MACHINE LEARNING APPROACH TO PREDICT E-COMMERCE CUSTOMER SATISFACTION SCORE

PONGTHANIN WANGKIAT

Graduate School Srinakharinwirot University

2022

การประยุกต์ใช้การเรียนรู้ของเครื่องเพื่อทำนายคะแนนความพึงพอใจของลูกค้าอีคอมเมิร์ซ

พงศ์ธนิน หวังเกียรติ

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2565

ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

MACHINE LEARNING APPROACH TO PREDICT E-COMMERCE CUSTOMER SATISFACTION SCORE

PONGTHANIN WANGKIAT

A Master’s Project Submitted in Partial Fulfillment of the Requirements
for the Degree of MASTER OF SCIENCE
(Data Science)
Faculty of Science, Srinakharinwirot University
2022

Copyright of Srinakharinwirot University

THE MASTER’S PROJECT TITLED

MACHINE LEARNING APPROACH TO PREDICT E-COMMERCE CUSTOMER SATISFACTION SCORE

BY

PONGTHANIN WANGKIAT

HAS BEEN APPROVED BY THE GRADUATE SCHOOL IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE MASTER OF SCIENCE
IN DATA SCIENCE AT SRINAKHARINWIROT UNIVERSITY

(Assoc. Prof. Dr. Chatchai Ekpanyaskul, MD.)
Dean of Graduate School

ORAL DEFENSE COMMITTEE

.... Major-advisor
(Asst. Prof.Chantri Polprasert, Ph.D.)

.... Chair
(Suttipong Thajchayapong, Ph.D.)

.... Committee
(Asst. Prof.Werayuth Charoenruengkit, Ph.D.)

Title	MACHINE LEARNING APPROACH TO PREDICT E-COMMERCE CUSTOMER SATISFACTION SCORE
Author	PONGTHANIN WANGKIAT
Degree	MASTER OF SCIENCE
Academic Year	2022
Thesis Advisor	Assistant Professor Chantri Polprasert , Ph.D.

This paper investigates the performance of machine learning (ML) to predict customer satisfaction scores from the sales dataset collected by Olist, the Brazilian e-commerce company. The customer satisfaction score is categorized into four classes: Low, Average, Good and Excellent. The majority of sales orders received an Excellent score. This was inspired by the fact that delivery duration and product rating score obtained from the purchases of other customers are one of the main factors that influenced the satisfaction scores of customers. A feature engineering method was proposed that creates delivery duration and an average product rating score, which are used as the main features in the ML model. Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbor (K-NN) were employed to predict customer satisfaction score and their performance was compared with the baseline model, which predicted the customer satisfaction score using the average product rating score. The results showed that the RF model yields the best performance with the average precision, recall, and macro F1 equal to 0.34, 0.36, and 0.32, respectively. In addition, RF achieves the best recall equal to 0.43, 0.33 and 0.33 for Low, Average and Good classes. The mean and SD of the product rating are two features with the highest feature importance equal to 0.313 and 0.087, respectively.

Keyword : E-commerce, Classification, Customer satisfaction, Machine learning, Rating prediction

ACKNOWLEDGEMENTS

I would like to express my gratitude and sincere thanks to my advisor, Asst. Prof. Chantri Polprasert, Ph.D. for his invaluable support, encouragement, and guidance in developing and completing this master's project.

My sincere appreciation is also extended to my proposal and oral examination committee, Dr. Suttipong Thajchayapong, Asst. Prof. Werayuth Charoenruengkit, Ph.D., and Subhorn Khonthapagdee, Ph.D. for their valuable suggestions and comments.

In addition, I would like to thank for partially supported of a research grant from the Graduate School, Srinakharinwirot University.

Finally, I would like to dedicate my thanks to my family, my colleagues at Shopee (Thailand), and my friends for their encouragement and support throughout my studies.

PONGTHANIN WANGKIAT

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

Chapter 1
Introduction

1.1 Research Background

The term "e-commerce" stands for electronic commerce. It refers to the online marketplace for purchasing or selling goods and services, including customer service and building relationships between sellers and end users [1]. Nowadays, the internet and its penetration play an essential role in everyday life and have created new chances for online business. Not only has the internet and network technology accelerated development, but e-commerce has become one of the leading sales channels and has grown over the past years. According to the worldwide retail e-commerce sales [2], global retail e-commerce sales is $5,211 billion in 2021. The increase in e-commerce players and the diversity of clients drive intense competition among businesses, which influences customer retention. Consequently, the demand for retaining existing clients has become one of the essential purposes of sellers on the e-commerce platform.

Customer satisfaction is one of the main marketing research parameters to understand customer behavior by evaluating a product or service based on the client's opinion [3]. It challenges every business model because the bargaining power of customers can force businesses to provide higher quality products or outstanding services that satisfy customers. This approach can enhance customers' trust and encourage customer loyalty [4]. Customer satisfaction and expectation depend on several factors, including the purchase, repeat purchase, product return, product attributes, inventory, logistics, and customer support [5]. Because of the tremendous competition in the e-commerce sector, many companies should concentrate on improving their operational performance since negative feedback could possibly impact sales and brand image. The big challenge for companies is identifying the critical factor influencing customer satisfaction leading to customer retention and enabling more productive use of organizational resources and enhanced operational performance [6].

Understanding consumers leads to competitive advantages in their business sectors. Companies can leverage insights, in particular, to predict customers' expectations more accurately by combining different types of data, such as transactional, demographic, and attitudinal data [7]. Customer satisfaction score is one of the main indicators that is used by vendors to determine customer’s impression with the online retail store. Customers give the satisfaction score based on their overall experience impression. Identifying factors contributing to poor or strong customer satisfaction scores is one of the most challenging problems that vendors are interested to investigate. Recently, machine learning (ML) has shown promising results to predict customers’ satisfying scores from massive customer’s data. Several articles utilized ML models to predict customer satisfaction scores or to determine factors that contribute to poor or strong customer’s score in many industries such as e-commerce, telecom and hotels [5], [8], [9]. Most of them yield high prediction accuracy due to its tendency to predict high customer satisfaction scores which are the majority of the scores obtained from customers. However, this approach ignores those with low or average customer satisfaction scores which could contain critical information that can improve sales performance of the vendors.

In this paper, we investigate the performance of ML to predict customers’ satisfaction score from the sales dataset collected by Olist, the Brazilian e-commerce company. Customer satisfaction score is categorized into 4 classes: Low, Average, Good and Excellent where majority of sales orders receive Excellent score. Inspired by the notion that one of the main factors that motivates customer’s purchase in e-commerce is the product rating score obtained from other customers’ purchase, we propose a feature engineering method that constructs the average and standard deviation (SD) of the product rating score which are used as the main features in the ML model. Different ML models such as Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbor (K-NN) are used to predict customers’ satisfaction score. Their performance is compared with the baseline model which predicts the customer satisfaction score using the average product rating score from other purchases.

Precision, recall and F1 score of every class are used as performance metrics of our ML model. Results show that the RF model yields the best performance with the average precision, recall, and macro F1 equal to 0.34, 0.36, and 0.32, respectively. In addition, RF achieves the best recall equal to 0.43, 0.33 and 0.33 for Low, Average and Good classes.

This paper is organized as follows. Chapter 2 discusses literature review on research related to online customer satisfaction score prediction using ML method. Methodology is explained in Chapter 3. Experimental results, discussions and conclusions are presented in Chapter 4 and 5, respectively.

1.2 Purpose of the study

In this study, the objectives are as follows:

1.2.1 To predict the customer satisfaction score of online purchasing services with e-commerce using machine learning methods.

1.2.2 To investigate the significant factors that affect customer review scores.

1.2.3 To experiment with different machine learning models to predict customer review scores and compare performance with the baseline model which using the average product rating score from other purchases.

1.2.4 To obtain customer insights data that assist in understanding consumers' demands while purchasing, increasing customer satisfaction.

1.3 Scope of the study

1.3.1 This study uses three years of Brazilian e-commerce public sales dataset by Olist from 2016 to 2018 from www.kaggle.com [10].

1.3.2 This study focuses on only successful orders that finished during 2016 and 2018.

1.3.3 New features are built to exploit the effect of the time duration in each step, such as delivery performance and a difference in timing between a predicted receiving date and an actual one.

1.3.4 Different oversampling and under-sampling techniques are investigated.

1.3.5 Customer satisfaction score is categorized into 4 classes: Low, Average, Good and Excellent where majority of sales orders receive Excellent score.

1.3.6 Unnecessary features are filtered using a Tree based feature importance or Mean Decrease in Impurity (MDI)

1.3.7 We evaluate the performance of our prediction model using a precision, recall and F1 score.

มหาวิทยาลัยศรีนครินทรวิโรฒ

Chapter 2
Literature review

This chapter includes the reviews of relevant literature on the topics as follows:

1. E-commerce
2. Customer satisfaction
3. Machine learning
4. Classification algorithms
5. Imbalanced dataset
6. Model evaluation

The review of research on customer satisfaction by using machine learning models and evaluation descriptions are given in Table 1,2 and 3, respectively.

Table 1 Description of Models

No.	Model	Description
1.	DT	Decision Tree
2.	RS	Random Forest
3.	NN	Neutral Networks
4.	SVM	Support Vector Machine
5.	NB	Naïve Bayes
6.	Ap	Apiorir algorithm
7.	XgB	XGBoost
8.	LR	Logistic Regression
9.	RF	Random Forest

Table 2 (Continued)

10.	K-NN	K-Nearest Neighbors
11.	LSTM	Long Short-Term Memory
12.	REPTree	Reduced Error Pruning Tree

Table 3 Description of Evaluation

No.	Evaluation	Description
1.	Ac	Accuracy Score
2.	AUC	Area under the Curve
3.	Re	Recall Score
4.	Pc	Precision Score
5.	F1	F1 Score
6.	ROC	Receiver Operating Characteristic
7.	Ss	Sensitivity
8.	Sc	Specificity

2.1 E-commerce

Electronic commerce, often known as e-commerce, uses digital information processing and electronic communications in commercial transactions to build, alter, and redefine value-generating interactions between people and organizations. There are six fundamental types of e-commerce [10] which can be schematized as in Figure 1:

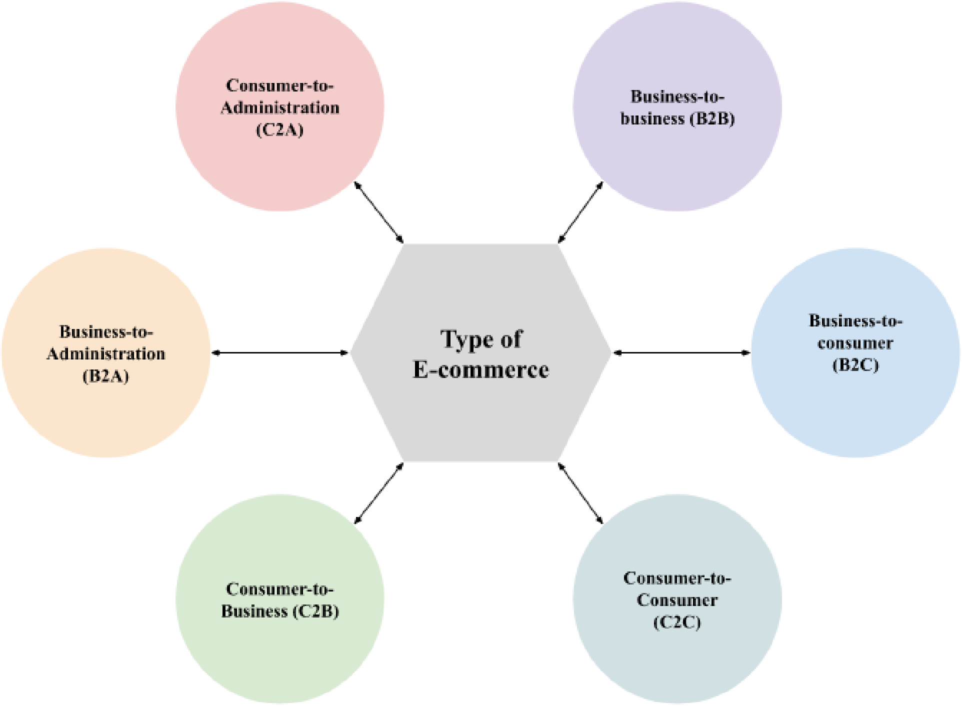

Consumer-to-Administration (C2A)
Business-to-business (B2B)
Business-to-Administration (B2A)
Type of E-commerce
Business-to-consumer (B2C)
Consumer-to-Business (C2B)
Consumer-to-Consumer (C2C)

Figure 1 Types of e-commerce [10]

- Business-to-Business (B2B): All business product or service transactions are considered B2B. This method is commonly used for online trade by manufacturers and traditional industrial wholesale businesses.
- Business-to-Consumer (B2C): business-to-consumer collaboration between the firms and the ending client company. Conventional retail transactions frequently undertake in the e-commerce shopping industry. Due to the rise of the internet, this industry has dramatically increased, with many online stores and businesses providing clients with a wide range of goods.
- Consumer-to-consumer (C2C): C2C e-commerce comprises any transactions involving the exchange of products or services between customers. This trade is often handled by a third party that provides an online transaction convention.

- Consumer-to-business (C2B): The typical trade situation of products invert in a C2B transaction. Companies that rely on crowdsourcing-based use this model frequently. The person also offers their services or products to businesses targeting particular categories of goods or services.
- Business-to-administration (B2A): The B2A segment includes all online transactions between the government and corporations. It comprises many programs, particularly taxes, social care, healthcare, legal, and documentation. Spending on e-government has dramatically expanded these modalities of delivery in recent years. By investing in e-government, these service delivery channels significantly grow.
- Consumer-to-administration (C2A): Every digital transaction between people and governments covers under the C2A concept. The application of C2A includes distance learning, information distribution of social security, filing tax returns and payments, information about illnesses, and the cost of health services.

Companies can rapidly expand their market share by utilizing the internet as a communication tool and distribution channel for marketing and selling products and services. Consequently, e-commerce is a growing trend that offers several advantages and is part of the engine of global economic growth. The following are the advantages of e-commerce for sellers and customers [10] , as shown in Table 4 and 5:

Table 4 Benefits of e-commerce for sellers and customers

No.	Benefits for sellers	Benefits for customers
1.	Expand to a new market and build revenue	Possible to purchase 24/7 daily

Table 5 (Continued)

2.	Reduce related costs such as operations and transportation	Saving time when buying online products
3.	Accelerate on selling products	Cheaper products and services compared with offline stores
4.	Easy to promote	Easy to search
5.	Make a stronger relationship with customers	Can buy products from aboard
6.		Able to look through existing customer reviews before making a decision

At present, Consumers globally can access the internet, which led to an annual increase in the number of online shoppers worldwide due to the advancement of mobile technologies that enable e-commerce to be more accessible and productive. According to the worldwide retail e-commerce sales [2], global retail e-commerce sales is $5,211 billion by 2021. It expects to expand by 50 percent over the next four years, reaching about $7,528 billion by 2025, as shown in Figure 2.

