WQD7005 Data Mining

Master of Data Science

Faculty of Computer Science & Information Technology

University of Malaya

Individual Assignment

Application of Data Mining in Movie Ratings Prediction

Tan Yin Yen

WQD180108

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Analysis Goal

A movie streaming company (Netflix) seeks to maximize customer's retention by recommending highly rated movies with DVD or streaming options available to their users. A movie can be predicted as good or bad based on the movie information, box office data and sentiment score of user reviews on the movie.

By predicting whether a movie is good or bad based on its reviews and box office achievement, the movie streaming company can filter out latest movies with DVD or streaming options available that are highly rated and recommend them to its users. Customers who are satisfied with the movie recommendations are more likely to subscribe to the movie streaming service in the next month.

1.2 Objectives

The main objective of this assignment is to come out with a model to predict whether a movie is good or bad based on its box office performance and user reviews. Besides that, the project also aims to recommend users the top and latest movies with DVD or streaming options available based on the prediction.

CHAPTER 2: METHODOLOGY

2.1 Tools and Programming Languages

This project is implemented using several tools and programming languages with their purposes shown as follows:

Tools and Programming Languages	Purpose
Python	Data scraping and data cleansing
Apache Hive	Data storage
SAS Enterprise Miner	Data cleansing, visualization and modelling

Table 1. Tools and Programming Languages Used

2.2 Data Acquisition

In this project, both the structured and unstructured data are used to predict whether a movie is good or bad. Since this project aims to recommend the latest movies to the users, only the data of the latest 5000 movies with Digital Versatile Disc (DVD) or streaming options available is scraped.

2.2.1 Structured Data

For structured data, movie information and box office data are scraped from *rottentomatoes.com* and *boxofficemojo.com* respectively. The movie information data contains features namely, Director, Genre, In Theater Date, On Streaming Date, Rating, Runtime, Studio, Writer, Audience Score, Critics Consensus, Title, Tomatometer, Tomatometer Count, Uniform Resource Locator (URL) and User Rating Count. The movie box office data contains features namely, In Release (total days), Widest Release (number of theaters), Domestic Gross, Foreign Gross, Opening Weekend Gross, Opening Statistics, Movie Title (foreign key) and Total Gross.

2.2.2 Unstructured Data

For unstructured data, movie reviews are scraped from *rottentomatoes.com*. The movie reviews data contains reviews in short paragraph and URL of the movie to act as foreign key.

2.3 Management of Data

The data scraped from Rotten Tomatoes and Box Office Mojo is stored into hive data warehouse.

2.4 Processing of Data

In this project, SEMMA which stands for Sample, Explore, Modify, Model and Assess is carried out for movie prediction. In processing of data stage, SEM (Sample, Explore, Modify) is performed on the input data.

1. Sample – Data is exported to a CSV file.
2. Explore – The attributes are explored using histograms to identify missing values, any inconsistencies in the data, or any hidden patterns.
3. Modify – Missing values are imputed using some pre-defined methods.

2.4.1 Feature Engineering

2.4.1.1 Rotten Tomatoes Data

The rotten tomatoes data is first exported to CSV from hive data warehouse. After the data is exported, Rotten Tomatoes Movie Info Data is checked to identify any missing values. The missing values for movie ‘genre’ column are manually imputed using the genre searched from Google, Wikipedia or IMDb. The ‘rating’ attribute which represents the MPAA (Motion Picture Association of America) film rating system is melted and converted into binary attributes such as G, NC17, NR, PG, PG_13, and R. For example, if G is true, then the movie’s

rating is classified as General Audience; while if NC17 is true, then the movie should not be viewed by children under the age of 17. The details of the ratings are as follows:

i. **G**: General audiences – All ages admitted
ii. **PG**: Parental guidance suggested – Some material may not be suitable for children.
iii. **PG-13**: Parents strongly cautioned – Some material may be inappropriate for children under 13.
iv. **R**: Restricted – Under 17 requires accompanying parent or adult guardian.
v. **NC-17**: No one 17 and under admitted.
vi. **NR:** Not Rated

Similarly, the 'genre' column is melted and clustered into 18 binary attributes, which are *'genre_Action', 'genre_Adventure', 'genre_Comedy', 'genre_Fantasy', 'genre_Horror', 'genre_Romance', 'genre_Sci-fi', 'genre_Special Interest', 'genre_Western', 'genre_FamilyKids', 'genre_AnimationManga', 'genre_FitnessSports', 'genre_DramaTele', 'genre_MusicalPerfarts', 'genre_ClassicsCult', 'genre_ArthouseInter', 'genre_ThrillMysSusp', 'genre_HistDocument'*. Using domain knowledge, similar genres are then grouped into similar genre clusters as follows:

i. genre_Action: Action (movies that exhibit action theme)
ii. genre_Adventure: Adventure (movies that exhibit adventure theme)
iii. genre_AnimationManga: Animation, Manga (movies that are animated or have japanese manga reference)
iv. genre_ArthouseInter: Art House, International (international movies)
v. genre_ClassicsCult: Classics, Cult Movies (movies that exhibit classical or are cult classics)
vi. genre_Comedy: Comedy (comedy movie)
vii. genre_DramaTele: Drama, Television (movies that are drama or TV series based)

