wise Rank & Report*. [online] DMER Haryana: Recruitment, News, Admit card, result. Available at: <https://dmerharyana.org/world-happiness-index/> .

[9] En.wikipedia.org. 2022. *List of countries by number of broadband Internet subscriptions - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/List_of_countries_by_number_of_broadband_Internet_subscriptions>

[10] Netflix Top 10. *Global Top 10, Weekly Top 10 lists of the most-watched TV and films*. Available at: < https://top10.netflix.com/>.

[11]Netflix Top 10 Weekly Dataset Available at: <https://www.kaggle.com/datasets/mikitkanakia/netflix-top-10-weekly-dataset>

[12]Brownlee, J., 2022. *How to Develop a Light Gradient Boosted Machine (LightGBM) Ensemble*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/light-gradient-boosted-machine-lightgbm-ensemble/>

[13]Lightgbm.readthedocs.io. 2022. *Parameters Tuning — LightGBM 3.3.2.99 documentation*. [online] Available at: <https://lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html>

6. Code Link

https://github.com/yzhou024/bt5153_project11

Appendix A

Correlation Matrix