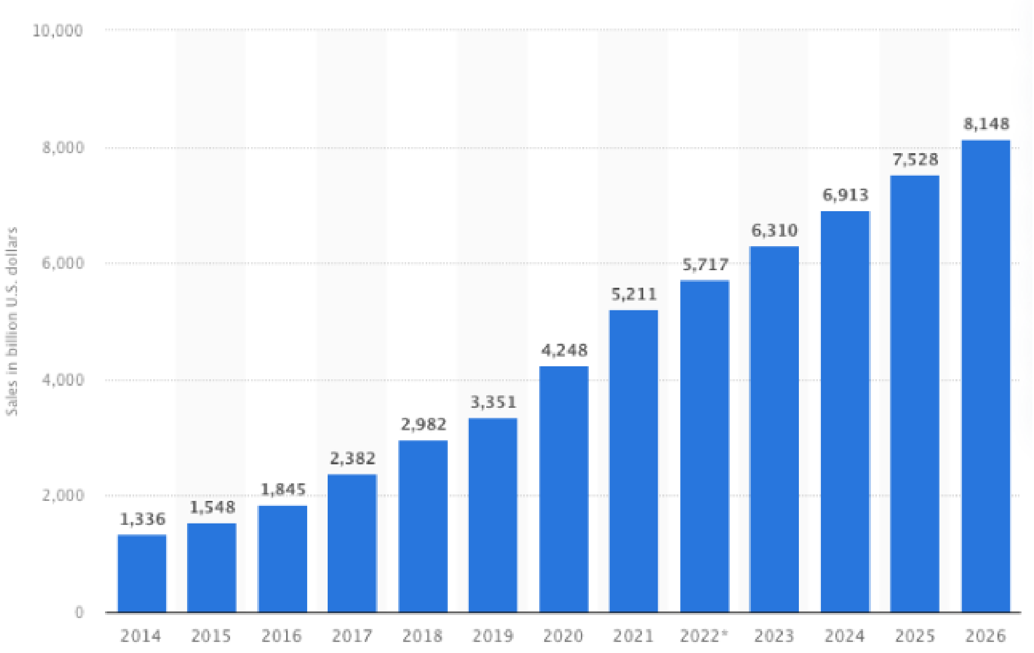


Figure 2 Retail e-commerce sales worldwide from 2014 to 2026 [2]

2.2 Customer Satisfaction

Customer satisfaction is the appraisal of products or services based on an individual's attitude or expectations for products or services before and after consumption. According to PwC's global consumer insights survey in 2022 [11], 73% of buyers feel that customer experience is a crucial factor in their purchasing decision. Therefore, the main point of differentiation for a brand is the customer experience. Businesses must enhance their customer service to provide a positive customer experience that will drive sales and loyalty.

The critical result of Lee's customer satisfaction model is a positive association between customer satisfaction and client retention. It integrates these several attributes of online shopping, implying that repurchasing is driven by customer satisfaction. Customer satisfaction levels in online businesses are affected by logistic support, customer service, pricing attractiveness, and website storefront [12], as shown in Figure 3.

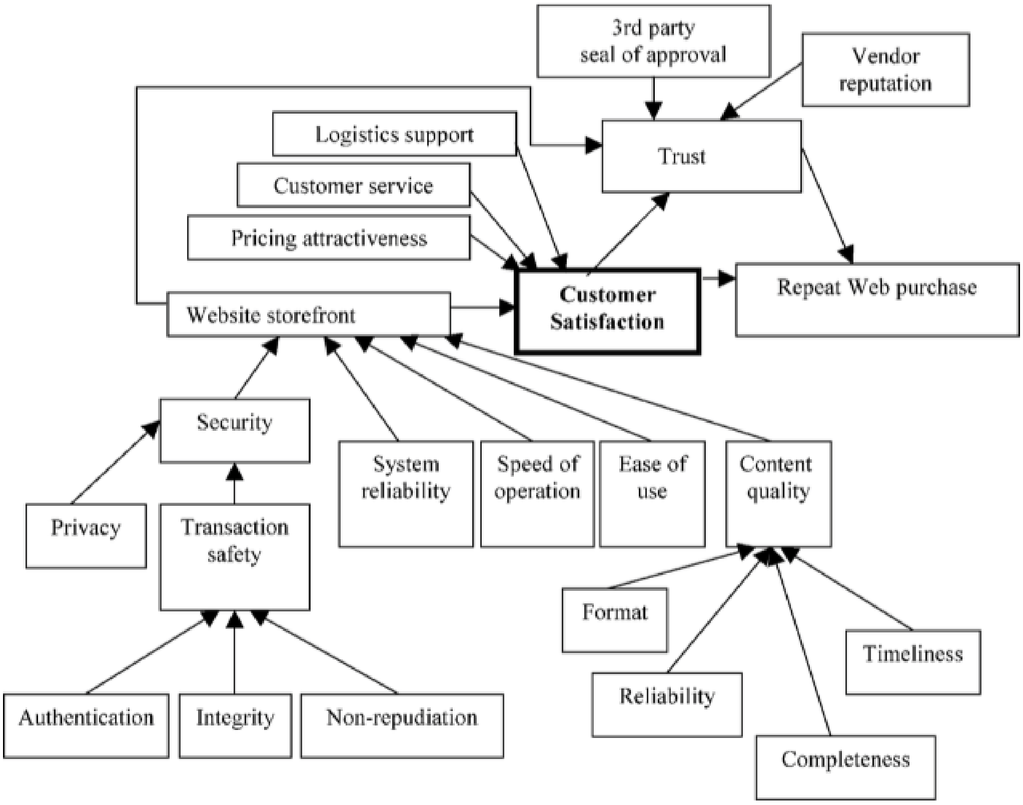


Figure 3 Lee’s model of internet consumer satisfaction [12]

Similarly, [13] said that various factors had an impact on customer satisfaction, including

- **Product quality:** Product ratings use to determine a product's quality baseline. The evaluation of a product's quality may summarize as determining whether or not it meets the consumer's demands and if it is up to the client's desired standard.
- **Service quality:** In e-commerce may be described as the total consumer assessment of the quality of online service delivery. Service quality is a critical factor in influencing online shoppers' purchasing intentions.
- **Product delivery:** Fulfillment dependability refers to the timely delivery of the correct items. Positive service evaluations and the customer's sentiment are two outcomes of delivery correctly.

- **Information Quality:** The relevance, correctness, comprehensiveness, and understandability of information offered by websites refer to information quality. Because information is the primary offering of e-retailer websites, it plays an essential role in satisfying consumers' visits.
- **Price:** Pricing perceptions directly impact customer satisfaction, but price fairness has an indirect impact. Then, buyers compare their expectations to the traditional purchasing scene.

2.3 Machine Learning

Machine Learning is an essential component of the expanding field of data science. Algorithms train to produce classifications or predictions and to reveal valuable insights in data mining projects using statistical methodologies. The algorithm's learning mechanism of machine learning can be divided into three major parts [14].

- **Decision process:** A prediction or classification performed using machine learning algorithms. The program will estimate a pattern in the input data, which may be labeled or unlabeled.
- **Error function:** An error function evaluates the model's prediction. If there are known instances, an error function can compare them to determine the model's correctness.
- **Model optimization process:** The weights are modified to decrease the difference between the known example and the model prediction If the model fits the data points in the training set better. The algorithm will repeat this "evaluate and optimize" procedure, automatically updating weights until an accuracy criterion is reached.

Machine learning models are divided into four types as follows [14]:

2.3.1 Supervised Machine Learning

The use of labeled datasets to train algorithms to categorize data or predict outcomes reliably. As input data supply into the model, the model changes its weights until it is well fitted. Then, as part of the cross-validation procedure to prevent the model from being overfitted or underfit. The techniques used in supervised learning include Neural Networks, Naïve Bayes, Linear Regression, Logistic Regression, Random Forest, and Support Vector Machines.

2.3.2 Unsupervised Machine Learning

Unsupervised learning analyzes and clusters unlabeled datasets. These algorithms identify hidden patterns or data clusters without the interference of a human. This method is beneficial for exploratory data analysis, cross-selling strategies, consumer segmentation, and image and pattern identification since it can find similarities and contrasts in information. Additionally, dimensionality reduction decreases the number of features in a model. Two standard techniques are singular value decomposition (SVD) and principal component analysis (PCA).

The techniques used in unsupervised learning include Neural networks, K-Means clustering, and probabilistic clustering.

2.3.3 Semi-Supervised Machine Learning

The semi-supervised method uses to train the dataset with labeled and unlabeled data. It is beneficial when extracting pertinent features from data and when the volume of data is enormous. It guides classification and features extraction during training from a more extensive, unlabeled data set using a smaller labeled data set. Moreover, it can solve the issue of insufficient labeled data for supervised learning algorithms. Semi-supervised learning is a suitable method for medical image analysis because a minimal quantity of training data may significantly increase accuracy.

2.3.4 Reinforcement Machine Learning

Even though algorithms do not train on sample data, reinforcement machine learning is a machine learning approach comparable to supervised learning. This model learns through trial and error. The sequence of successful results will reinforce the establishment of the optimal recommendation or policy for a specific situation.

2.4 Classification algorithms

Classification is a supervised learning technique applied to structured and unstructured data. It divides the data into classes to identify the class of new observations based on training data and determine the possibility that new data will belong to one of the previously categorized classes. A classifier is an algorithm that performs classification on a dataset. Classifications can be divided into two types [15]:

- Binary Classifier: It's used when the classification task has just two potential outcomes, such as yes or no, male or female, spam or not spam, cat or dog.
- Multi-class Classifier: It uses when a classification task includes more than two outcomes, such as classifications of types of crops and types of music.

2.4.1 Random Forest (RF)

A supervised machine learning algorithm called random forest comprises several independent decision trees that work as an ensemble. This algorithm has three significant hyperparameters that must be set prior to training. These variables include node size, number of trees, and characteristics sampled. The random forest classifier uses to address regression or classification issues. The class with the most significant votes is selected as the prediction by model from among the individual class predictions produced by each tree in the random forest [16].

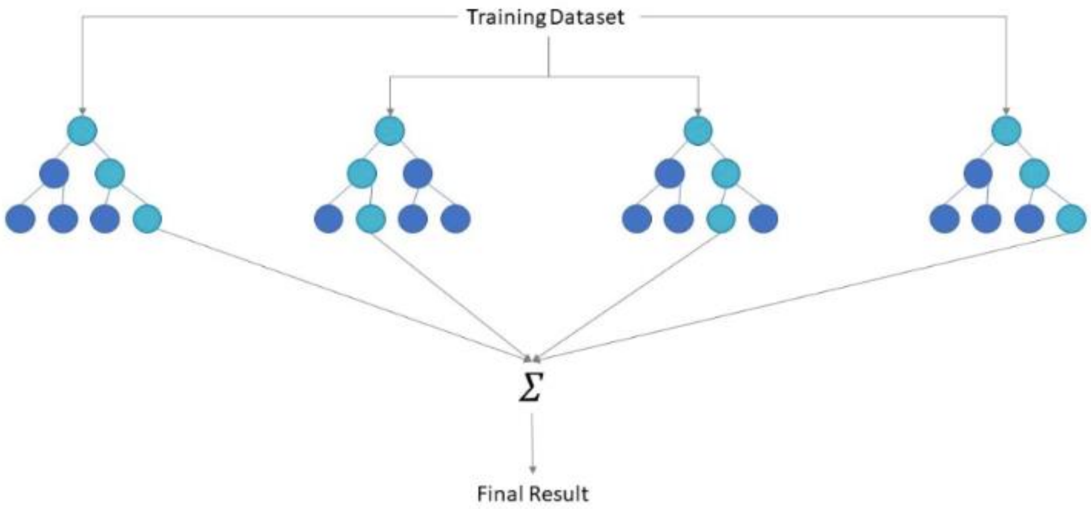


Figure 4 RF classifier diagram [16]

2.4.2 K-Nearest Neighbors (K-NN)

K-Nearest Neighbors algorithm is a non-parametric, supervised learning classifier that employs closeness to create classifications or predictions about an individual data point grouping. Even though it applies to classification or regression issues, it is a classification method assuming that similar points can discover close to one another [17]. In Figure 5, category A and category B exist and have a new data point; hence this data point will fall into one of these categories. The K-NN algorithm solves this problem and can determine the category or class of a dataset [18].

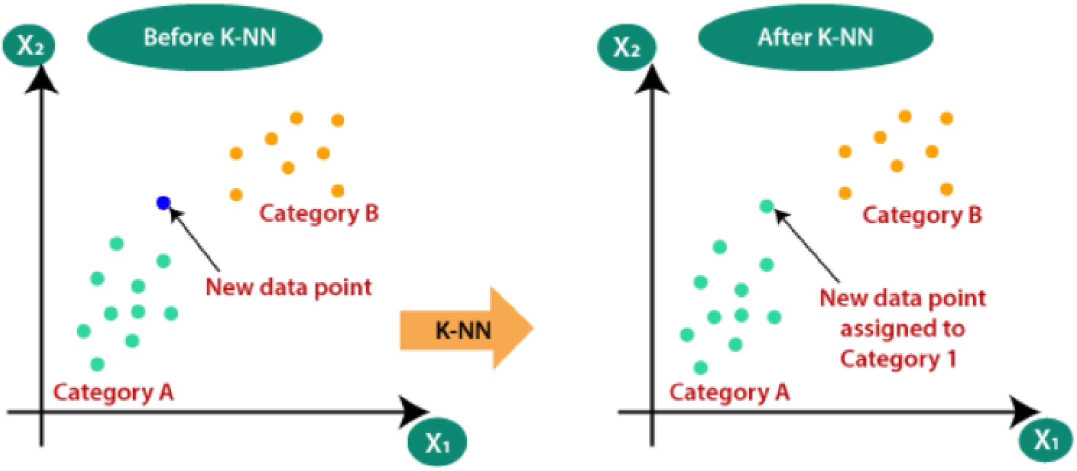


Figure 5 K-NN diagram [18]

2.4.3 Logistic Regression (LR)

Logistic regression is a popular Machine Learning algorithm that belongs to the Supervised Learning technique. It is used to predict the categorical dependent variable from a set of independent variables. Logistic regression is used for solving the classification problems. Logistic regression predicts the outcome of a categorical dependent variable. As a result, the outcome must be a categorical or discrete value. It can be Yes or No, 0 or 1, true or False, etc., but instead of giving the exact values as 0 and 1, it gives the probabilistic values that fall between 0 and 1.