viii. genre_FamilyKids: Family, Kids (movies for family and kids)

ix. genre_Fantasy: Fantasy (movies that exhibit a fantasy theme)

x. genre_FitnessSports: Fitness, Sports (movies that exhibit fitness or sports theme)

xi. genre_HistDocument: History, Documentary (documentary films or movies that are based on history)

xii. genre_Horror: Horror (horror movie)

xiii. genre_MusicalPerfarts: Musical, Performing Arts (movies that exhibit musical or performings arts theme)

xiv. genre_Romance: Romance (movies that exhibit a romance theme)

xv. genre_Sci_fi: Sci-fi (Science fiction movies)

xvi. genre_Special_Interest (miscellaneous movies)

xvii. genre_ThrillMysSusp: Thriller, Mystery, Suspense (movies that exhibit thriller, mystery or suspense theme)

xviii. genre_Western: Western (movies that exhibit a western theme)

Then, the genre columns that are not significant and columns that are not useful the movie prediction are dropped.

2.4.1.2 Box Office Data

Similarly, the box office data is first exported to CSV from hive data warehouse. Then, columns such as *domestic_gross, domestic_opening, foreign_gross* and *total_gross* are converted to numeric. Number of markets' exposure is also extracted from the *markets* column. To indicate the missing values in *markets*, a new column called '*markets_missing*' is added. The missing values in *total_gross* are replaced by taking the sum of *domestic_gross* and *foreign_gross*. Finally, all the columns that are not useful for the movie prediction are dropped.

2.4.1.3 Sentiment Analysis

In this section, the reviews data collected from Rotten Tomatoes is converted to sentiment score. The model is trained using the Large Movie Review Dataset v1.0 http://ai.stanford.edu/~amaas/data/sentiment/ (Maas et al., 2011). The code in the article by Aaron Kub is adapted to train a model to classify the sentiment of a review, and the model is used to classify the collection of Rotten Tomatoes reviews data and output an aggregate review's sentiment score for each movie (Kub, 2018). The sentiment score for each movie is aggregated as follows:

i. ss_mean: mean of sentiment score (ss)
ii. ss_median: median of ss
iii. ss_p25: 25th percentile of ss
iv. ss_p75: 75th percentile of ss
v. ss_std: standard deviation of ss
vi. ss_count: Total number of reviews for that movie

2.4.2 Table Properties

Next, all the features from Rotten Tomatoes Data, Box Office Data and Sentiment Analysis are merged into a data frame. The final input data has a total of 4878 rows (observations) and 44 columns (variables / attributes) as shown in Figure 1.

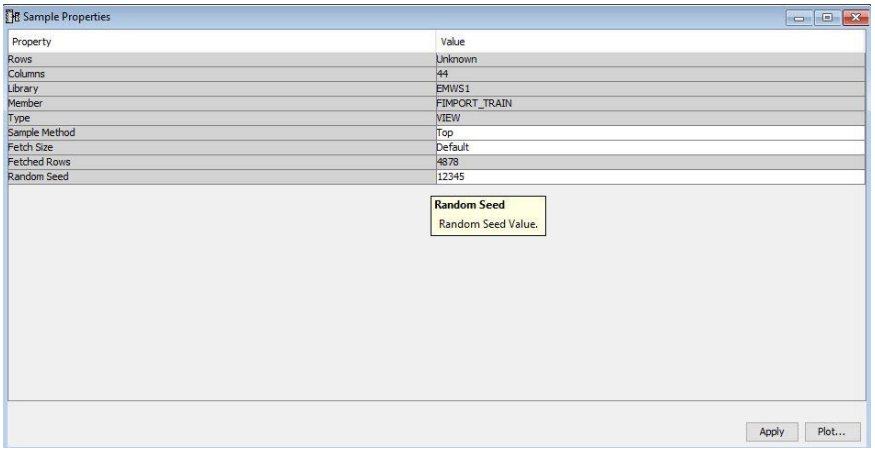

Figure 1. Sample Properties

The information of all the columns are as follows:

NAME	ROLE	LEVEL	DESCRIPTION
G	INPUT	BIARY	General audiences – All ages admitted
NC17	INPUT	BINARY	No one 17 and under admitted
NR	INPUT	BINARY	Not Rated
PG	INPUT	BINARY	Parental guidance suggested – Some material may not be suitable for children.
PG_13	INPUT	BINARY	Parents strongly cautioned – Some material may be inappropriate for children under 13.
R	INPUT	BINARY	Restricted – Under 17 requires accompanying parent or adult guardian.
audience_score	REJECTED	INTERVAL	The audience rating in rottentomatoes.com
audience_score_positive	TARGET	BINARY	A binary indicator that indicates whether the movie is good or not (in the perspective of the audience)
domestic_gross	INPUT	INTERVAL	Total domestic gross amount ($)
domestic_opening	INPUT	INTERVAL	Total domestic opening gross amount ($)
foreign_gross	INPUT	INTERVAL	Total foreign gross amount ($)
genre_Action	INPUT	BINARY	Action genre
genre_Adventure	INPUT	BINARY	Adventure genre
genre_AnimationManga	INPUT	BINARY	Animation or Manga genre
genre_ArthouseInter	INPUT	BINARY	Art House or International genre
genre_ClassicsCult	INPUT	BINARY	Classics or Cult Movies genre
genre_Comedy	INPUT	BINARY	Comedy genre
genre_DramaTele	INPUT	BINARY	Drama or Television genre

