MULTIMEDIA UNIVERSITY

FACULTY OF COMPUTING AND INFORMATICS

BACHELOR IN COMPUTER SCIENCE (HONS)

MACHINE LEARNING - CDS6354

TRIMESTER II, SESSION 2024/2025

Prediction of Starbucks Customers Loyalty Status

By:

Ooi Li Yoong (1211306826)

Acknowledgment

We are immensely thankful to our instructor, Dr. Wong Ya Ping, whose insightful guidance has been pivotal throughout this project’s journey.

Table of Content

Acknowledgment 2
Table of Content 3
Chapter 1: Project Overview 4
1.1 Background 4
1.2 Problem Statement and Project Objective 4
1.3 Chapter Outline 5
Chapter 2: Methodology 6
2.1 Introduction 6
2.3 Data Pre-processing 7
2.4 Classification 7
Chapter 3: Findings 10
3.1 Introduction 10
3.2 Exploratory Data Analysis (EDA) 10
3.3 Classification 16
Chapter 4: Conclusion 17

Chapter 1: Project Overview

1.1 Background

In today's competitive F&B market, understanding customer spending behavior and maintaining customer loyalty are vital to ensure a sustainable business. Starbucks, a global leader in the coffee industry, has been continuously improving its customer relationship management and customer retention strategies. For such purposes, it is necessary to have the ability to accurately identify the loyalty status of each customer and understand the factors that influence loyalty so that effective business strategies for boosting customer experience can be formulated, and therefore, new customers can be developed to be loyal and existing customers can remain attracted to the business.

It is striking that the first step to emphasize is the prediction of customer loyalty status. Therefore, this study aims to develop predictive models to categorize customers into loyal and non-loyal customers based on various features such as income, time spent at Starbucks, and many more. By utilizing classification techniques, we will create models to accurately classify the loyalty status of customers; thus, allowing Starbucks to effectively plan its marketing strategies.

Our framework consists of several stages, and it will start with data cleaning to result in an analytical dataset of optimal quality. Then, we will construct different classification models under various experimental settings. Following that, the models will be evaluated and compared using metrics such as accuracy, precision, recall, and F_1-score. From there, we will identify the optimal classifier that can be used for predicting the loyalty status of customers in the future.

The subsequent process should be formulating different sets of strategies tailored to loyal and non-loyal customers respectively. The detailed strategy planning requires the aid of marketing experts, and therefore, will not be covered in this study. However, we will provide brief and basic strategies to showcase the potential of this project.

1.2 Problem Statement and Project Objective

This project mainly tackles the part of predicting the loyalty status of Starbucks customers. Considering the large number of customers, manual analysis can be extremely resource intensive. Therefore, automation is preferred, and our objective is to identify an optimal classifier to predict the loyalty status of Starbucks customers.

1.3 Chapter Outline

The rest of this report is organized as the following: Chapter 2 describes the methods leveraged to achieve the objectives; chapter 3 presents the findings; chapter 4 summarizes this study’s achievements and future directions.

Chapter 2: Methodology

2.1 Introduction

This chapter explains the methods employed in this study. Figure 2.1.1 shows our workflow. Firstly, we perform an exploratory data analysis (EDA) on our dataset to understand its properties. Then, we carry out necessary data pre-processing to prepare it for the upcoming classification. We attempt some classification algorithms to predict the loyalty status of Starbucks customers. Following that, we evaluate the constructed classifiers with evaluation metrics and eventually decide the best-performing classifier. Notably, the random states are always set to 42 whenever applicable to ensure that the results are reproducible.

