Title: Classification and clustering of various laptop brands of the project

Group members: [redacted] and Natibundit Suntraranusorn (300360042)

PROJECT OBJECTIVES

The objective project is to extract meaningful information from the laptops purchase dataset by using R programming and Tableau to explore the data and build models that can predict the price category of laptops based on specifications and ratings. The processes will include exploring the data, preprocess, building different models, evaluating its performance, and presenting the results.

The key goals:

- Data Exploration and Cleaning to initially understand a descriptive data and apply data cleaning techniques to ensure the dataset is suitable for analysis.
- Explore the relationships between various variables in the dataset to find patterns or trends.
- Apply scoring and classification techniques such as Logistic regression, Random Forest including optimizing the predictive model by tuning hyperparameters and evaluating performance to efficiently increase accuracy for prediction.
- Apply K-Mean clustering to find the customer segmentation based on their laptop preferences.
- The findings and insights derived from the analysis will help in understanding customer preferences and predicting laptop prices.

DATA SET OVERVIEW

The dataset contains 23 columns and 896 records of purchased laptops, which incorporate all several key factors of laptop such as brand name, CPU, storage, weight, screen size, price, star rating. The original data was collected by using an automated chrome web extension tool called Instant Data Scrapper from flipkart.com. However, The dataset is downloaded from https://www.kaggle.com/datasets/kuchhbhi/latest-laptop-price-list?resource=download.

DATA EXPLORATION

The dataset was explored, which aims to better understanding about the aspect for the dataset. In addition, the exploration encompassed as the following descriptive data:

Exploration

- The top 6 brands include Acer, Asus, Dell, HP, Lenovo and MSI, which contributes 91% of the total data.
- Average price of laptops is compared with the average discount and the average star rating to initially examine the assumption of lower price and greater discount can dramatically impact the customer satisfaction. For example, the chart indicates that Lenovo adopts a pricing strategy of setting their prices at an average level and provides the highest discount when compared to its competitors. Despite this, the star rating show in average, which stand at 2.9. On the other

hand, MSI sold laptops at a higher price with an average discount. Surprisingly, customer satisfaction for MSI laptops is better than that of most other brands, except for HP. As a result, discount and price and discount are not essentially correlated with customer satisfaction.

- The boxplot demonstrates the dispersion of data that shows the interquartile range and Minimum & Maximum data to understand the overall data point of the top 4 brands, especially an outliner.
- the distribution of price lengths by displaying as a histogram, indicating that the data exhibits a left-skewed pattern. The majority of prices fall within the range of 0 to 2,000, contributing to approximately 90% of the total price length.
- For the correlation heat map, ram_gb, ssd, graphic_card_gb are selected to be independent variables based on their high correlation with the target variable, which exceed a positive value of 0.4 (rounded to one digit). Notably, the ssd exhibits the strongest positive correlation with the price, with a correlation coefficient of 0.59. In addition to the selected independent variables, star_rating and brand are included due to these two factors could affect pricing despite the star_rating has a negative correlation of -0.2 and the brand is a categorical value which could not plot in the heat map.
- MSI (the leading brand in high-end gaming) and Apple are the worldwide reputation for quality, showing over 4 star rating that define high customer satisfaction. Furthermore, high processors such as Core i9, Core m3, GeForce RTX, Genuine Windows, Pentium Quad and Ryzen9 is the one important factor to indicate the high impressive for client's star rating.
- Upgrading the quality of the graphics card would sightly increase the laptop price.

DATA PRE-PROCESSING

The data pre-processing was performed as follows:

1. The missing values were filled with "Missing" therefore we firstly replaced "NA" to "Missing" to capture the missing values.
2. There were inconsistent names of the brand, Lenovo, in terms of upper and lowercase.
3. Computer components columns i.e., SSD, HDD, processor generation, display size, and ram, could be converted into numeric columns therefore subset function was applied to remove characters and set the number into numeric.
4. The data was scrapped from an Indian website therefore the currency in this data is Indian Rupee. Thus, the laptop price column was converted into CAD with the fixed currency exchange rate 1 CAD: 62 Rupee.
5. There were 666 missing values; 95 of model values, 239 of processor generation values, 332 of display size values; We handled missing values using "vtreat" library.
6. Outliers' removal was achieved by using boxplot.
7. Create a correlation heatmap to view the correlation in order to determine the independent variables and ensure no multicollinearity e.g., discount, old price.
8. In labelling the dependent variable, latest laptop price, into 3 price ranges based on the result of the clustering method.