NAME	ROLE	LEVEL	DESCRIPTION
genre_FamilyKids	INPUT	BINARY	Family or Kids genre
genre_Fantasy	INPUT	BINARY	Fantasy genre
genre_FitnessSports	INPUT	BINARY	Fitness or Sports genre
genre_HistDocument	INPUT	BINARY	History or Documentary genre
genre_Horror	INPUT	BINARY	Horror genre
genre_MusicalPerfarts	INPUT	BINARY	Musical or Performing Arts genre
genre_Romance	INPUT	BINARY	Romance genre
genre_Sci_fi	INPUT	BINARY	Science Fiction genre
genre_Special_Interest	INPUT	BINARY	Special Interest genre
genre_ThrillMysSusp	INPUT	BINARY	Thriller, Mystery or Suspense genre
genre_Western	INPUT	BINARY	Western genre
markets	INPUT	INTERVAL	Number of markets exposure
markets_missing	INPUT	BINARY	Missingness indicator of 'markets' column
movie_score	REJECTED	INTERVAL	The tomatometer rating in rottentomatoes.com
movie_score_positive	REJECTED	BINARY	A binary indicator that indicates whether the movie is good or not (in the perspective of the movie critics)
runtime	INPUT	INTERVAL	Movie length in minutes
ss_count	INPUT	INTERVAL	Number of text reviews
ss_mean	INPUT	INTERVAL	Mean sentiment scores
ss_median	INPUT	INTERVAL	Median sentiment scores
ss_p25	INPUT	INTERVAL	25th percentile of sentiment scores
ss_p75	INPUT	INTERVAL	75th percentile of sentiment scores
ss_std	INPUT	INTERVAL	Aggregate standard deviation of sentiment scores
title	ID	NOMINAL	Movie title (ID)

NAME	ROLE	LEVEL	DESCRIPTION
tomatometer_count	INPUT	INTERVAL	Number of ratings given by movie critics rottentomatoes.com
total_gross	INPUT	INTERVAL	Total gross amount ($)
user_rating_count	INPUT	INTERVAL	Number of ratings given by verified users in rottentomatoes.com

Table 2. Metadata Table

2.4.3 Exploration of Data

To explore the data, histograms are created for each of the columns as follows:

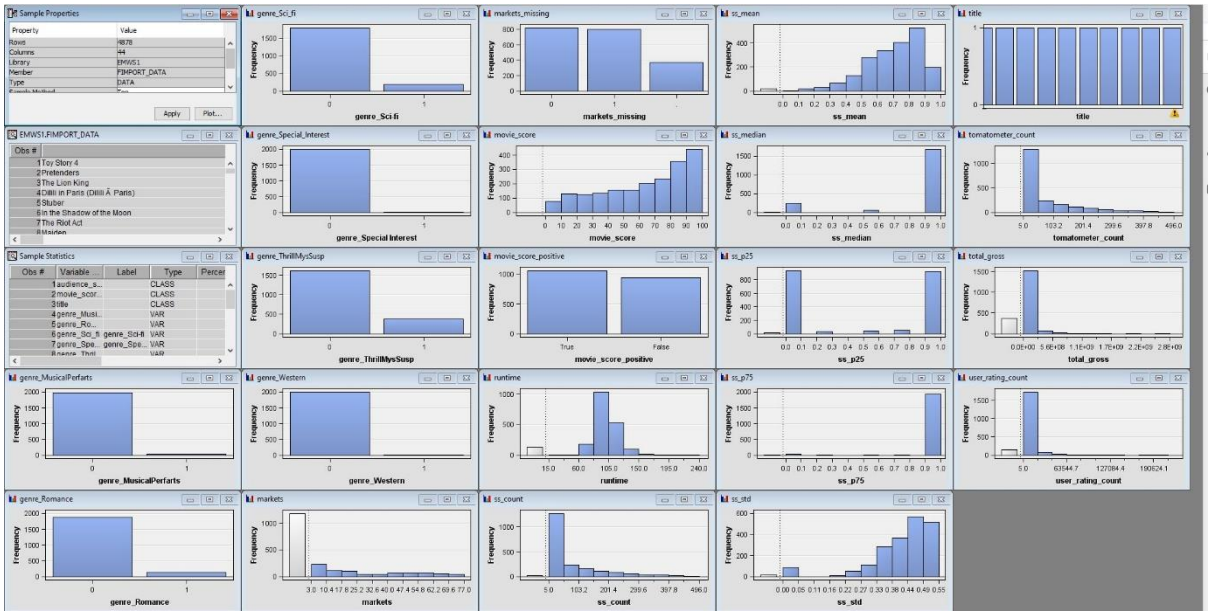

Figure 2. Histograms of Data Columns

From histograms shown in Figure 2, missing data. For example, *domestic_opening, foreign_gross, audience_score, domestic_gross, total_gross, runtime, user_rating_count, markets, ss_mean, ss-median, ss_p25, ss_p75, ss_count, ss_std* have missing values.