Logistic Regression is an important machine learning algorithm because it can provide probabilities and classify new data using both continuous and discrete datasets. Logistic Regression can be used to classify observations using various types of data and can easily determine which variables are most effective for classification. The logistic regression value must be between 0 and 1, and it cannot exceed this limit, forming a curve similar to the "S" form. The Sigmoid function or logistic function is another name for the S-form curve [19]. The logistic function is illustrated in Figure 6 below:

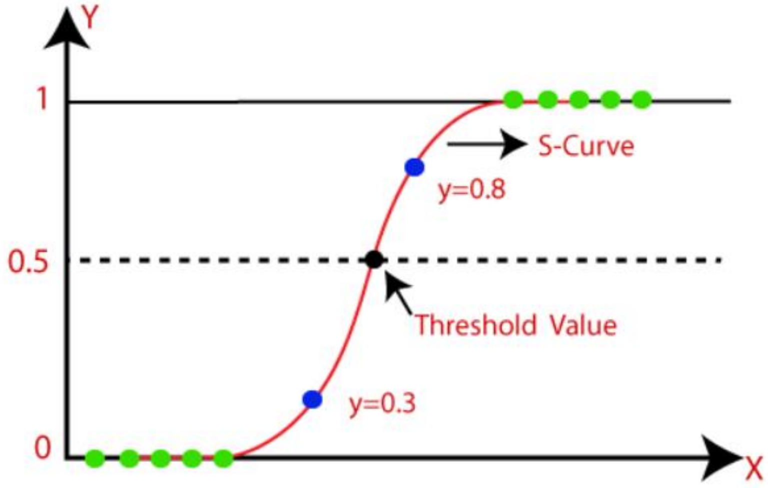


Figure 6 Logistic Function [19]

2.5 Imbalanced Dataset

The term "imbalanced data" describes datasets where the target class has an unequal distribution of observations that the one class label has a very high number of observations. In contrast, the other has a very low number of observations.

For example case, a bank offers its clients credit cards. The bank is now concerned that some fraudulent transactions are occurring, and when they checked their data, they discovered that only 30 cases of fraud were detected for every 2000 transactions. So, the amount of fraudulent transactions per 100 transactions is less than 2% or more than 98% of transactions are "No Fraud." The class "No Fraud" is referred to as the majority class, while the considerably smaller "Fraud" class is referred to as the minority class [20].

2.5.1 Approach to deal with the imbalanced dataset problem

It is essential to identify the minority class appropriately. As a result, the model should not be biased toward detecting the majority class but should provide equal weight or importance to the minority class [20]. Here in Figure 8 is an example of the mechanic for both techniques of imbalance data treatment that generally used which are oversampling and undersampling.

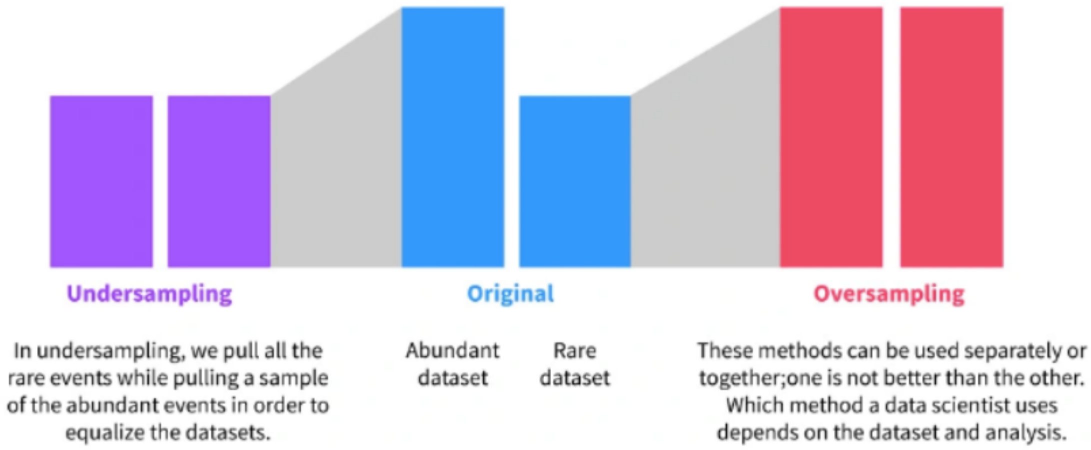

Undersampling
Original
Oversampling
In undersampling, we pull all the rare events while pulling a sample of the abundant events in order to equalize the datasets.
Abundant dataset
Rare dataset
These methods can be used separately or together;one is not better than the other. Which method a data scientist uses depends on the dataset and analysis.

Figure 7 Oversampling and undersampling diagram [20]

2.5.1.1 Oversampling

This method is used to sample the minority or majority class. When we have an unbalanced dataset, we use replacement to oversample the minority class. This method is known as oversampling. After sampling the data, we obtained a balanced dataset for both the majority and minority classes. Therefore, if both classes have similar entries in the dataset, we can expect that the classifier will assign equal weight to both classes [20].

2.5.1.1.1 SMOTE (Synthetic Minority Over-sampling Technique)

Another method for oversampling the minority class is SMOTE. Adding duplicate minority class entries to the model often doesn't provide new data. SMOTE creates new instances by synthesizing the data already available. Simply put, SMOTE searches for examples of the minority class, uses k nearest neighbors to choose a random nearest neighbor, and generates a synthetic instance randomly in feature space [20].

2.5.1.2 Undersampling

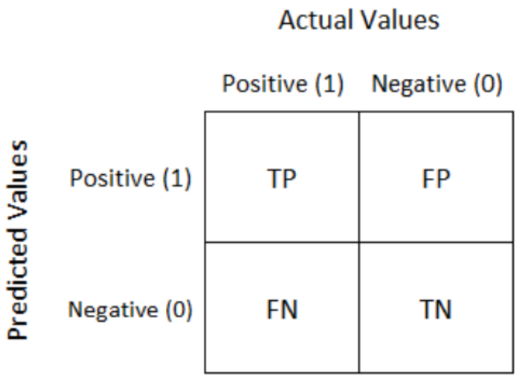
Undersampling is the process of randomly deleting rows from the majority class in order to match them with the minority class. After sampling the data, we obtained a balanced dataset for both the majority and minority classes. Therefore, if both classes have similar entries in the dataset, we can expect that the classifier will assign equal weight to both classes [20].

2.6 Model evaluation

Model evaluation analyzes a machine learning model's performance, strengths, and limitations using various evaluation criteria. During the initial stages of research, it is crucial to evaluate models to determine their efficacy [21].

2.6.1 Confusion matrix

Correct and incorrect classifications for each class provide a more thorough description in the confusion table. A confusion matrix can help grasp the difference between classes, mainly if the value of misclassification differs or has significantly more test data on one class than the other [22]. When making classification predictions, there are four possible outcomes as follows [23]



Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Figure 8 Confusion matrix diagram [23]

- **True Positives (TP):** predicting that an observation belongs to a class and belongs to that class.
- **True Negative (TN):** predicting that observation does not belong to a class and it truly does not belong to that class.
- **False Positives (FP):** predicting that an observation belongs to a class whereas it does not.
- **False Negatives (FN):** predicting that observation does not belong to a class but does.

These four outcomes frequently represent a confusion matrix. For example, the following confusion matrix is an example of binary classification as shown in Figure 8. The matrix is obtained after making predictions on test data and identifying each prediction as one of the four probable outcomes above.

2.6.1.1 Accuracy

The proportion of accurate predictions made using the test data is known as accuracy. It readily calculates by dividing the number of correct predictions by the total number of predicts [22]. Accuracy was calculated using (1).

$$\text{accuracy} = \frac{\text{correct predictions}}{\text{all predictions}} \tag{1}$$

2.6.1.2 Recall

The model's capacity to identify positive samples is measured by recall. It is computed as the proportion of positive samples classified correctly as positive to all positive samples. More positive samples are found when recall is higher [22]. Recall was calculated using (2).

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \tag{2}$$

2.6.1.3 Precision

Precision defines as the proportion of examples that are genuinely positive and relevant among all of the examples that project to belong in a certain class. Precision allows us to visualize the machine learning model's dependability in classifying the model as positive [22] as shown in (3).

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{3}$$

2.6.1.4 F1 Score

Measurement that computes the score by considering both recall and accuracy. The F1 score may be considered a weighted average of the accuracy and recall values, with the greatest and worst values occurring at 1 and 0, respectively [22] as shown in (4).

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}. \quad (4)$$

2.7 Literature review on customer satisfaction using machine learning

Research article 1: Optimising e-commerce customer satisfaction with machine learning [5]

This article predicts the key drivers that influence satisfaction. Four classification machine learning algorithms, DT, RF, ANN, and SVM, are selected to classify customer satisfaction based on 3-year data with 112,00 orders from a Brazilian e-commerce retailer, and performance is compared using the accuracy, sensitivity, specificity, F1-score, and computation time. The results indicate that RF achieves the best result with the highest accuracy and reasonable processing time. In addition, the outcome yields that; the estimated delivery date and the number of days to deliver an order are the top two important factors affecting customer satisfaction.

Research article 2: An advanced intelligence system in customer online shopping behavior and satisfaction analysis [24]

In this study, researchers collected 40,000 data from customer reviews, online surveys, and customer feedback on online shopping sites using Google Sheets which most of the people are using daraz.com, bikroy.com, rokomari.com, and amazon.com. They used machine learning methods to determine their work's accuracy and analyze customer online shopping satisfaction. NB, Apiorir algorithm, DT, and RF classification algorithms are used for this analysis. The best result showed 88% accuracy using the NB algorithm and 87% accuracy using the Apiorir algorithm as shown the overview in Table 6.

Table 6 Algorithms and results obtained by research article 2

Algorithm	Accuracy
NB	88.44%
Apiorir algorithm	87.89%
DT	82.69%
RF	84.25%

Research article 3: A consumer behavior prediction method for e-commerce application [25]

This study uses a hybrid classification method that considers K-NN and DT based on prediction analysis. Customer behavior predicts by using 1963 reviews of consumers of the Amazon e-commerce from the UCI collection, which includes evaluations of a range of Amazon products. In an earlier study, NB was used for consumer behavior analysis, but the accuracy was poor. This work proposes a hybrid K-NN and DT classification to improve accuracy. Accuracy, precision, recall, F1 score, and execution time are the performance indicators used in this test. There are three classes for classification that are positive (-1), neutral (0), and negative (+1) compared to NB. Results showed that the hybrid classification has higher accuracy by accuracy of the Naïve Bayes is 74.11%. In contrast, the accuracy of the hybrid classifier is 90.75%. Moreover, the computing time of the hybrid classification is faster than the NB as shown the overview in Table 7.

Table 7 Algorithms and results obtained by research article 3

Algorithm	Classes	Precision	Recall	F1-score	Accuracy	Execution time (Second)
Hybrid	-1	0.61	0.99	0.76	0.907	0.7
	0	0.00	0.00	0.00		
	1	0.97	0.35	0.51		
NB	-1	0.72	0.81	0.76	0.741	1
	0	0.00	0.00	0.00		
	1	0.76	0.69	0.73		

Research article 4: Analytics in support of e-commerce systems using machine learning [26]

This study intends to use machine learning to assess the sentiments from online consumer reviews on an e-commerce platform. However, reviews are text data and must be prepared in a format machine learning can comprehend. Therefore, the e-commerce dataset used National Language Processing (NLP) approaches to prepare data for the study. Then, the dataset is evaluated using NB, LR, SVM, and NN classifiers to execute sentiment analysis. Data is split into 80/20 train and test datasets. The six classifiers' performances are compared regarding ROC curves, accuracy, precision, recall, F1 score, and execution time. Finally, the performance of LR outperforms other classifiers for predicting customers' sentiments from their text reviews as shown the overview in Table 8.

Table 8 Algorithms and results obtained by research article 4

Algorithm	Precision	Recall	F1-score	Accuracy
LR	0.88	0.79	0.83	0.91
NB	0.91	0.51	0.48	0.83
SVM	0.85	0.76	0.80	0.90
NN	0.80	0.78	0.79	0.88

Research article 5: Machine learning algorithms to empower Indian women entrepreneur in e-commerce clothing [27]

This article investigated 23,486 customer reviews with 11 attributes along with class attributes from Indian women's clothes on the e-commerce site to understand the thoughts of the needy customers. Machine learning algorithms such as Reduced Error Pruning (REPTree) and SVM are used to classify customer reviews and compare their accuracy, precision, and recall performance. Results show that REPTree has an accuracy of 91.43%. Furthermore, REPTree has accuracy and recall scores of 93.75% and 96.04%, respectively. Which shown the overview in Table 9.

Table 9 Algorithms and results obtained by research article 5

Algorithm	Accuracy	Precision	Recall
REPTree	91.43	93.75	96.04
SVM	85.94	91.01	92.29

Research article 6: Forecast the rating of online products from customer text review based on machine learning algorithms [8]

This study aims to utilize machine learning to predict product ratings from online customer text reviews on products comprising 71,045 reviews that contain 25 columns. Dataset, named “GrammarandProductReviews” is provided by Datafiniti. The review includes opinions on the product and a rating of 1 to 5. The dataset is then analyzed using RF, LR, and XgB. Accuracy, precision, recall, and F1 score are used to compare and evaluate the three classifiers' performance. The RF method then demonstrated the best performance compared to others, achieving an accuracy of 94%, precision, recall, and f1-scores of 0.94, 0.94, and 0.94 respectively. As shown the overview in Table 10.

Table 10 Algorithms and results obtained by research article 6

Algorithm	Accuracy (%)	Precision	Recall	F1-score
RF	94.2	0.94	0.94	0.94
LR	92.1	0.92	0.92	0.92
XgB	84.7	0.85	0.85	0.85

Research article 7: Performance analysis of different machine learning in customer prediction [28]

Based on the performance analysis of customer prediction, this study applies machine learning that takes into consideration K-NN, LR, and NB. The Australian credit customer dataset used in this study, obtained from the UCI repository, shows whether a consumer has good or bad credit. Additionally, three models' performances are shown utilizing a variety of metrics, including precision, sensitivity, accuracy, and TP rates. The experiment's results suggest that NB archives a more significant prediction of 84.7826 than LR and K-NN. Further, compared to LR and NB, K-NN has greater specificity of 0.8903 and higher accuracy of 0.849. As shown the overview in Table 11.

Table 11 Algorithms and results obtained by research article 7

Algorithm	Accuracy (%)	Precision	Sensitivity	Specificity	TP rate	FP rate
NB	84.7826	0.841	0.811	0.8772	0.811	0.123
LR	83.913	0.814	0.827	0..8485	0.827	0.151
K-NN	83.6232	0.849	0.769	0.8903	0.769	0.110

Research article 8: Prediction of customer propensity based on machine learning [29]

This study's objective is to utilize machine learning to predict customer propensity to buy the product from web browsing behavior data of Kaggle competitions comprising 455,401 samples and 24 features. The dataset is analyzed using RF and LR to solve the problem. During developing the model using LR, the researcher used two regularization approaches, L1 and L2, to generate the model using LR.

The hyperparameter setting of L1 and L2 was listed as follows:

1. L1 Logistic Regression
    a. Penalty: l1
    b. tol : 0.001
    c. C: 13.0
    d. Max_iter: 1000
    e. Multi_class: auto
    f. Verbose: 1
    g. N_jobs: -1
2. L2 Logistic Regression
    a. Penalty: l2
    b. Tol : 0.001
    c. C: 13.0
    d. Random_state: 34
    e. Max_iter: 100
    f. Multi_class: auto
    g. Verbose: 1
    h. N_jobs: -1

These two approaches for determining penalty terms avoid overfitting by adding parameters. The L1 model is equivalent to introducing a Laplacian before the model's parameters from a Bayesian perspective, and the L2 model is equal to introducing a Gaussian prior. Comparing the results of models L1, L2, and RF used accuracy, precision, recall, F1 score, and ROC curves. It showed that L1 and L2's prediction effects are similar. The accuracy of LR and RF predictions is the same, according to comparisons of their predictive effects. Additionally, while recall, F1 score, and ROC of LR is better compared to RF, the accuracy of RF is higher compared to LR.