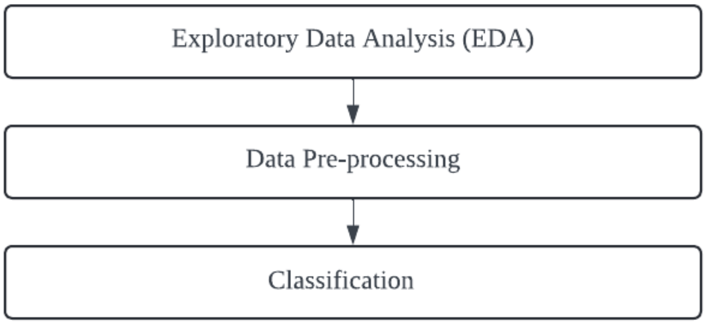

Figure 2.1.1: Workflow

2.2 Exploratory Data Analysis (EDA)

EDA is crucial to understand more about the dataset. We try to understand the available feature with the aid of the provided description from the data source (https://www.kaggle.com/datasets/mahirahmzh/starbucks-customer-retention-malaysia-survey). In addition, we check for any missing value and duplicated row. From there, data cleaning is performed if necessary. Finally, we visualize the distribution of the features to gain more information if possible.

2.3 Data Pre-processing

One-hot encoding (OHE) is performed to transform categorical data into numerical data because all classifiers involved in this study only accept numerical inputs. Additionally, train-test split is performed in each setting where 80% training data and 20% testing data are generated. Notably, SMOTE is applied to some settings because we found out that our class distribution is imbalanced.

2.4 Classification

We attempt three classification algorithms, namely logistic regression (LR), k-nearest neighbor (kNN), and random forest (RF). Table 2.4.1 shows a summary of different experimental settings attempted.

Experimental Setting	**Pre-processing**	**Classification Algorithm**
M_1	OHE + Train-test Split	LR
M_2	OHE + Train-test Split + SMOTE	LR
M_3	OHE + Train-test Split	kNN
M_4	OHE + Train-test Split + SMOTE	kNN
M_5	OHE + Train-test Split	RF
M_6	OHE + Train-test Split + SMOTE	RF

Table 2.4.1: Experimental Setups Attempted

Subsequently, for the evaluation, we use metrics such as accuracy, precision, recall and F_1-score. These metrics also incorporate measures of True Positives (TP), False Positive (FP), False Negative (FN), and True Negatives (TN) to assess the classifier's performance. TP means that when a positive outcome is predicted, and the actual outcome is indeed positive. FP happens when a positive outcome is predicted, but the actual outcome is negative, which is also called Type 1 Error. FN is when a negative outcome is predicted, but the actual outcome is positive, which is known as a Type 2 Error. TN occurs when a negative outcome is predicted, and the actual outcome matches that prediction.

Accuracy measures the proportion of correct predictions made by classifier model. However, researchers often avoid using accuracy as a performance metric in classification tasks involving class imbalance, as it may not effectively highlight the importance of identifying rare cases. Equation 1. shows the formula of accuracy.

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

Precision measures the percentage of predicted positive cases that are truly positive, indicating fewer false positives. It can help to control false positives, but it is not sufficient for evaluating models in class imbalance situations, as it is sensitive to the number of false positives. Equation 2. shows the formula of precision.

2. $$\Pr e cision = \frac{TP}{TP+FP}$$

Recall measures the number of actual positive cases identified by a model. It is not influenced by class imbalance since it focuses only on the positive group. Recall is vital for determining true positive rates and detecting actual positives. Equation 3. shows the formula of recall.

3. $$Recall = \frac{TP}{TP+FN}$$

F_1-score is an evaluation metric for class imbalance problems, representing the harmonic mean of precision and recall, then combining both measures. Equation 4. shows the formula of F_1-score.

4. $$F1 - score = \frac{TP}{TP+0.5(FP+FN)}$$

Finally, we select an optimal classifier that can be used to predict the loyalty status of Starbucks customers in future.

Chapter 3: Findings

3.1 Introduction

This chapter illustrates findings extracted during the EDA and classification.

3.2 Exploratory Data Analysis (EDA)

Upon initial checking, we found that the raw dataset contained 21 features. The data was collected from 122 respondents via a survey. Table 3.2.1 describes these features.