2.4.4 Cleansing of Data

The missing values are imputed using the Tree Surrogate method in SAS Enterprise Miner.

2.4.5 Proportion of Target Variable

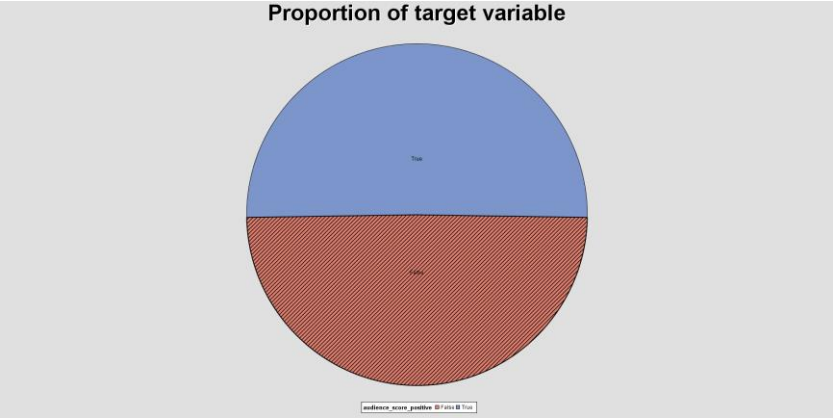

Figure 3. Pie Chart of Proportion of Target Variable

A pie chart is created to examine the proportion of the target variable. The pie chart in Figure 3 shows that the target variable (*audience_score_positive*) has a proportion of approximately 50% True (good movie) and 50% False (bad movie). This shows that the target class is balance and can be fed into a model for training.

2.5 Descriptive Analysis

Descriptive analysis is an important step for conducting an analysis. It helps in providing a basic understanding of the distribution of the data, detecting outliers and errors and identifying the relationships among the variables of the data. In descriptive analysis, E stage of SEMMA is performed on input data.

2.5.1 Clustering

2.5.1.1 Segment Size

Hierarchical clustering is used to explore the data as the number of clusters no need to be specified.

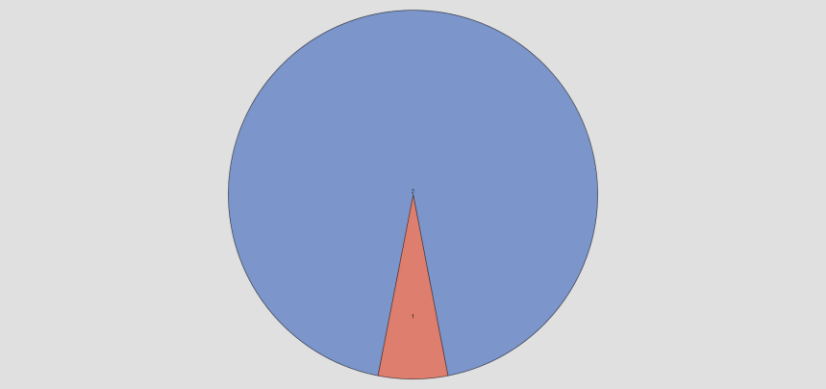

Figure 4. Segment Size Plot

The distribution of the input variables is shown by clusters. As shown in Figure 4, two clusters are formed, where the red segment represents good movie and blue segment represents bad movie.