มหาวิทยาลัยศรีนครินทรวิโรฒ

Both algorithms effectively predicted outcomes (the accuracy rates are above 90%), so they can accurately predict whether consumers would make purchases based on their online browsing behaviors. As shown the overview in Table 12.

Table 12 Algorithms and results obtained by research article 8

Algorithm	Accuracy	Precision	Recall	F1-score	ROC
RF	0.993	0.874	0.969	0.929	0.981
L1	0.993	0.872	0.987	0.926	0.990
L2	0.993	0.872	0.987	0.926	0.990

Research article 9: A XGBoost method based on telecom customer satisfaction enhancement strategy [30]

The literature uses the XgB machine learning method to help operators predict the satisfaction score to increase customer satisfaction to discover the relationship between the customer satisfaction index and network experience. This paper uses customer surveys of over 600 thousand records and links each record to a primary service base station according to customers' resident locations. Then, XgB evaluated the accuracy of satisfaction prediction with the LR model to validate the model's performance. The results demonstrate that XgB outperforms the standard LR model while using parameters configuration as follows:

1. n_estimators: 100
2. Max_depth: 6
3. Learning_rate: 0.3
4. Gamma: 5
5. alpha : 1

6. Lambda: random
7. Subsample: 0.8

Finally, the accuracy of prediction of the XgB model and LR model is 94.97% and 75.36%, respectively.

Research article 10: A machine learning approach for opinion mining online customer reviews [9]

This study aims to apply machine learning to predict online customer reviews from a hotel website “Agoda.com” in Vietnam 39,976 reviews containing 25 columns. At the labeling step, the dataset is given a rating score and classified review score on two levels; less than 7.0 are negative, and more significant than 7.0 are positive. Next step, The dataset is then analyzed using six algorithms; RF, LR, NB, SVM, DT, and NN. Finally, the model's performance is evaluated based on accuracy, precision, recall, and F1 score to compare and consider. The training results demonstrate that LR is the most effective model among those trained, proving that LR is the model that works best with the training set of data. As shown the overview in Table 13 and 14.

Table 13 Algorithms and results obtained by research article 10

Algorithm	Accuracy	Precision	Recall	F1-score
NB	0.50	0.70	0.50	0.50
SVM	0.79	0.79	0.80	0.79
LR	0.81	0.80	0.81	0.79
NN	0.79	0.79	0.79	0.77

Table 14 (Continued)

DT	0.74	0.74	0.75	0.74
RF	0.69	0.48	0.69	0.57

In addition, Table 15 and 16. summarizes all the literature and includes information of the data, algorithms, and results of the best performance model from each research.

According to most previous researches have used various procedures to predict customer behavior or satisfaction, they have only tested their algorithms on a small scale. They classified products using a limited number of categories. They have previously used a small number of dependent variables. As a result, more variables must be investigated in order to classify the features. Also most studies have also failed to connect the purchasing time with product categories or even shipping duration, which should be considered as one of the most important factors influencing customer satisfaction scores.

Our study intends to test the effect of any feature on customer review scores, particularly timing duration, which is one of the most important features that customers will give to sellers when they receive poor service from deliverers. For added complexity, our target classes review score will be divided into four groups to assist sellers in estimating the satisfaction score that they will receive from customers after the purchasing process is completed.

Table 15 Data, algorithms, and results from research articles.

No.	Citation of machine learning used in customer satisfaction prediction	Data information	Model	Evaluation	Best performance model
1.	[5]	3-year data from the "Brazilian E-Commerce Public Dataset by Olist" with 112,000 orders	DT, RF, ANN, and SVM	Ac, Ss, Sc, F1, and computation time	- RF (Ac= 87.00%) - The outcome provided insights; the estimated delivery date and the number of days to deliver an order are the top two important factors affecting customer satisfaction.
2.	[24]	40,000 data from customer reviews, online surveys, and customer feedback on online shopping sites	NB, Ap, DT, and RS	Ac	- Ap (Ac= 88.00%)
3.	[25]	1,963 reviews of consumers of the Amazon e-commerce from the UCI collection	hybrid classification of K-NN and DT, and NB	Ac, Re, Pc, F1 and computation time	- Hybrid classification (Ac= 90.75%)
4.	[26]	Online consumer reviews on an e-commerce platform data	NB, LR, SVM, and NN	Ac, Re, Pc, F1, ROC curve and computation time	- LR (Ac= 91.00%)

Table 16 (Continued)

5.	[27]	23,486 customer reviews with 11 attributes from Indian women's clothes on the e-commerce site	SVM and REPTree	Ac, Re, and Pc	- REPTree (Ac= 91.43%)
6.	[8]	71,045 reviews of the product dataset named “GrammarandProductReviews” provided by Datafiniti	XgB, RF, and LR	Ac, Re, Pc, and F1	- RF (Ac= 94.20%)
7.	[28]	The Australian credit customer dataset from the UCI collection	LR, K-NN, and NB	Ac, Pc, Ss, and TP rates	- NB (Ac= 84.78%)
8.	[29]	Web browsing behavior data of Kaggle competitions comprising 455,401 samples and 24 features	LR and RF	Ac, Re, Pc, F1, and ROC curve	- Both algorithms (Ac= 99.30%)
9.	[30]	Customer surveys over 600 thousand records and links each record to a primary service base station according to customers' resident locations	XgB and LR	Ac	- XgB (Ac= 94.97%)
10.	[9]	Online customer reviews from a hotel website in Vietnam, Agoda.com 39,976 reviews containing 25 columns	RF, LR, NB, SVM, DT, and NN	Ac, Pc, Re, and F1	- LR (Ac= 81.00%)

Chapter 3
Methodology

This chapter explains the methodology and tools for data analysis used in this study on the following topics:

1. Dataset and Data Understanding
2. Data Cleansing
3. Exploratory Data Analysis (EDA) and Visualization
4. Split Data
5. Feature Engineering
6. Feature Selection
7. Modeling
8. Model and Evaluation

This study aims to predict e-commerce customer satisfaction and identify features affecting satisfaction scores. The whole process of implementation starts with data understanding to overview the dataset and find some insight. Then follow up by data cleansing and EDA before splitting data for further modeling.

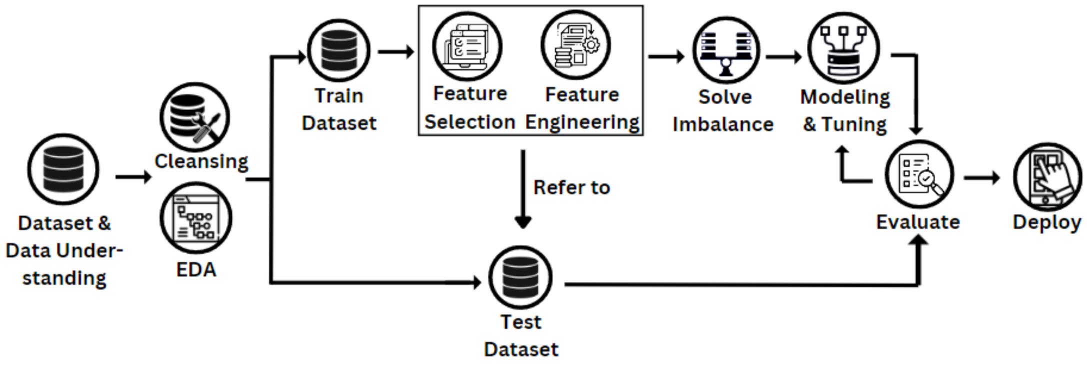


Figure 9 The prediction process schematic

The next step is implementing the algorithm, building models, and measuring its performance. The training dataset and test dataset obtained in the previous step are used in the step of cleansing, feature engineering, and feature selection separately. For the reason of prevention of data leakage in the feature engineering step. Results from feature selection are also used to select features in both train and test sets. Then, evaluate, and improve the imbalanced data and tune hyperparameters in each algorithm to optimize the model performance and evaluate the result compared with the baseline method which predicts the score using the average product rating obtained from other customer's purchases and iterate the process to modeling and tuning again after got a low performance otherwise the model will be applied to e-commerce business implementation or deployment. The prediction process was summarized in Figure 9.

3.1 Dataset and Data Understanding

Datasets are collected by the Brazilian e-commerce company Olist Shops, the biggest department store in Brazilian markets owns the sales historical dataset from 2016 to 2018 which has been used for this research and provided to www.kaggle.com [10].

The dataset contains approximately 112,000 orders from 2016 to 2018 in 34 columns and 102,710 rows. In addition, it includes 8 tables of data on the order status, price, payment, freight performance to customer location, product attributes, and reviews written by customers as follows:

- olist_order_payments_dataset
- olist_products_dataset
- Product_category_name_translation
- olist_order_reviews_dataset
- olist_orders_dataset
- olist_order_items_dataset
- olist_sellers_dataset

- olist__order_customers_dataset
- olist_geolocation_dataset

The data is separated into different datasets to facilitate comprehension, as illustrated in Figure 10, which explains and shows the relationship between each table. This study, geolocation table is not used in the calculation since it contains with latitude-longitude of the state in Brazil which we do not concentrate on in this study.

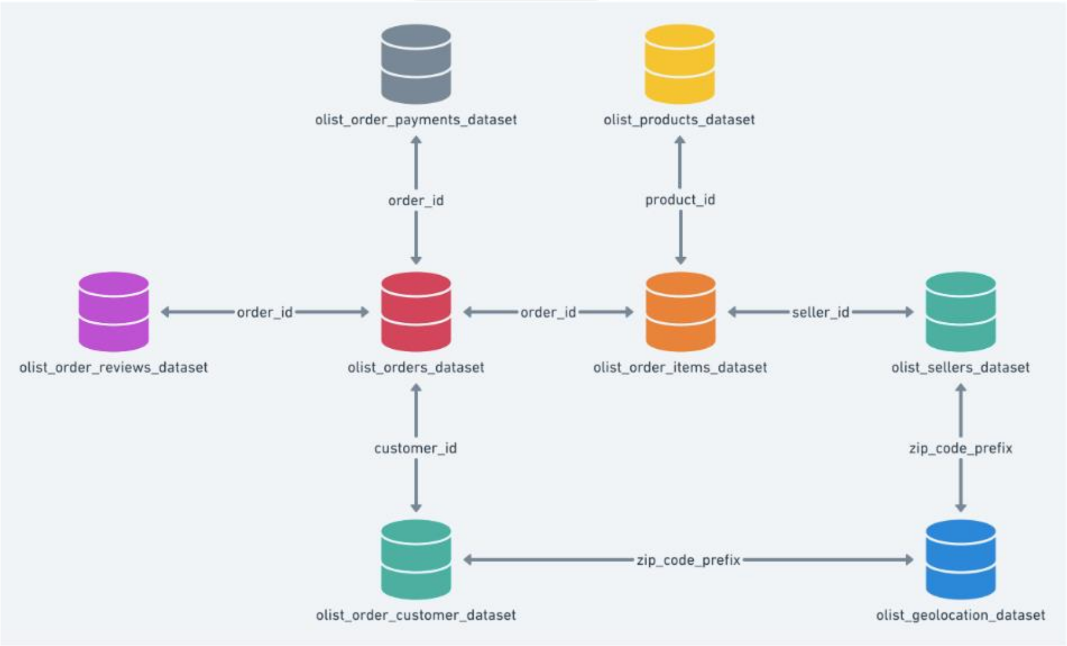


Figure 10 Data schematic of Brazilian e-commerce public by Olist shop [31]

3.2 Data Cleansing

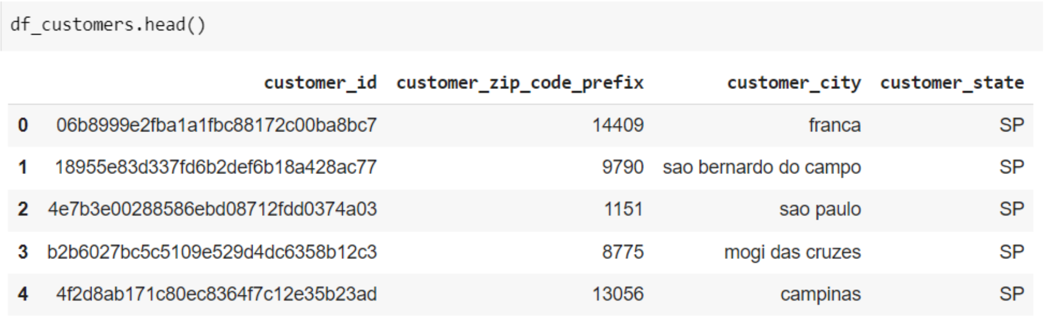
The dataset would be cleaned when there are contain dirty and missing records in every table that may possibly find. Our assumption is to concentrate on only completed data and will drop the others. Then, the table has been cleaned by a specific method from each table before merging to prevent duplicate data as following details.