Feature	Description
Timestamp	Describes the time when the customer answers the survey
1. Your Gender	Describes the customer’s gender
2. Your Age	Describes which age group the customer belongs to. The pre-specified options include “Below 20”, “From 20 to 29”, “From 30 to 39”, and “40 and above”
3. Are you currently ....?	Describes the customer’s occupation. Unique values found are “Employed”, “Self-employed, “Student, and “Housewife”
4. What is your annual income?	Describes the customer’s annual income
5. How often do you visit Starbucks?	Describes the customer’s visiting frequency to Starbucks. This is a closed-ended question where the pre-specified answers include “Never, “Rarely”, “Monthly”, “Weekly”, and “Daily”
6. How do you usually enjoy Starbucks?	Describes the customer’s preferred visiting method, such as “Dine in”, “Drive-thru”, and others
7. How much time do you normally spend during your visit?	Describes how much time the customer usually spends at Starbucks. This is a closed-ended question where the pre-specified answers include “Below 30

	minutes", "Between 30 minutes to 1 hour", "Between 1 hour to 2 hours", "Between 2 hours to 3 hours", and "More than 3 hours"
8. The nearest Starbucks's outlet to you is...?	Describes the distance between the customer's home and the nearest Starbucks shop from it. This is a closed-ended question where the pre-specified answers include "within 1km", "1km - 3km", and "more than 3km"
9. Do you have Starbucks membership card?	States if the customer has a Starbucks membership card
10. What do you most frequently purchase at Starbucks?	Describes the product most frequently purchased by the customer
11. On average, how much would you spend at Starbucks per visit?	Describes the average spending at Starbucks by the customer in one visit. This is a closed-ended question where the pre-specified answers include "Zero", "Less than RM20", "Around RM20 – RM40", and "More than RM40"
12. How would you rate the quality of Starbucks compared to other brands (Coffee Bean, Old Town White Coffee..) to be:	Describes how the customer rates Starbucks compared to its competitor. This question uses a rating scale with values between 1 and 5 inclusively
13. How would you rate the price range at Starbucks?	Describes how the customer rates the prices of Starbucks products. This question uses a rating scale with values between 1 and 5 inclusively
14. How important are sales and promotions in your purchase decision?	Describes how the customer rates the importance or influence of promotions in purchasing Starbucks products. This question uses a rating scale with values between 1 and 5 inclusively

15. How would you rate the ambiance at Starbucks? (lighting, music, etc...)	Describes how the customer rates the ambience at Starbucks from numerous perspectives such as lighting and music. This question uses a rating scale with values between 1 and 5 inclusively
16. You rate the WiFi quality at Starbucks as..	Describes how the customer rates the quality of Starbucks Wi-Fi. This question uses a rating scale with values between 1 and 5 inclusively
17. How would you rate the service at Starbucks? (Promptness, friendliness, etc..)	Describes how the customer rates the service at Starbucks. This question uses a rating scale with values between 1 and 5 inclusively
18. How likely you will choose Starbucks for doing business meetings or hangout with friends?	Describes how the customer view Starbucks as a suitable place to have business meetings or hangouts with friends. This question uses a rating scale with values between 1 and 5 inclusively
19. How do you come to hear of promotions at Starbucks? Check all that apply.	Describes how the customer knows ongoing Starbucks promotions. Options include “social media”, “emails”, and others
20. Will you continue buying at Starbucks?	Describes the customer’s desire to make subsequent purchases at Starbucks based on numerous factors. This is the target variable.

Table 3.2.1: Description of Features in the Raw Dataset

It is clear that the names of these raw features are too lengthy. Therefore, we renamed the columns before further analysis and pre-processing. Table 3.2.2 shows the relationship between the original features and the renamed features.