2.5.1.2 Segment Profile

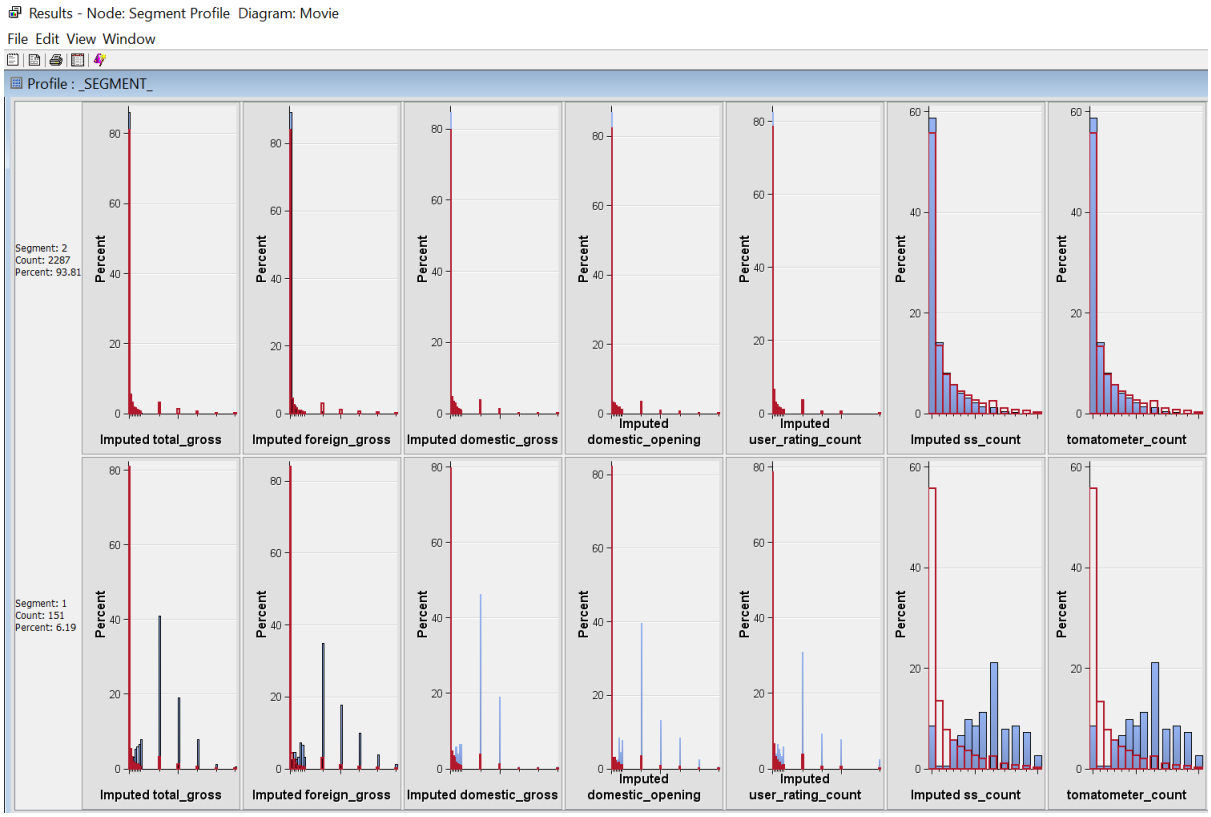

Figure 5. Segment Profile

As seen in Figure 5, box office and number of reviews are more significant for clustering compared to the genres of the movie. The red graph represents the original distribution and the blue graph represents the distribution of the clustered data. Segment 1 contains only 151 cases whereas segment 2 contains 2287 cases. For segment 1, the overall box office distribution and number of reviews are higher than average. For segment 2, the overall box office distribution and number of reviews are lower than average. Therefore, segment 1 can be considered as good and recommended movie as it contains only very few cases other than having a higher-than-average box office and higher-than-average number of reviews. Segment 2 contains a lot of cases and slants towards bad movies, hence it requires further investigation.

2.5.2 Scatter Plot

A scatter plot is also created to investigate the relationship between the mean sentiment score and audience score.

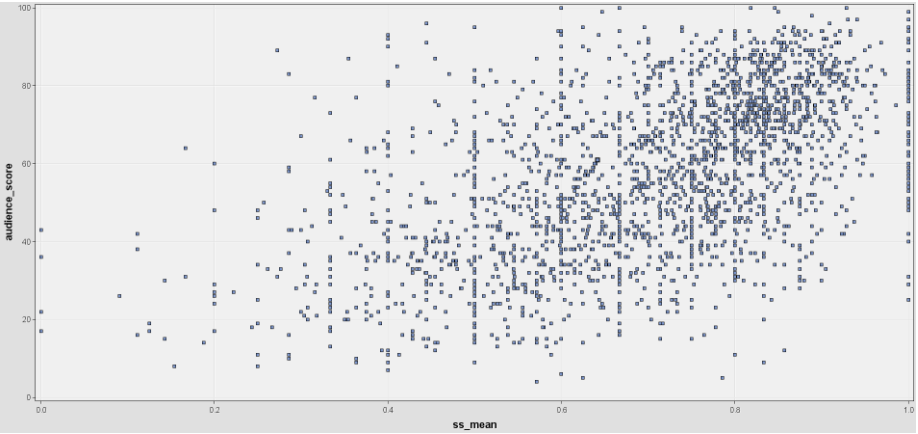

Figure 6. Scatter plot of audience_score against ss_mean

From Figure 6, it can be seen that the audience score increases with the mean sentiment score. This indicates that the audience rating has a positive relationship with the sentiment score which is derived from the movie reviews.

2.6 Diagnostic Analysis

Diagnostic analysis takes the insight found from descriptive analysis and drills down to find the cause of that outcome. The top 20 positive and negative features are extracted using natural language processing as follows:

Positive Sentiment	Negative Sentiment
Excellent	Worst
Perfect	Awful
Great	Boring
Wonderful	Waste
Amazing	Bad
Superb	Poor
Enjoyable	Terrible
Best	Dull
Today	Poorly
Fun	Disappointment
Enjoyed	Disappointing
Brilliant	Unfortunately
Must see	Worse
Loved	Stupid
Fantastic	Horrible
Liked	Mess
Incredible	Nothing
Funniest	Lame
Wonderfully	Lacks
Better than	Save

Table 3. Top Positive and Negative Features

Positive features are the words that typically appear in the reviews of a good movie, whereas negative features are the words that typically appear in the reviews of a bad movie.

2.7 Predictive Analysis

Predictive analysis extracts information from the input data in order to determine patterns and predict future outcomes and trends. In predictive analysis, M stage of SEMMA which applies a model to the data is performed.

2.7.1 Decision Tree

An interactive decision tree is developed as the model is simple and easy to interpret.