Customers Table

Table 17 Description of the customer column name

No.	Variable	Description
1.	customer_id	Key to the orders dataset. Each order has a unique customer_id
2.	customer_unique_id	Unique identifier of the customer
3.	customer_zip_code_prefix	First five digits of the customer's zip code
4.	customer_city	Customer city name
5.	customer_state	Customer state

Table 17 shows customers' contact information and location data, including zip code, city, and state. The customer_id used to place each order. Therefore, it implies that it will receive various IDs for various orders, in contrast to customer_unique_id, which identifies each customer who made a repurchase at the store. Otherwise, customer_unique_id has been removed. Using a customer_id instead, as shown in Figure 11.



```
df_customers.head()
```

```
                         customer_id  customer_zip_code_prefix         customer_city  customer_state
0   06b8999e2fba1a1fbc88172c00ba8bc7                     14409                franca              SP
1   18955e83d337fd6b2def6b18a428ac77                      9790  sao bernardo do campo              SP
2   4e7b3e00288586ebd08712fdd0374a03                      1151             sao paulo              SP
3   b2b6027bc5c5109e529d4dc6358b12c3                      8775        mogi das cruzes              SP
4   4f2d8ab171c80ec8364f7c12e35b23ad                     13056             campinas              SP
```

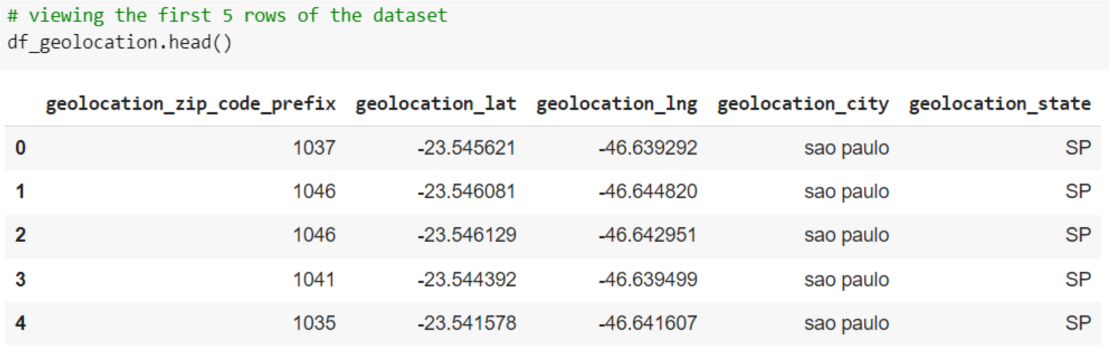
Figure 11 Overview of the customer table

Geo-location Table

Table 18 Description of the geolocation column name

No.	Variable	Description
1.	geolocation_zip_code_prefix	First 5 digits of the zip code
2.	geolocation_lat	Latitude
3.	geolocation_lng	Longitude
4.	geolocation_city	City
5.	geolocation_state	State

According to Brazilian states, Latitude/Longitude coordinates, and zip codes as shown in Table 18. Duplication data is removed, which reduces the row's shape from 1,000,163 to 738,332.



```
# viewing the first 5 rows of the dataset
df_geolocation.head()
```

	geolocation_zip_code_prefix	geolocation_lat	geolocation_lng	geolocation_city	geolocation_state
0	1037	-23.545621	-46.639292	sao paulo	SP
1	1046	-23.546081	-46.644820	sao paulo	SP
2	1046	-23.546129	-46.642951	sao paulo	SP
3	1041	-23.544392	-46.639499	sao paulo	SP
4	1035	-23.541578	-46.641607	sao paulo	SP

Figure 12 Overview of Geo-location Table

Order Items Table

Table 19 Description of item column name

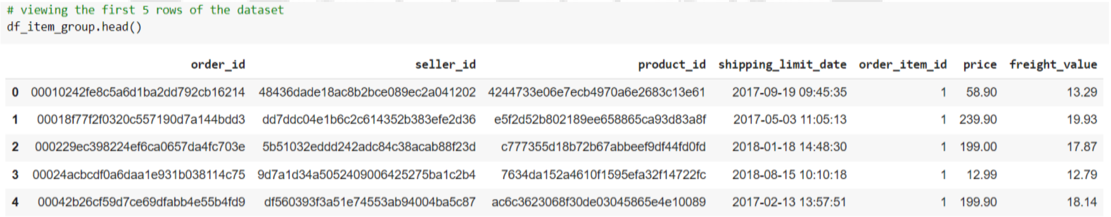
No.	Variable	Description
1.	order_id	Order unique identifier
2.	order_item_id	Sequential number identifying the number of the items included in the same order
3.	product_id	Product unique identifier
4.	seller_id	Seller unique identifier
5.	shipping_limit_date	Seller shipping limit date for handing the order over to the logistic partner

มหาวิทยาลัยศรีนครินทรวิโรฒ

Table 20 (Continued)

6.	price	Item price
7.	freight_value	Item freight value item (if an order has more than one item, the freight value is split between items)

Each order includes the number of items as shown in Table 19 and 20. The limitation is that each order's different sellers, shipment dates, and item count may vary. For instance, the order has id "ca3625898fbd48669d50701aba51cd5f," consisting of 8 different items from 2 stores that ship at different times. The calculation is based on the item's weight and total quantity; therefore, grouping summarize is used to sum overall freight expense and item price of each order as shown in Figure 13.



```
# viewing the first 5 rows of the dataset
df_item_group.head()
```

	order_id	seller_id	product_id	shipping_limit_date	order_item_id	price	freight_value
0	00010242fe8c5a6d1ba2dd792cb16214	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2017-09-19 09:45:35	1	58.90	13.29
1	00018f77f2f0320c557190d7a144bdd3	dd7ddc04e1b6c2c614352b383efe2d36	e5f2d52b802189ee658865ca93d83a8f	2017-05-03 11:05:13	1	239.90	19.93
2	000229ec398224ef6ca0657da4fc703e	5b51032eddd242adc84c38acab88f23d	c777355d18b72b67abbeef9df44fd0fd	2018-01-18 14:48:30	1	199.00	17.87
3	00024acbcdf0a6daa1e931b038114c75	9d7a1d34a5052409006425275ba1c2b4	7634da152a4610f1595efa32f14722fc	2018-08-15 10:10:18	1	12.99	12.79
4	00042b26cf59d7ce69dfabb4e55b4fd9	df560393f3a51e74553ab94004ba5c87	ac6c3623068f30de03045865e4e10089	2017-02-13 13:57:51	1	199.90	18.14

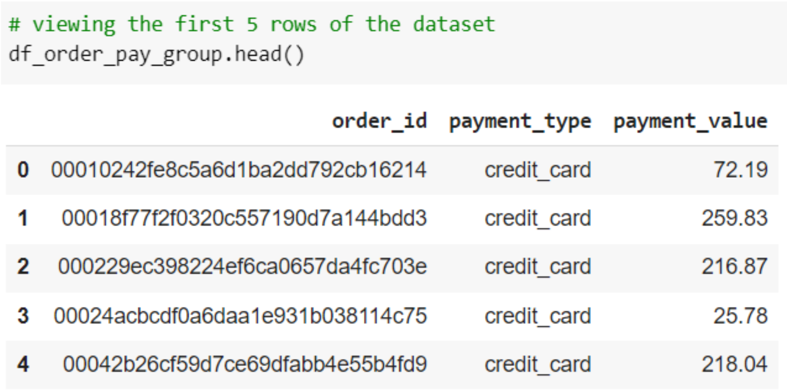
Figure 13 Result of order item table after grouping

Payments Table

Table 21 Description of the payment column name

No.	Variable	Description
1.	order_id	Unique identifier of an order.
2.	payment_sequential	A customer may pay an order with more than one payment method, and will be created to accommodate all payments.
3.	payment_type	Method of payment chosen by the customer.
4.	payment_installments	The number of installments chosen by the customer.
5.	payment_value	Transaction value.

Essential details about the available order payment options are given in Table 21. Each payment option has a section for payment installments that indicates how frequently the customer must pay. So, order payment values are summarized in each payment type and overall transaction value as shown in Figure 14.



```python
# viewing the first 5 rows of the dataset
df_order_pay_group.head()
```

```
                           order_id  payment_type  payment_value
0  00010242fe8c5a6d1ba2dd792cb16214   credit_card          72.19
1  00018f77f2f0320c557190d7a144bdd3   credit_card         259.83
2  000229ec398224ef6ca0657da4fc703e   credit_card         216.87
3  00024acbcdf0a6daa1e931b038114c75   credit_card          25.78
4  00042b26cf59d7ce69dfabb4e55b4fd9   credit_card         218.04
```

Figure 14 Result of payment table after grouping

Order Reviews Table

Table 22 Description of the reviews column name

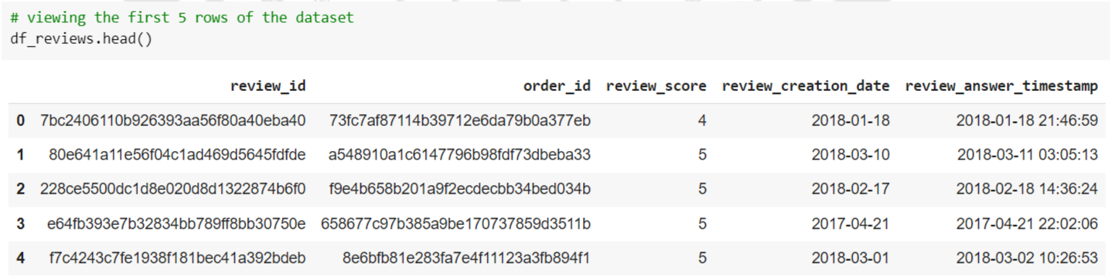
No.	Variable	Description
1.	review_id	Unique review identifier
2.	order_id	Unique order identifier
3.	review_score	The customer gives notes ranging from 1 to 5 on a satisfaction survey.
4.	review_comment_title	Comment title from the review left by the customer in Portuguese.
5.	review_comment_message	Comment message from the review left by the customer in Portuguese.
6.	review_creation_date	The date on the satisfaction survey was sent to the customer.

Table 23 (Continued)

7.	review_answer_timestamp	Satisfaction survey answer timestamp.

Customer review information is provided in Table 22 and 23. When a customer makes a purchase, the seller is notified and must complete the transaction by fulfilling the order. Then, the customer receives an email with a satisfaction survey when their item is finished, or the projected delivery date gets closer. In this survey, they can share their feedback on the shopping experience.

This study intends to learn about a feature's effect on customer satisfaction scores. Thus, review_comment isn't used, and it's removed from the table and reported as Figure 15.



```
# viewing the first 5 rows of the dataset
df_reviews.head()
```

	review_id	order_id	review_score	review_creation_date	review_answer_timestamp
0	7bc2406110b926393aa56f80a40eba40	73fc7af87114b39712e6da79b0a377eb	4	2018-01-18	2018-01-18 21:46:59
1	80e641a11e56f04c1ad469d5645fdfde	a548910a1c6147796b98fdf73dbeba33	5	2018-03-10	2018-03-11 03:05:13
2	228ce5500dc1d8e020d8d1322874b6f0	f9e4b658b201a9f2ecdecbb34bed034b	5	2018-02-17	2018-02-18 14:36:24
3	e64fb393e7b32834bb789ff8bb30750e	658677c97b385a9be170737859d3511b	5	2017-04-21	2017-04-21 22:02:06
4	f7c4243c7fe1938f181bec41a392bdeb	8e6bfb81e283fa7e4f11123a3fb894f1	5	2018-03-01	2018-03-02 10:26:53

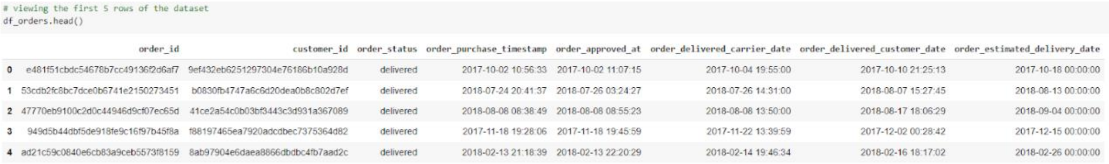
Figure 15 Result of review table after grouping

Order Table

Table 24 Description of the orders column name

No.	Variable	Description
1.	order_id	Unique identifier of the order.
2.	customer_id	Key to the customer dataset. Each order has a unique customer_id.
3.	order_status	Reference to the order status (delivered, shipped, etc.)
4.	order_purchase_timestamp	The purchase timestamp.
5.	Order_approved_at	The payment approval timestamp.
6.	order_delivered_carrier_date	The order posting timestamp. When the logistic partner handled it.
7.	order_delivered_customer_date	The actual order delivery date to the customer.
8.	order_estimated_delivery_date	The estimated delivery date was informed to the customer at the purchase moment.

The order table shows all other information, particularly timing, such as the date of purchase or the date it was delivered in each order as shown in Table 24. All features are needed.



```
# viewing the first 5 rows of the dataset
df_orders.head()
```

```
                           order_id                       customer_id order_status order_purchase_timestamp order_approved_at order_delivered_carrier_date order_delivered_customer_date order_estimated_delivery_date
0  e481f51cbdc54678b7cc49136f2d6af7  9ef432eb6251297304e76186b10a928d    delivered      2017-10-02 10:56:33 2017-10-02 11:07:15          2017-10-04 19:55:00           2017-10-10 21:25:13           2017-10-18 00:00:00
1  53cdb2fc8bc7dce0b6741e2150273451  b0830fb4747a6c6d20dea0b8c802d7ef    delivered      2018-07-24 20:41:37 2018-07-26 03:24:27          2018-07-26 14:31:00           2018-08-07 15:27:45           2018-08-13 00:00:00
2  47770eb9100c2d0c44946d9cf07ec65d  41ce2a54c0b03bf3443c3d931a367089    delivered      2018-08-08 08:38:49 2018-08-08 08:55:23          2018-08-08 13:50:00           2018-08-17 18:06:29           2018-09-04 00:00:00
3  949d5b44dbf5de918fe9c16f97b45f8a  f88197465ea7920adcdbec7375364d82    delivered      2017-11-18 19:28:06 2017-11-18 19:45:59          2017-11-22 13:39:59           2017-12-02 00:28:42           2017-12-15 00:00:00
4  ad21c59c0840e6cb83a9ceb5573f8159  8ab97904e6daea8866dbdbc4fb7aad2c    delivered      2018-02-13 21:18:39 2018-02-13 22:20:29          2018-02-14 19:46:34           2018-02-16 18:17:02           2018-02-26 00:00:00
```

Figure 16 Overview of the customer table

Products Table & Product Category Info

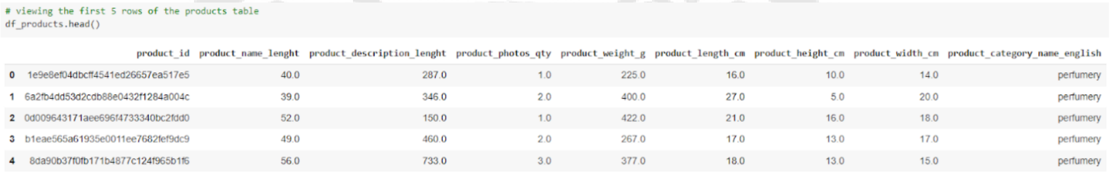
Table 25 Description of the product column name

No.	Variable	Description
1.	product_id	Unique product identifier
2.	product_category_name	Root product category, in Portuguese.
3.	product_name_lenght	The number of characters extracted from the product name.
4.	product_description_lenght	The number of characters extracted from the product description.
5.	product_photos_qty	Number of product-published photos
6.	product_weight_g	Product weight measured in grams.
7.	product_length_cm	Product length measured in centimeters.
8.	product_height_cm	Product height measured in centimeters.

Table 26 (Continued)

9.	product_width_cm	Product width measured in centimeters.