Once the dataset was preprocessed to handle missing values and outliers, normalizing the data was performed including splitting into a training set (60%) and a test set (40%) for model evaluation.

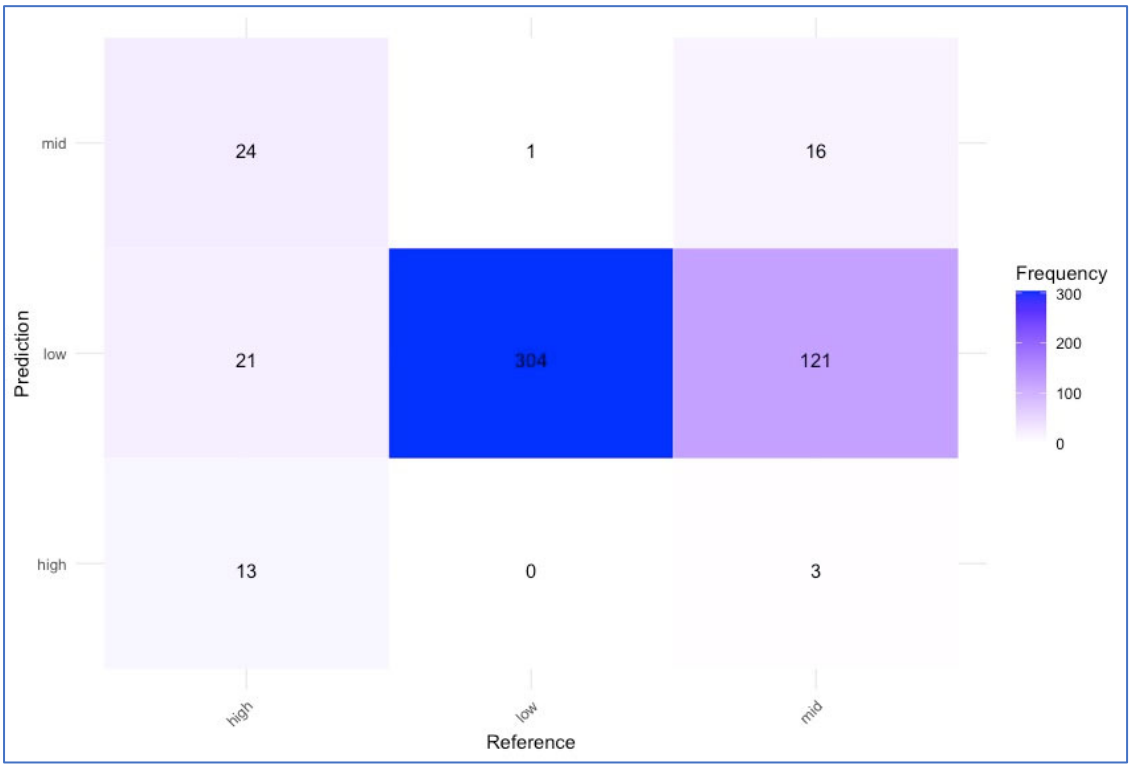
INSIGHTFUL FINDINGS

Classification

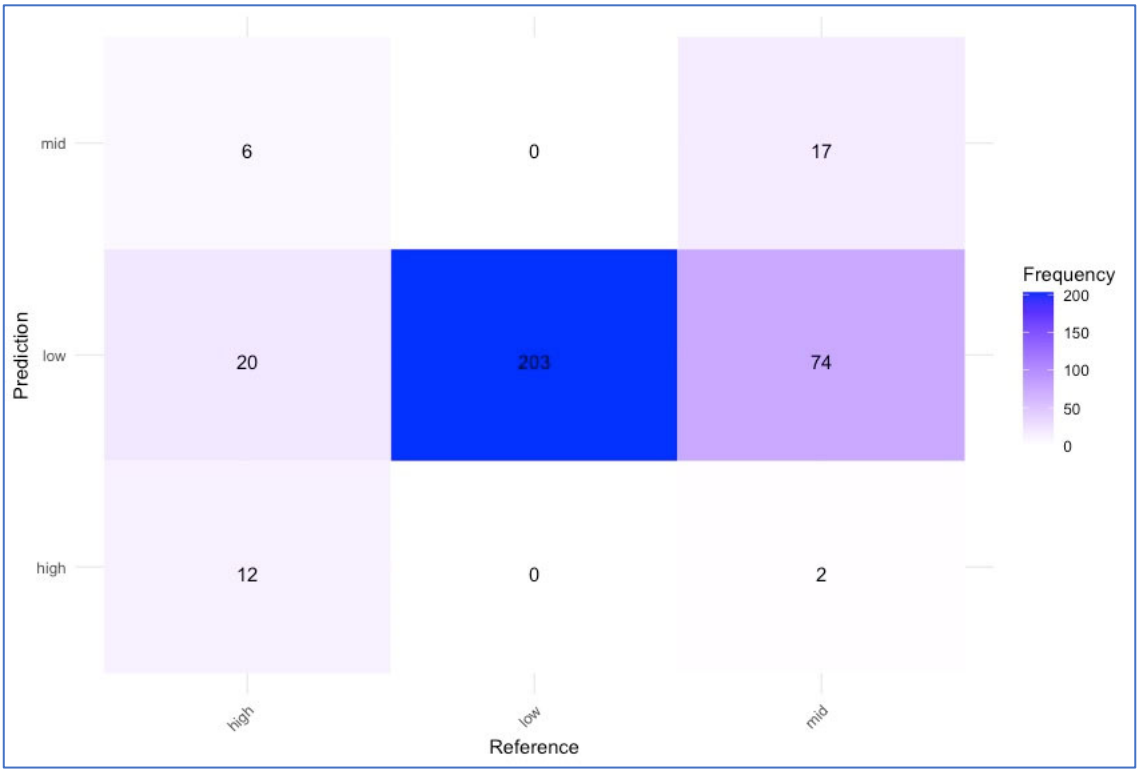
1. Logistic Regression: accuracy on Training Data: 66.0%, accuracy on Test Data: 69.2%
2. Random Forest with tuning and different type:
   2.1 Random Forest with Type Class: Accuracy on Training Data: 89.3% and 75.2% on Test Data.
   2.2 Random Forest with Type Response: Accuracy on Training Data: 88.1% and 74.9% on Test Data.
   2.3 Random Forest with tuning ntree, mtry, nodesize: Accuracy on Training Data: 80.7% and 75.7% on Test Data.

The following visualization illustrated the confusion matrix of the tuned Random Forest model:

Confusion matrix of train set



Confusion matrix of test set



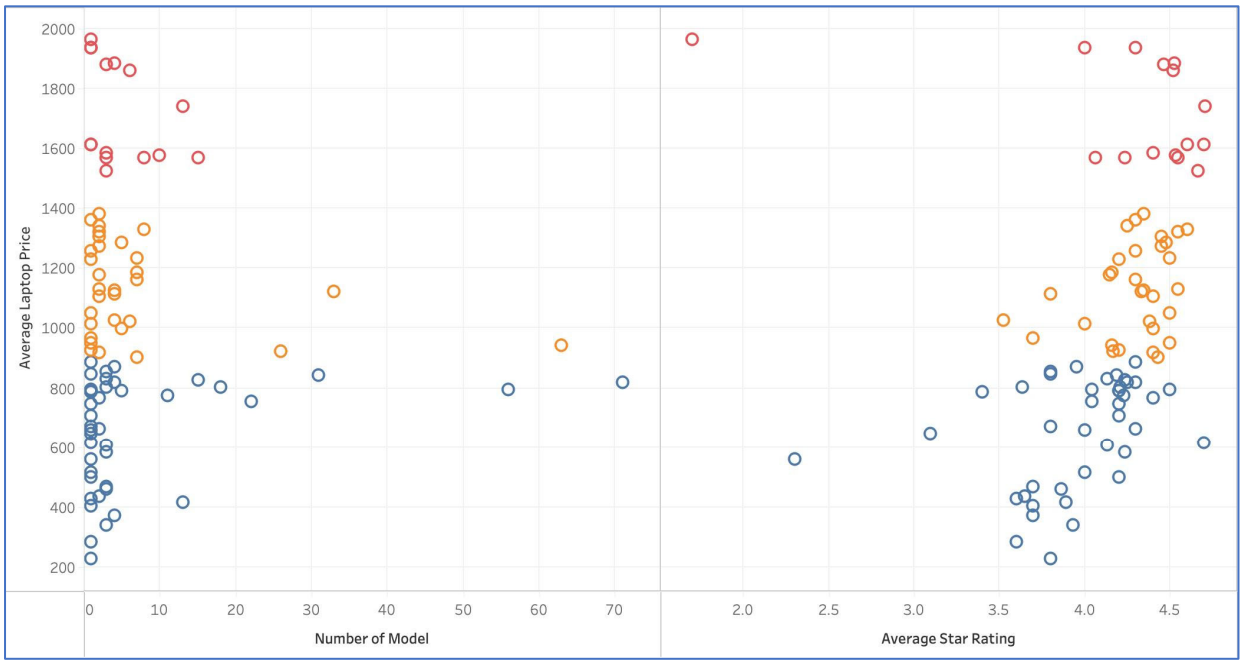
Random Forest Model has given the best result comparing to others models therefore it was employed and tuned in various settings and types. There were two versions of the random forest model were employed. The first version used the “class” predictions and the second used the “response” given probabilities. Both models gave an accuracy around 75% on the test set. Next, a hyperparameter tuning was applied to the random forest model to optimize the performance. The number of trees (ntree), the number of features (mtry), and the minimum size of terminal nodes (nodesize) were tuned. The best model achieved an accuracy of approximately 80.7% on the training set and 75.7% on the test set.

The results of the models are shown as follows:

```
> result_modeling
                                        method accuracy_train accuracy_test
1                                        logit      0.6580517     0.6916168
2                  Random Forest with type class      0.8926441     0.7514970
3              Random Forest with type response      0.8807157     0.7485030
4 Tuning Random Forest (ntree, mtry, nodesize)      0.8071571     0.7574850
```

Clustering

In Tableau, K-means clustering are examined in different k numbers (3,4,5) to find the properly meaningful clusters by using the number of models and the average star rating as the independent variables while the average laptop price as the dependent variable to form a clustering model. As a result, k equals 3 is the best representative for each cluster that separate laptop into 3 segmentations include “low price”, “medium price”, “high price”.



For R studio, high correlation variables with laptop price are selected to build the K-Mean model. Plotting WSS for a Range of K (Elbow), the Calinski-Harabasz index (ch) and the average silhouette width (asw) are applied to consider the best k number, which suggests k=2 to 4, k=3, and k = 10 in relatively different techniques. After the different k number are tested the result of experiment show that k equals 3 can efficiently classify 3 types of laptop, which incorporate “High feature laptop”, “Medium feature laptop”, “Low feature laptop”. The high feature laptop are the higher processor and storage in comparison with other groups, which contain 116 transactions from 587 transactions in total.

CONCLUSION

This report presents the evaluation of different machine learning models applied to a dataset containing information about laptop prices. The predictive models include logistic regression, random forest with variations in parameter settings are applied, which Random Forest Model (ntree, mtry, nodesize) has given the best accuracy. The clustering method nominated for a few k numbers and 3 clusters was selected resulting in predicting the price category of laptops based on three classes: “low price”, “medium price”, “high price”.