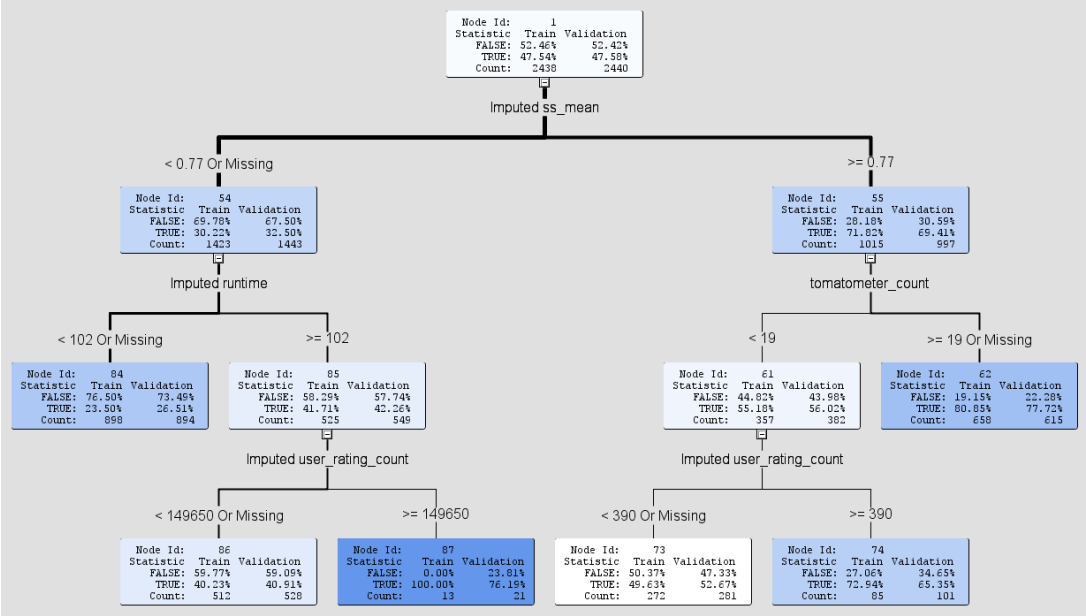

Figure 7. Decision Tree

The nodes of decision tree are split based on the information gain and interpretability of the variables. The variable with highest information, which is the highest logworth value is normally selected. However, variable with second or third highest information gain may be selected if it has higher interpretability as it is easier to interpret.

As seen in Figure 7, the first split is based on the mean sentiment score. The training data is partitioned into two subsets. The first subset, corresponding to movies with mean sentiment score lower than 0.77 have a higher than average concentration of being a bad movie, whereas the second subset, corresponding to movies with mean sentiment score greater than or equal to 0.77 have a higher than average concentration of being a good movie.

The first subset is further split based on movies running time. Movies with running time shorter than 102 minutes have a concentration of higher than 70% to be bad, while movies with running time longer than or equal to 102 minutes require further split. The next split is based on the number of ratings given by verified rotten tomatoes users. Movies with lesser than 149650

users giving the ratings are likely to be bad movies whereas movie with at least 149650 users giving the ratings are likely to be good movies.

The second subset is further split based on the number of ratings given by rotten tomatoes movie critics. Movies with at least 19 critics giving the ratings have a concentration of higher than 60% to be good whereas movies with less than 19 critics giving the ratings require further split. The next branch is then split based on the number of ratings given by verified rotten tomatoes users. Movies with lesser than 390 users giving the ratings are likely to be bad whereas movie with at least 390 users giving the ratings are likely to be good.

2.7.2 Logistic Regression

Logistic regression is a statistical regression model used to model the probability of a discrete set of classes. A logistic regression is also developed for movie prediction as the model is simple to implement and efficient to train.