Original Feature	Renamed Feature
Timestamp	timestamp
1. Your Gender	gender
2. Your Age	age
3. Are you currently ....?	occupation
4. What is your annual income?	annualIncome
5. How often do you visit Starbucks?	visitingFrequency
6. How do you usually enjoy Starbucks?	visitingMethod
7. How much time do you normally spend during your visit?	timeSpent
8. The nearest Starbucks's outlet to you is...?	distanceToStarbucks
9. Do you have Starbucks membership card?	hasMembership
10. What do you most frequently purchase at Starbucks?	favouriteProduct
11. On average, how much would you spend at Starbucks per visit?	averageExpensePerVisit
12. How would you rate the quality of Starbucks compared to other brands (Coffee Bean, Old Town White Coffee..) to be:	overallRating
13. How would you rate the price range at Starbucks?	priceRating
14. How important are sales and promotions in your purchase decision?	needPromotion
15. How would you rate the ambiance at Starbucks? (lighting, music, etc...)	ambienceRating
16. You rate the WiFi quality at Starbucks as..	wifiRating

17. How would you rate the service at Starbucks? (Promptness, friendliness, etc..)	serviceRating
18. How likely you will choose Starbucks for doing business meetings or hangout with friends?	forMeetingHangout
19. How do you come to hear of promotions at Starbucks? Check all that apply.	promotionMarketingMedia
20. Will you continue buying at Starbucks?	loyaltyStatus

Table 3.2.2: Mapping of Original Features to Renamed Features

Upon initial checking on the renamed features, we discovered that there was only one row with missing values. Given such minimal missing values, we simply drop this row. Subsequently, we found out that there was no duplicated row. We dropped the “timestamp” column as it serves no purpose in disclosing the opinions of customers.

Moving on to visualizations of distributions, we utilized bar charts. All features and the target variables are categorical. Although some features such as “overallRating”, “wifiRating” and others seem to hold numerical values, they are representing categories; in these columns, a higher discrete value such as 5 simply means a greater satisfaction with the subject in question, while a lower discrete value such as 1 implies a lower satisfaction with the subject in question. Figure 3.2.1 visualizes the distributions via bar charts.

It is obvious that the target variable has an imbalanced class distribution. Therefore, techniques such as SMOTE are expected to effectively improve the classification results.