Table 25 and 26 provides details on the products sold at Olist stores, including the root category of the product in Portuguese, and the product category info tables are translated into English and then merged.



```
# viewing the first 5 rows of the products table
df_products.head()
```

```
                         product_id  product_name_lenght  product_description_lenght  product_photos_qty  product_weight_g  product_length_cm  product_height_cm  product_width_cm  product_category_name_english
0  1e9e8ef04dbcff4541ed26657ea517e5                 40.0                       287.0                 1.0             225.0               16.0               10.0              14.0                      perfumery
1  6a2fb4dd53d2cdb88e0432f1284a004c                 39.0                       346.0                 2.0             400.0               27.0                5.0              20.0                      perfumery
2  0d009643171aee696f4733340bc2fdd0                 52.0                       150.0                 1.0             422.0               21.0               16.0              18.0                      perfumery
3  b1eae565a61935e0011ee7682fef9dc9                 49.0                       460.0                 2.0             267.0               17.0               13.0              17.0                      perfumery
4  8da90b37f0fb171b4877c124f965b1f6                 56.0                       733.0                 3.0             377.0               18.0               13.0              15.0                      perfumery
```

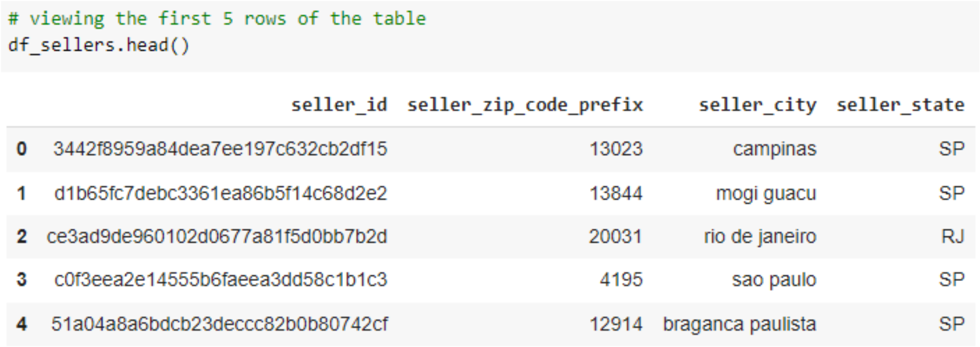
Figure 17 Overview of Products Table

Sellers Table

Table 27 Description of the seller column name

No.	Variable	Description
1.	seller_id	Seller unique identifier
2.	seller_zip_code_prefix	First 5 digits of the seller’s zip code
3.	seller_city	Seller city name
4.	seller_state	Seller state

This table contains information about the sellers who filled orders placed at the Olist shop as shown in Table 27. All features are needed.

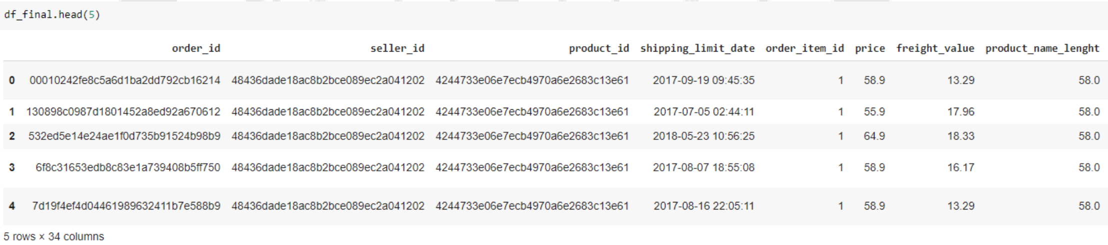


```
# viewing the first 5 rows of the table
df_sellers.head()
```

	seller_id	seller_zip_code_prefix	seller_city	seller_state
0	3442f8959a84dea7ee197c632cb2df15	13023	campinas	SP
1	d1b65fc7debc3361ea86b5f14c68d2e2	13844	mogi guacu	SP
2	ce3ad9de960102d0677a81f5d0bb7b2d	20031	rio de janeiro	RJ
3	c0f3eea2e14555b6faeea3dd58c1b1c3	4195	sao paulo	SP
4	51a04a8a6bdcb23deccc82b0b80742cf	12914	braganca paulista	SP

Figure 18 Overview of seller Table

Inner merging combines all 8 tables by joining through the relation key as in Figure 10, To generate the final table shown in Figure 19.



```
df_final.head(5)
```

	order_id	seller_id	product_id	shipping_limit_date	order_item_id	price	freight_value	product_name_lenght
0	00010242fe8c5a6d1ba2dd792cb16214	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2017-09-19 09:45:35	1	58.9	13.29	58.0
1	130898c0987d1801452a8ed92a670612	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2017-07-05 02:44:11	1	55.9	17.96	58.0
2	532ed5e14e24ae1f0d735b91524b98b9	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2018-05-23 10:56:25	1	64.9	18.33	58.0
3	6f8c31653edb8c83e1a739408b5ff750	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2017-08-07 18:55:08	1	58.9	16.17	58.0
4	7d19f4ef4d04461989632411b7e588b9	48436dade18ac8b2bce089ec2a041202	4244733e06e7ecb4970a6e2683c13e61	2017-08-16 22:05:11	1	58.9	13.29	58.0

5 rows × 34 columns

Figure 19 Example of the data set (pre-cleaning)

Next, missing values in data are identified, and we figured out how we handled them before utilizing the missing data in the next step of model development.

Before starting the process, The data attributes need to be ensured that the description is correct and need to be converted by using astype() method based on usage in Table 28.

Table 28 Description of each data type

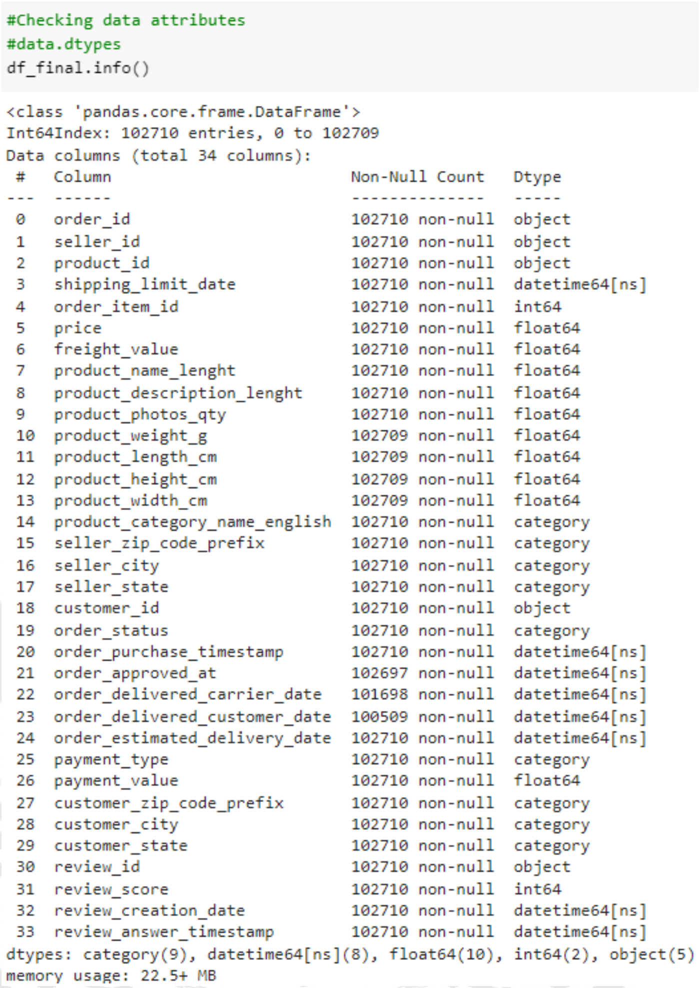
Pandas dtype	Python type	Numpy type	Usage
Object	Str or mixed	String.. , unicode.., mixed types	Text or mixed numeric and non-numeric values
int64	int	int.., int8, int16, int32, int64, uint8, uint16, uint32, uint64	Integer numbers
float64	float	float.., floot16, float32, float64	Floating point numbers
bool	bool	bool_	True/False values
datetime64	NA	datetime64[ns]	Date and time values
timedelta[ns]	NA	NA	Differences between two date times
category	NA	NA	Finite list of text values

Referring to this table, the following group of a feature needs to change type as detailed below.

- Geolocation information identifies a list of local state, city, and zip codes for both customers and seller's side.
    - "Object" to "Category"
        - Customer_state

        - Customer_city
        - Seller_city
        - Seller_state
    - “Integer” to “Category”
        - Customer_zip_code_prefix
        - Seller_zip_code_prefix
- Date & time information that identifies the timing of each activity
    - “Object” to “Datetime[ns]”
        - Shipping_limit_date
        - Review_creation_date
        - Review_answer_timestamp
        - Order_purchase_timestamp
        - Order_approved_at
        - Order_delivered_carrier_date
        - Order_delivered_customer_date
        - Order_estimated_delivery_date
- Payment type information which identifies a type of payment from each customer that selected to pay their orders
    - “Object” to “Category”
        - Payment_type
        - Product category name
    - “Object” to “Category”
        - Product_category_name_english

Figure 20 shows a result after transforming the type of all features and being ready to process the following process to clean data.

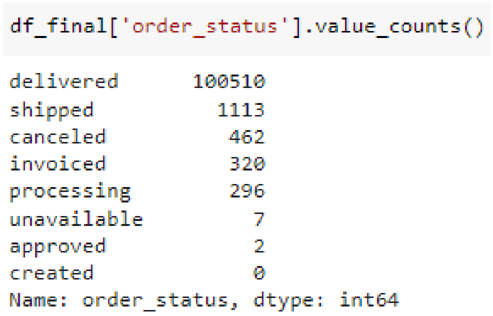


```
#Checking data attributes
#data.dtypes
df_final.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 102710 entries, 0 to 102709
Data columns (total 34 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       102710 non-null  object
 1   seller_id                      102710 non-null  object
 2   product_id                     102710 non-null  object
 3   shipping_limit_date            102710 non-null  datetime64[ns]
 4   order_item_id                  102710 non-null  int64
 5   price                          102710 non-null  float64
 6   freight_value                  102710 non-null  float64
 7   product_name_lenght            102710 non-null  float64
 8   product_description_lenght     102710 non-null  float64
 9   product_photos_qty             102710 non-null  float64
 10  product_weight_g               102709 non-null  float64
 11  product_length_cm              102709 non-null  float64
 12  product_height_cm              102709 non-null  float64
 13  product_width_cm               102709 non-null  float64
 14  product_category_name_english  102710 non-null  category
 15  seller_zip_code_prefix         102710 non-null  category
 16  seller_city                    102710 non-null  category
 17  seller_state                   102710 non-null  category
 18  customer_id                    102710 non-null  object
 19  order_status                   102710 non-null  category
 20  order_purchase_timestamp       102710 non-null  datetime64[ns]
 21  order_approved_at              102697 non-null  datetime64[ns]
 22  order_delivered_carrier_date   101698 non-null  datetime64[ns]
 23  order_delivered_customer_date  100509 non-null  datetime64[ns]
 24  order_estimated_delivery_date  102710 non-null  datetime64[ns]
 25  payment_type                   102710 non-null  category
 26  payment_value                  102710 non-null  float64
 27  customer_zip_code_prefix       102710 non-null  category
 28  customer_city                  102710 non-null  category
 29  customer_state                 102710 non-null  category
 30  review_id                      102710 non-null  object
 31  review_score                   102710 non-null  int64
 32  review_creation_date           102710 non-null  datetime64[ns]
 33  review_answer_timestamp        102710 non-null  datetime64[ns]
dtypes: category(9), datetime64[ns](8), float64(10), int64(2), object(5)
memory usage: 22.5+ MB
```

Figure 20 Type of each feature

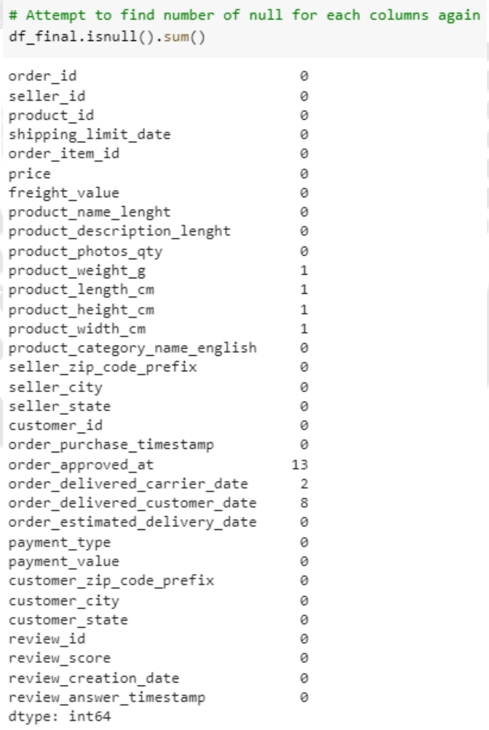
According to the scope of this study and values count in each order status in Figure 21, The delivered order is only used, which means the others will be removed due to the assumption that the order buyers received can only be rated a satisfaction score and decide to drop this column.



```
df_final['order_status'].value_counts()

delivered      100510
shipped          1113
canceled          462
invoiced          320
processing        296
unavailable         7
approved            2
created             0
Name: order_status, dtype: int64
```

Figure 21 Amount of values in each order status

From the missing value in Figure 22, the following process of cleaning the data is to drop off a null row due to dataset show remaining null value below than 1% which not affect too much on dataset.



```
# Attempt to find number of null for each columns again
df_final.isnull().sum()

order_id                          0
seller_id                         0
product_id                        0
shipping_limit_date               0
order_item_id                     0
price                             0
freight_value                     0
product_name_lenght               0
product_description_lenght        0
product_photos_qty                0
product_weight_g                  1
product_length_cm                 1
product_height_cm                 1
product_width_cm                  1
product_category_name_english     0
seller_zip_code_prefix            0
seller_city                       0
seller_state                      0
customer_id                       0
order_purchase_timestamp          0
order_approved_at                13
order_delivered_carrier_date      2
order_delivered_customer_date     8
order_estimated_delivery_date     0
payment_type                      0
payment_value                     0
customer_zip_code_prefix          0
customer_city                     0
customer_state                    0
review_id                         0
review_score                      0
review_creation_date              0
review_answer_timestamp           0
dtype: int64
```

Figure 22 Amount of missing values left in each feature

3.3 Exploratory Data Analysis (EDA) and Visualization

The exploratory data analysis section analyses dataset and highlights vital characteristics. It assists in determining how to effectively modify data sources to obtain answers, making it simpler to identify patterns, detect errors, test a hypothesis, or validate assumptions.

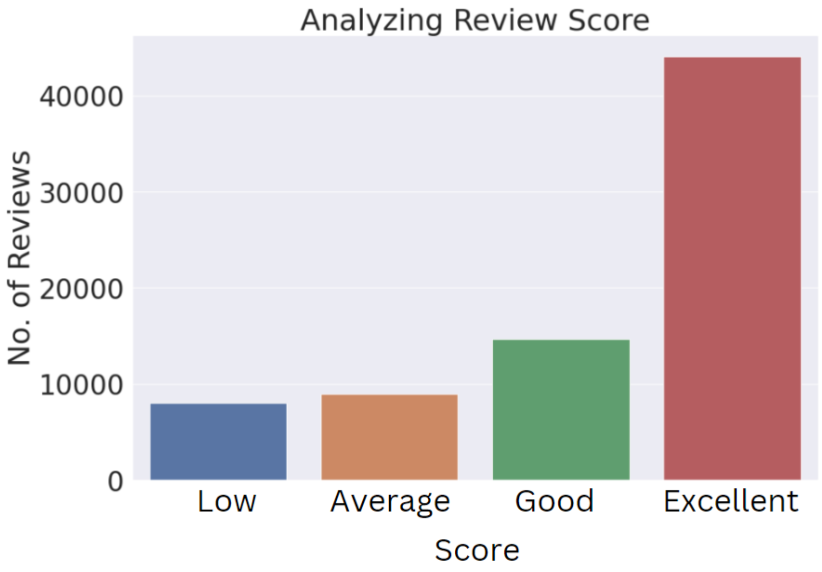


Figure 23 Analyzing review score after grouping in train dataset

Figure 23 shows Customer satisfaction score is grouped into 4 categories: Low, Average, Good, and Excellent to prevent of initiative unbalancing in review scores for a score with a very low sample need to treat and following the assumption of customer satisfaction score. From the picture, most orders receive Excellent review scores, followed by Good, Average, and Low.