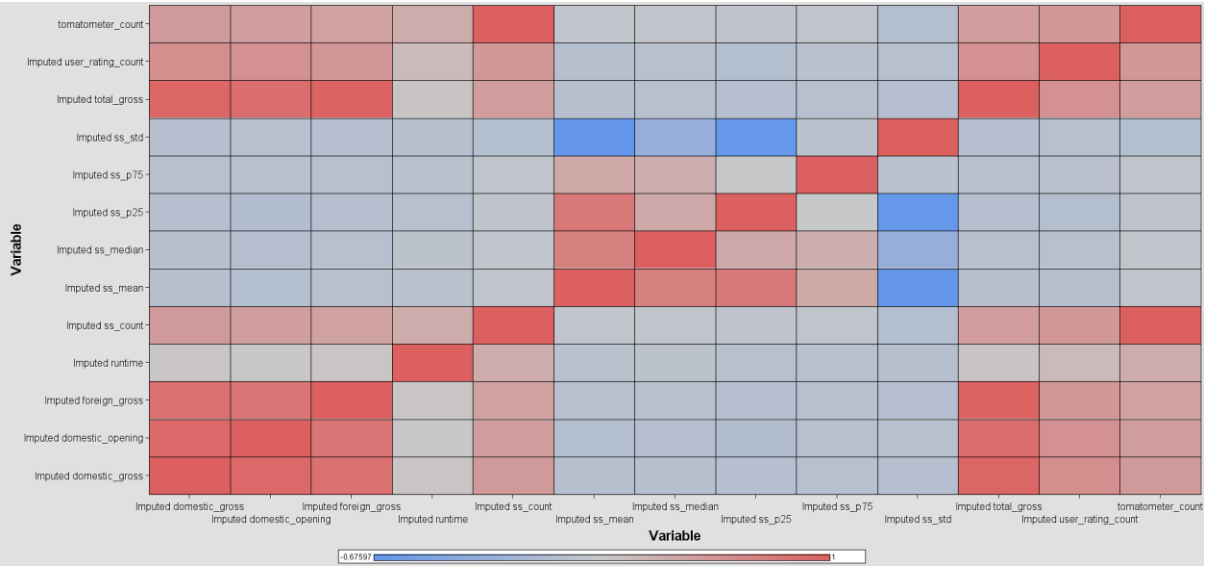

Figure 8. Correlation between Variables

Since logistic regression assumption includes absence of multicollinearity, the correlation between variables is first examined to remove features that are highly correlated to each other. A heatmap with correlation matrix between variables as shown in Figure 8 is created and sorted by colour, where red represents high collinearity and blue box represents low collinearity.

Imputed total_gross, *Imputed foreign_gross*, *Imputed domestic_opening* and *Imputed domestic_gross* are highly correlated to each other, and *Imputed total_gross* is selected to represent all these variables as it includes both the foreign and domestic earnings. As for sentiment score, *Imputed ss_p75*, *Imputed ss_p25*, *Imputed ss_median* and *Imputed ss_mean* are highly correlated, and *Imputed ss_mean* is selected as representation. *tomato_count* is selected to represent *Imputed user_rating_count*, *tomato_count* and *Imputed ss_count*. This is because movie critics watch more movies and have more professional views as compared to general audience. All the other variables seem to be significant and useful for the movie prediction and are preserved.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	-4.5011	0.5828	59.65	<.0001		0.011
IMP_runtime		1	0.0154	0.00328	22.13	<.0001	0.1593	1.016
IMP_ss_mean		1	5.1110	0.3489	214.62	<.0001	0.4918	165.837
R	0	1	0.2007	0.0562	12.74	0.0004		1.222
genre_AnimationManga	0	1	-0.3999	0.1367	8.56	0.0034		0.670
genre_DramaTele	0	1	-0.2091	0.0565	13.68	0.0002		0.811
genre_FitnessSports	0	1	-0.9172	0.3881	5.59	0.0181		0.400
genre_HistDocument	0	1	-0.5957	0.0797	55.93	<.0001		0.551
genre_Horror	0	1	0.2259	0.0943	5.74	0.0166		1.253
genre_Sci_fi	0	1	0.1687	0.0936	3.25	0.0716		1.184
genre_ThrillMysSusp	0	1	0.1377	0.0673	4.18	0.0409		1.148
tomatometer_count		1	0.00581	0.000726	64.03	<.0001	0.2692	1.006

Figure 9. Analysis of Maximum Likelihood Estimates

Both the entry and stay significance level of the model are set to be 0.1 to include more variables in the estimation. The selected variables are fitted into the logistic regression and the fitted model can be written as

$$log\frac{p}{1-p} = -4.5011 + 0.0154x_1 + 5.1110x_2 + 0.2007x_3 - 0.3999x_4 - 0.2091x_5 - 0.9172x_6$$

$$- 0.5957x_7 + 0.2259x_8 + 0.1687x_9 + 0.1377x_{10} + 0.00581x_{11}$$

where p represents the predicted probability of the outcome, x_1 , x_2 , x_3 , x_4 , x_5 , x_6 , x_7 , x_8 , x_9 , x_{10}, x_{11} represent *IMP_runtime, IMP_ss_mean, R, genre_AnimationManga, genre_DramaTele, genre_FitnessSports, genre_HistDocument, genre_Horror, genre_Sci_fi, genre_ThrillMysSusp* and *tomatometer_count* variables respectively and the slope of each

variable is the regression coefficient. The regression coefficient represents an average increase of log odds per unit increase of variable. A positive regression coefficient indicates that the mean of the dependent variable increases with the value of independent variable, hence increases the likelihood of being a good movie, whereas a negative regression coefficient indicates that the mean of the dependent variable decreases with the value of independent variable, hence reduces the likelihood of being a good movie. The higher the regression coefficient, the higher the influence of the variable on the log odds. As seen from the fitted model equation, mean sentiment score with the regression coefficient value of 5.1110 has the highest positive influence on log odds, which indicates that movie with higher mean sentiment score is likely to be a good movie. Restricted movie which is movie with MPAA of R has the third highest positive influence as the content of the movies are often more interesting and thrilling. As for movie genre, horror movie has the highest positive influence, followed by science fiction movie. This might be due to horror movie is often combined with science fiction movie. For instance, movies with theme of human threatened by aliens like Alien often capture and hold the audience's attention. On the other hand, sports movie with the regression coefficient value of -0.9172 has highest negative influence and is likely to be a bad movie. Audience are less interested in sports movies as they think that the plots of the movies are very similar and boring. Besides that, audience also think that watching sports movies is a waste of time as it is not meaningful watching others win their game.