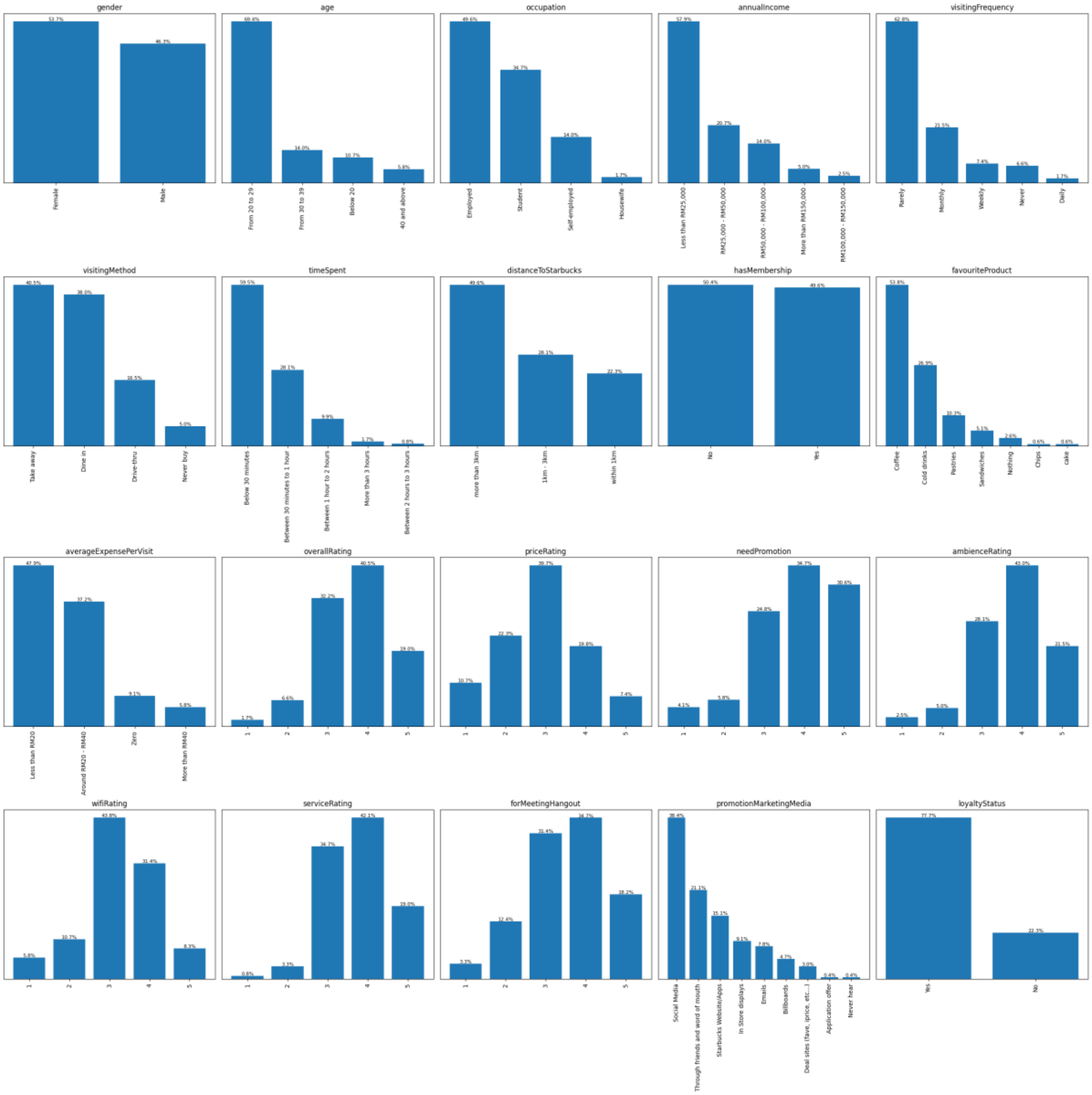

Figure 3.2.1: Bar Charts

3.3 Classification

The performances of the constructed classifiers are summarized in Table 3.31.

Model	Training Performance				Testing Performance			
	Accuracy	Precision	Recall	F_1-score	Accuracy	Precision	Recall	F_1-score
M_1	0.95	0.94	1.00	0.97	0.76	0.81	0.89	0.85
M_2	0.97	0.97	0.97	0.97	**0.80**	0.89	0.84	**0.86**
M_3	0.88	0.88	0.97	0.92	0.76	0.81	0.89	0.85
M_4	0.85	0.98	0.72	0.83	0.76	**0.93**	0.74	0.82
M_5	**1.00**	**1.00**	**1.00**	**1.00**	0.76	0.79	**0.95**	**0.86**
M_6	**1.00**	**1.00**	**1.00**	**1.00**	0.72	0.77	0.89	0.83

Table 3.3.1 Performances of Classifiers

When comparing the training and testing performance, we observe that RF models experienced the greatest drop in almost all metric values except recall. Therefore, it is valid to state that RF models are slightly overfitting our training dataset.

Considering the usefulness of SMOTE, we observe that models leveraging SMOTE such as M_2 and M_4 yielded relatively more desired results than their respective models without SMOTE, M_1 and M_3, respectively. The only exception is the recall scores. However, recall scores usually concern less regarding the class imbalance issue as it solely focuses on the positive group, and therefore, may not be a good metric here. In essence, SMOTE is useful in our study.

Concerning the choice of classification algorithms, we believe that the optimal choice is LR. Since we have class imbalance issue, F_1-score should be prioritized over all other metrics. Following that, it is striking that the optimal model is M_2 that utilizes a combination of SMOTE and LR.

Chapter 4: Conclusion

This study set forth only one objective: to identify an optimal classifier to predict the loyalty status of Starbucks customers. The findings revealed that the optimal classifier is a specialized LR model leveraging SMOTE technique as a solution to tackle the class imbalance issue. This optimal classifier can be further enhanced by fine-tuning the hyperparameter values. Eventually, it can be supported by the business to identify the loyalty status of customers.