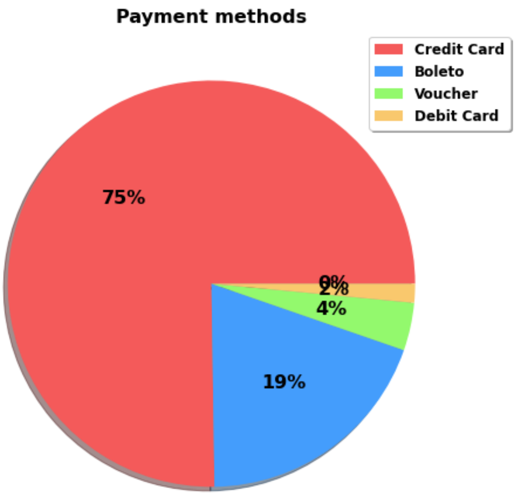


Figure 24 Distribution of payment method

Figure 24 shows customer payment methods, most of the customers around 75% use credit card, 19% use Boleto (bank ticket in Brazil), 4% use voucher, and 2% use debit card to purchase goods.

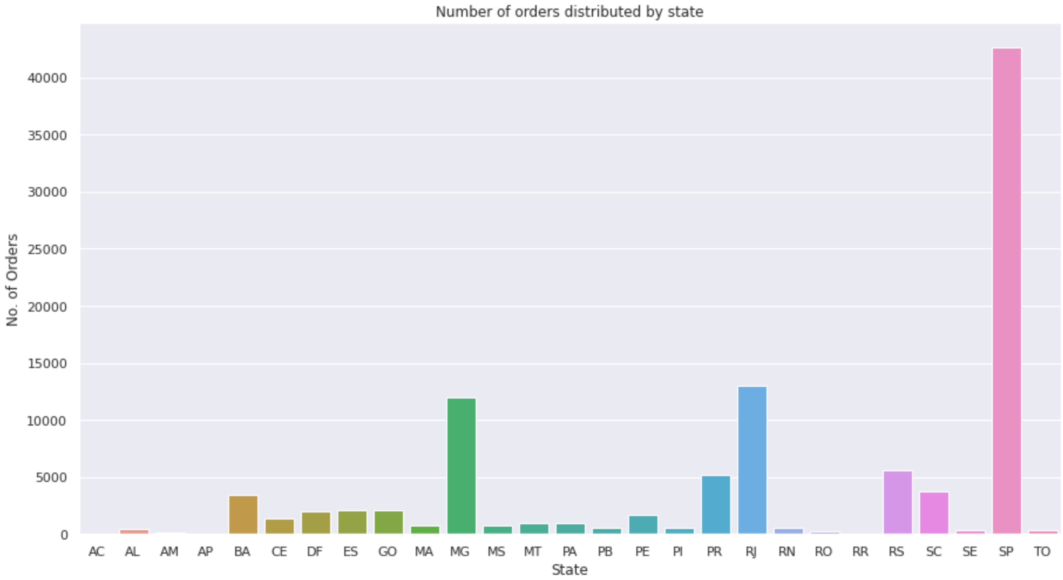


Figure 25 Number of orders distributed by the state

Figure 25 illustrates the states with the highest number of customers. The graph clearly shows that citizens of the state of So Paulo (SP) have placed the greatest number of orders. This state had over 42000 orders or 42.1% of total orders over the Olist shop. Customers from Rio de Janeiro and Minas Gerais have placed the next most orders for the following conditions.

3.4 Split Data

Dataset is split into train and test data sets in a ratio of 75:25 w4here there are 75,803 and 25,268 records in the train and test datasets, respectively. One-hot encoding is used to convert categorical variables into dummies/indicator variables for both train and test sets.

3.5 Feature Engineering

Two types of features are created from the datasets to improve the performance of ML models to predict customer satisfaction score.

3.5.1 Process Duration

As mentioned in [3], most clients feel more satisfied when the ordered packages arrive faster than expected. This type of information could implicitly indicate the customer's satisfaction score. Based on the above assumption, the following features are created from the dataset.

- **Total shipping date**: The actual delivery duration from the seller to the buyers.
- **Estimate the total shipping date**: The estimated delivery duration from the seller to the customer.
- **Delivery performance**: The timing difference between the actual and estimated receiving date.

3.5.2 Average Product and User Rating

Inspired by the fact that either product or seller rating scores obtained from other customers' purchases is one of the main factors that influence customer's satisfaction score, a feature engineering method is employed to calculate the product and seller rating scores. Rating of the seller and product are created by using the average historical review score based on the seller id and product id. The following features are obtained and will be used in the ML model.

- **Seller rating:** Average review score obtained by grouping each seller id
- **Product rating:** Average review score obtained by grouping each product id
- **Seller rating standard deviation:** SD of the review score obtained from the train dataset in each seller id
- **Product rating standard deviation:** SD of the review score obtained from the train dataset in each product id

To prevent data leakage, the average rating and SD are obtained from the train dataset only and will be substituted on the test dataset based on the seller id and product id in the test dataset.

Figure 26 shows a heat map of the correlation matrix of every feature used in the ML model. The top 4 features that have the high correlation with the customer's review score are product_rating, seller_rating, delivery_ performance, and total_shipping_date, respectively.

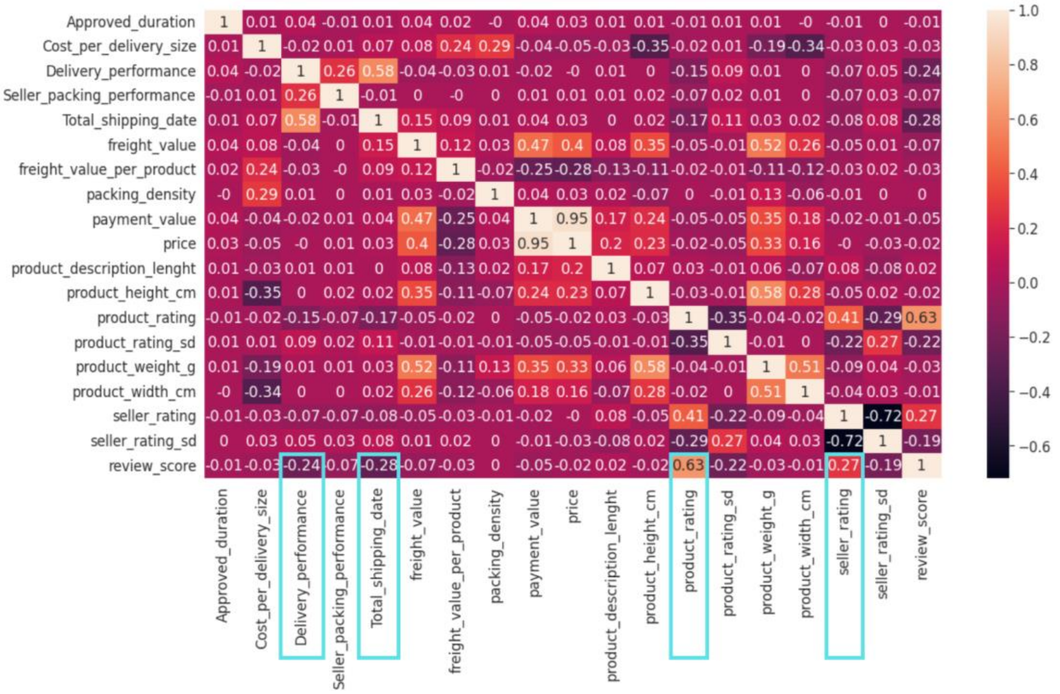


Figure 26 Correlation heatmap of features

Figure 27a-27d display box plots of product_rating, seller_rating, total_shipping_date and delivery_performance", respectively.

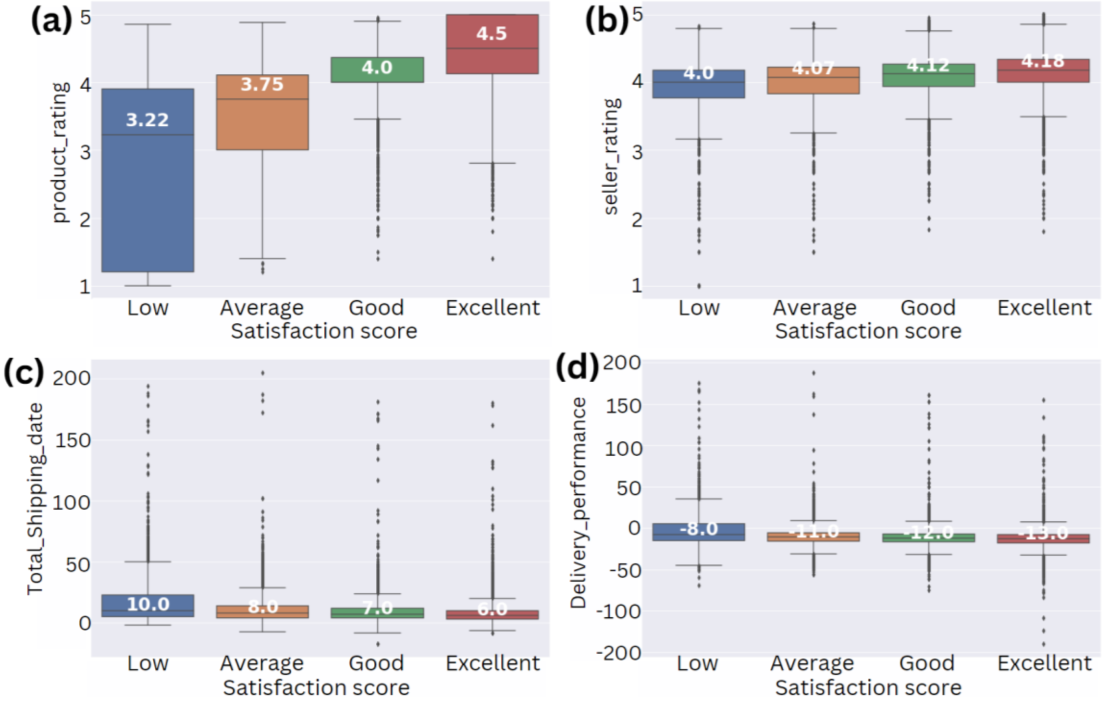


Figure 27 Box plot of a new feature with the satisfaction score (a) product rating (b) seller rating (c) total shipping date (d) delivery performance

As presented in Figure 27a, product_rating exhibits a strong tendency to differentiate 4 classes of review score with some overlapped duration between Low and Average classes. From Fig. 27c and 27d, total_shipping_date and deliver_performance are capable of differentiating parts of Low class from others. seller_rating in Fig. 27b doesn't exhibit any obvious capability to differentiate any classes.

3.6 Feature Selection

This step is implemented to remove unnecessary features that have a low correlation compared to the review score. The Information gained is used to calculate the reduction in entropy from the transformation of a dataset. It can be used for feature selection by evaluating the Information gained from each variable in the context of the target variable. 50% of random data has been used to reduce calculation time.

Information gain is used to determine which features/attributes provide the most information about a class. It complies with the concept of entropy while attempting to reduce the level of entropy. The difference in entropy before and after the split is computed as information gain, which specifies the impurity in-class elements.

The information gain (Gain(S,A)) of an attribute A relative to a collection of data set S, is defined as

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where Values(A) are all possible values for attribute A, and Sv is the subset of S for which attribute A has value v.

Features that have mutual information below the quartile or 25% are removed due to low correlation with a review score, a result describing that features which have low information gain contain the following features: “Estimate total shipping date”, “day of the week", "payment type", "product box size", "product category name English", "product height cm", "product length cm", "product width cm" as shown in Figure 28.

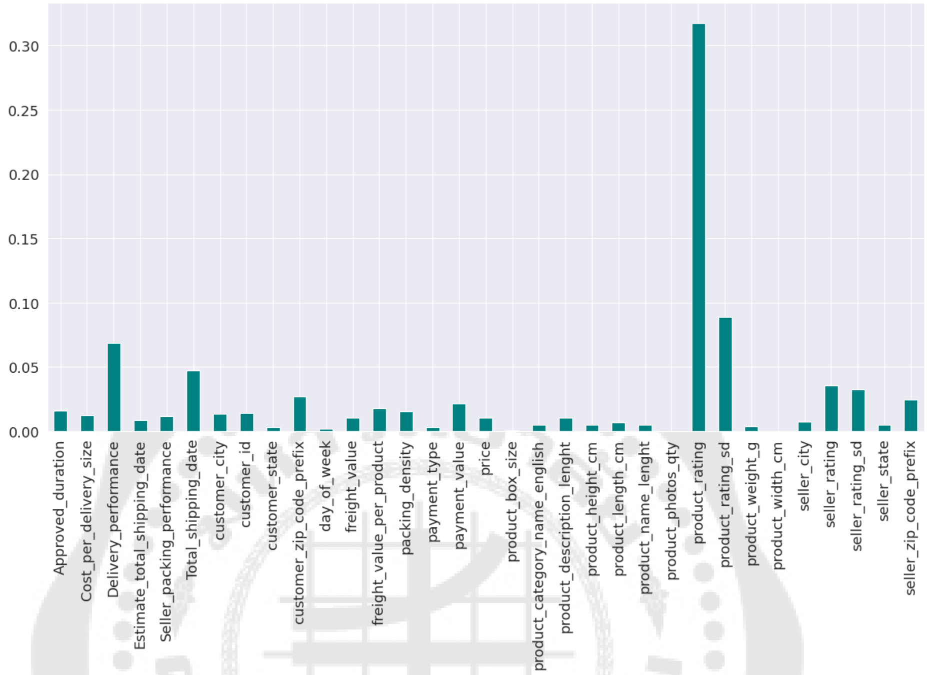


Figure 28 Feature important by mutual information

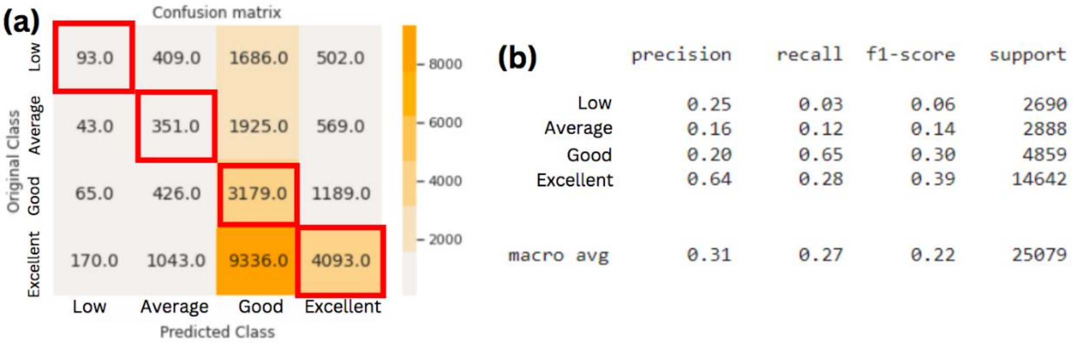
Furthermore, additional features such as "Zip code", "Customer zip code" and "Seller zip code" are discarded due to high computational complexity.