2.7.3 Performance of Models

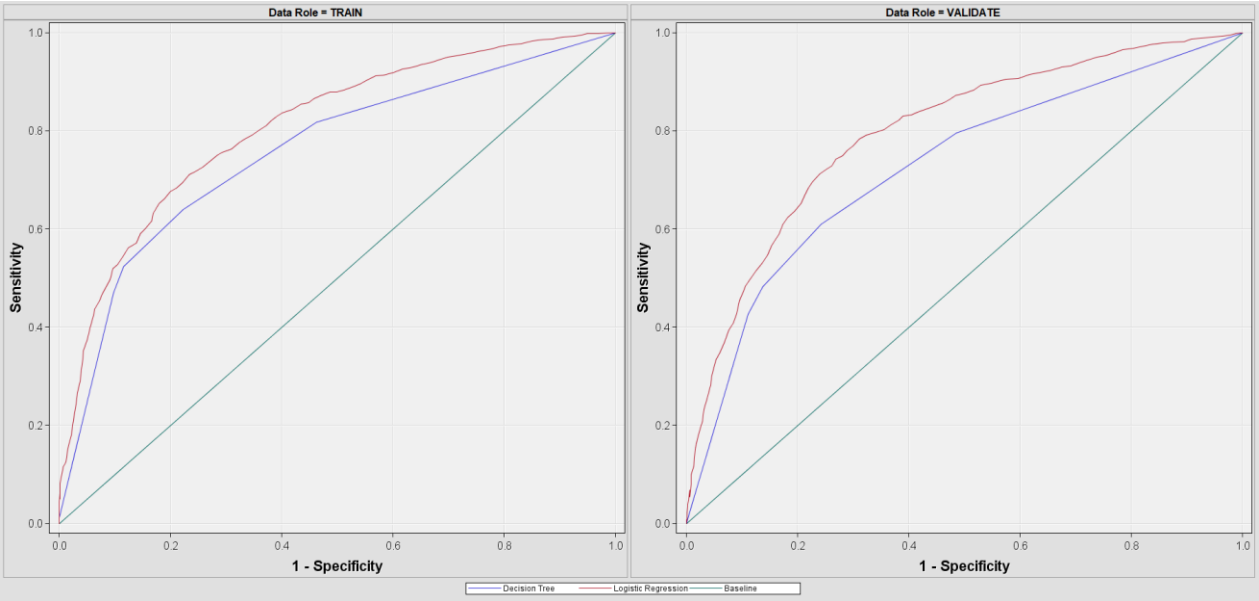

Figure 10. Receiver Operating Characteristic Chart of Decision Tree and Logistic Regression

Model	Sample	Accuracy (%)	Sensitivity (%)	Specificity (%)
Decision Tree	Training	71.25	52.37	88.35
	Validation	68.11	48.23	86.16
Logistic Regression	Training	73.38	72.56	74.12
	Validation	73.36	72.87	73.81

Table 4. Model Evaluation of Decision Tree and Logistic Regression

The performances of both the decision tree and logistic regression model are compared in order to select the better model for prescriptive analysis. The performances of the models are compared in terms of accuracy, sensitivity and specificity using the formulas as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative respectively. Logistic regression outperforms decision tree in terms of accuracy and sensitivity. Although decision tree achieves a higher specificity as compared to logistic regression, the model compromises too much on the sensitivity. Besides that, the difference between the training and validation set is also smaller in logistic regression, indicating that there is no issue on overfitting. The overall performance of logistic regression is better than that of decision tree. Hence, logistic regression is selected for prescriptive analysis.

2.8 Prescriptive Analysis

Prescriptive analysis focuses on finding the best course of action in a scenario given the available data. In prescriptive analysis, A stage of SEMMA is performed to assess whether the model is useful. Based on the logistic regression, movies with high sentiment score will be recommended to the users as movies with positive reviews are likely to be good movies. Movies with genre of horror and science fiction will also be recommended to the users as movies with both of these genres capture the attention of audience and are likely to be good movies. Besides that, movies with MPAA of R (Restricted) will be recommended to the users provided that the age of the user is above 17. Hence, it is suggested that the users that subscribe to the movie streaming service should register with their identity number so that appropriate movies can be recommended to the users.

2.9 Source Code

The source code for data scraping, data storage, feature engineering, sentiment analysis and modelling is available at: *https://github.com/yinyen/DataMiningProject* .

CHAPTER 3: CONCLUSION

In this project, a movie recommender system is developed using decision tree and logistic regression models. The results show that logistic regression is robust and produces a good performance for the movie prediction. Logistic regression is simple to implement, and the regression coefficients can be easily interpreted. Interpretability is important for decision makers to understand the influence of variables on the outcome of the prediction, which is whether the movie is good or bad. Hence, movies can be predicted as good or bad based on the box office performance and user reviews using logistic regression. Typically, words such as excellent, perfect and great appear in a good movie review while words such as worst, awful and boring appear in a bad movie review. Movies with high sentiment scores and genres of horror and science fiction are likely to be good movie and should be recommended to the users. Besides that, restricted movies should also be recommended to the users of age 17 and above.

REFERENCES

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 142-150). Association for Computational Linguistics.

Kub, A. (2018). IMDb sentiment analysis. Retrieved December 22, 2019, from GitHub website: https://github.com/aaronkub/machine-learning-examples/tree/master/imdb-sentiment-analysis