3.7 Modeling

Due to their simplicity and strong performance, RF, K-NN, and LR are three ML models that are used to predict customer satisfaction score. because, in related research of this study.

Before implement the models, data scaling and column transform are used to scaling only numeric columns that represent in the table to remove bias of difference range in each columns.

The average customer review score of each product id obtained from other purchases in the train dataset is used as the baseline model. The confusion matrix and classification report of the average score method is shown in Figure 29. From the figure, Good class yields the best recall and Excellent class yields the best precision score. This is due to the fact that most customers give either a Good or Excellent review score.



```
              precision    recall  f1-score   support

         Low       0.25      0.03      0.06      2690
     Average       0.16      0.12      0.14      2888
        Good       0.20      0.65      0.30      4859
   Excellent       0.64      0.28      0.39     14642


   macro avg       0.31      0.27      0.22     25079
```

Figure 29 (a) Confusion matrix & (b) Classification report of the average score method (baseline)

3.8 Model and Evaluation

From the train dataset, this research dataset will be imbalanced data. So, imbalance treatment is to be used to fix and evaluate the model more efficiently.

3.5.1 Imbalanced data set treatment

By bringing the train data to test into 3 parts were Existing Data, the data that has been added (Over-Sampling) with SMOTE technique, and data reduction (Under-Sampling) with RandomUnder technique and test by running model using default hyper-parameter setting.

3.5.1.1 Existing Data

Figure 30, Show the distribution of data in each class as follow

1) Class “Low” contain with 8061 rows
2) Class “Average” contain with 8940 rows
3) Class “Good” contain with 14709 rows
4) Class “Excellent” contain with 44093 rows

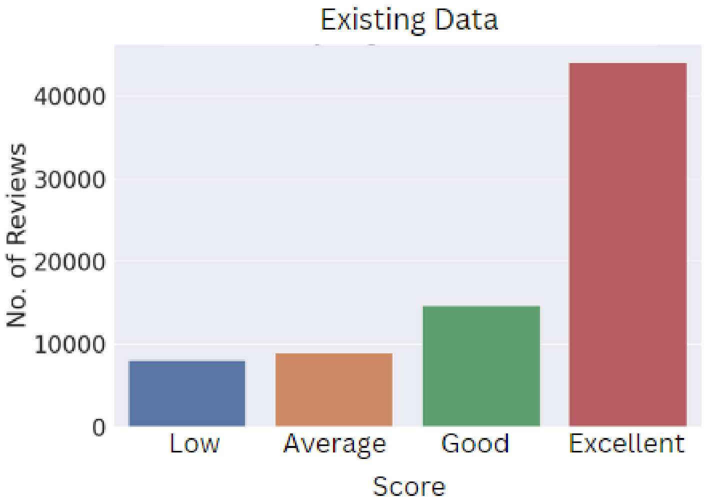


Figure 30 Distribute of existing data

3.5.1.2 Under sampling by RandomUndersampling

Figure 31, Show the distribution of data in each class as follow

5) Class “Low” contain with 8061 rows

6) Class “Average” contain with 8061 rows

7) Class “Good” contain with 8061 rows

8) Class “Excellent” contain with 8061 rows

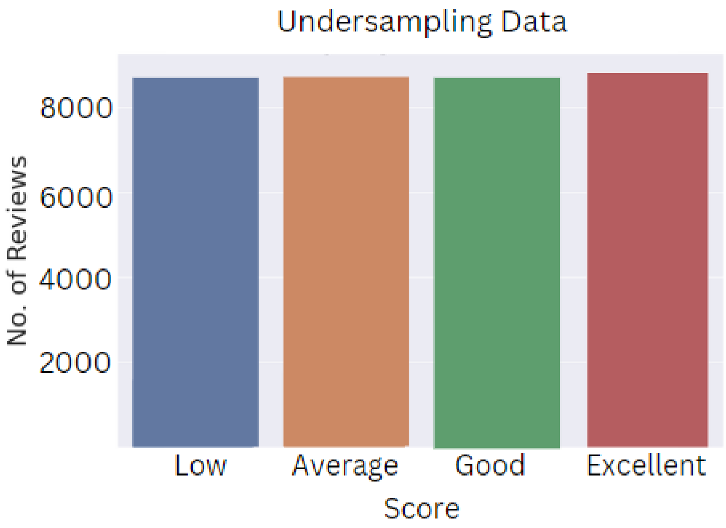


Figure 31 Distribution of under sampling data

3.5.1.3 Over sampling by SMOTE

Figure 32, Show the distribution of data in each class as follow

9) Class “Low” contain with 44093 rows

10) Class “Average” contain with 44093 rows

11) Class “Good” contain with 44093 rows

12) Class “Excellent” contain with 44093 rows

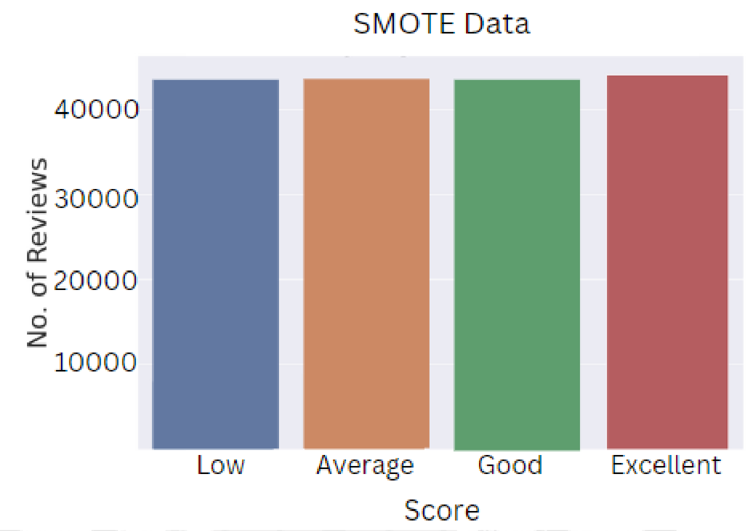


Figure 32 Distribution of SMOTE Data

3.5.2 improvements to hyperparameters with GridSearch technique

Before taking the whole dataset of 3 models, creating the model should be hyper-improved parameters (Hyper-Parameter) so that each parameter is suitable for each model. GridSearchCV technique for imbalance data (StratifiedKFold with n_splits = 5) combined with each algorithm and configured F1-Score is a measure of efficiency by choosing a value Hyper-Parameter related to that algorithm as following table 29.

Table 29 List of Hyper-parameter tuning in each models.

Imbalance treatment	Model	Parameter name	Parameter value
Random UnderSampler SMOTE	RF	min_samples_split	2,5,10,20
		n_estimators	100,150,200
	K-NN	n_neighbors	range(3,30,2)
	LR	C	0.01, 0.1, 1, 10, 100

3.5.3 Model Evaluation:

Because each measurement value indicates the model's performance, choosing the performance measuring of the simulation must be consistent with what needs to be known from the model. As a result, various measurement values will be chosen, including accuracy, completeness (Recall), precision (Precision), and overall model performance (F1-Score).

When the parameters were updated, all 3 types were Existing Data, data that was enhanced (Over-Sampling) with SMOTE techniques, and data that was downgraded (Under-Sampling) with the RandomUnderSampler technique.

The efficiency of the three models was determined by combining test data with models and benchmarks. Validity Must (Accuracy), completeness (Recall), accuracy (Precision), and overall efficiency (F1-Score) The three models were then compared with the average score method to determine which one performed the best.

Chapter 4
Experimental Result

For data imbalance, under sampling (RandomUnder Sampler) and oversampling (SMOTE) are used to treat data and increase more in terms of recall before running through hyperparameter tuning to find the best parameter in each model and show the result in the topics of

1. Result of model performance after imbalance treatment
2. Result of hyper-parameter tuning
3. Result of model evaluation
4. Recheck feature important from best model

4.1 Result of model performance after imbalance treatment

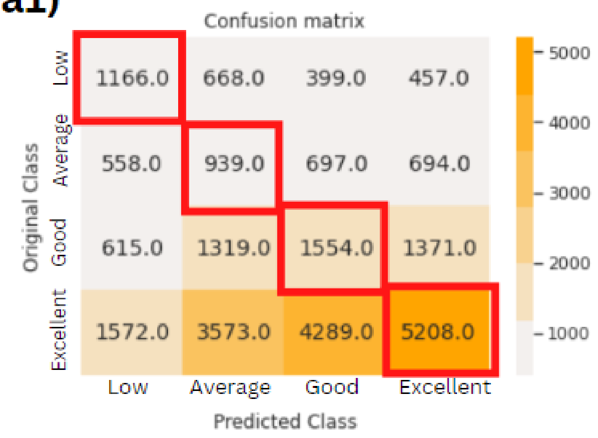
Both data types that treat by under-Sampling technique, and by SMOTE technique, it was trained by RF, K-NN, and LR models using default parameters, and then, the performance of all models was measured using the following metrics, and the result is shown:

1. Precision
2. Recall
3. F1-Score

4.1.2 Under-sampling by RandomUnderSampler technique

Data after treatment by RandomUnderSampler was trained by RF, K-NN, and LR models and report by confusion matrix and all above metrices and performance is shown in Figure 33.

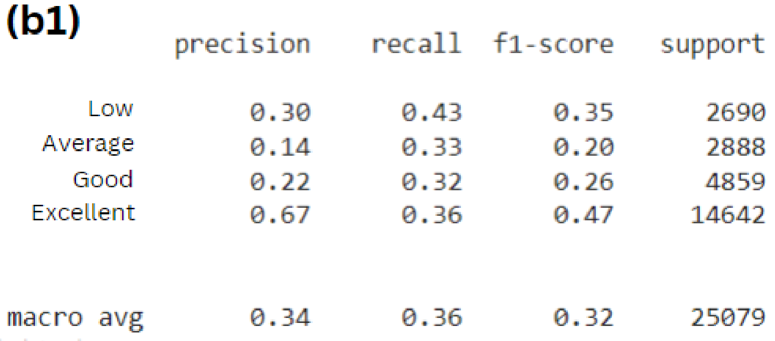
(a1)



Confusion matrix

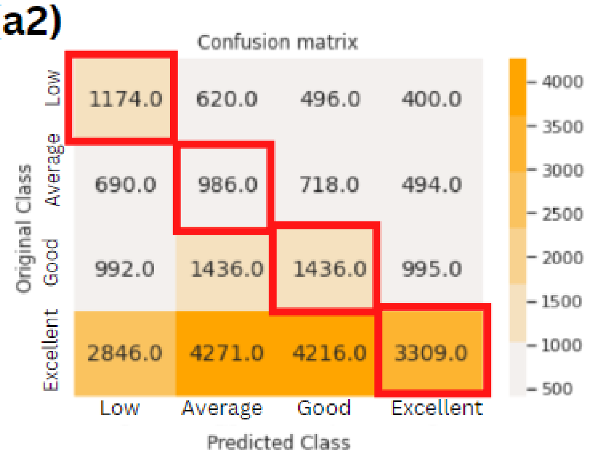
Original Class \ Predicted Class	Low	Average	Good	Excellent
Low	1166.0	668.0	399.0	457.0
Average	558.0	939.0	697.0	694.0
Good	615.0	1319.0	1554.0	1371.0
Excellent	1572.0	3573.0	4289.0	5208.0

(b1)



```
              precision    recall  f1-score   support

         Low       0.30      0.43      0.35      2690
     Average       0.14      0.33      0.20      2888
        Good       0.22      0.32      0.26      4859
   Excellent       0.67      0.36      0.47     14642


   macro avg       0.34      0.36      0.32     25079
```

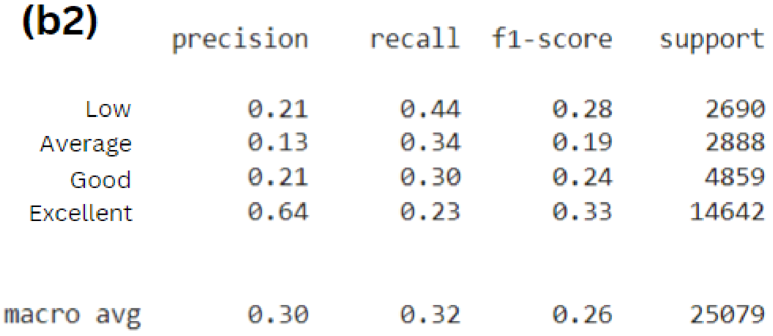
(a2)



Confusion matrix

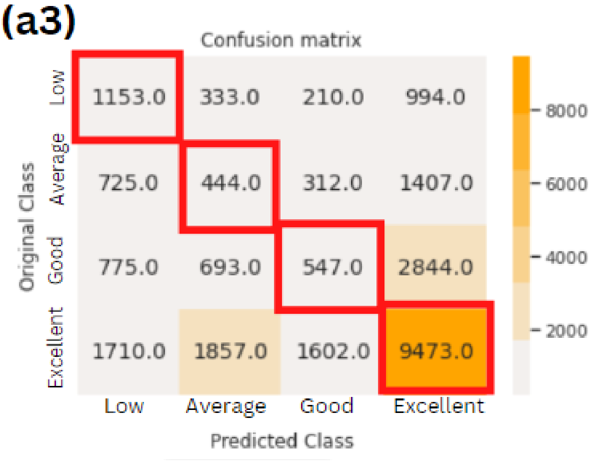
Original Class \ Predicted Class	Low	Average	Good	Excellent
Low	1174.0	620.0	496.0	400.0
Average	690.0	986.0	718.0	494.0
Good	992.0	1436.0	1436.0	995.0
Excellent	2846.0	4271.0	4216.0	3309.0

(b2)



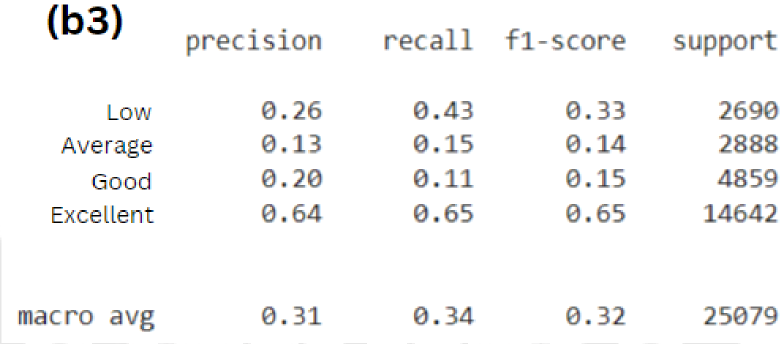
```
              precision    recall  f1-score   support

         Low       0.21      0.44      0.28      2690
     Average       0.13      0.34      0.19      2888
        Good       0.21      0.30      0.24      4859
   Excellent       0.64      0.23      0.33     14642


   macro avg       0.30      0.32      0.26     25079
```

(a3)



	Low	Average	Good	Excellent
Low	1153.0	333.0	210.0	994.0
Average	725.0	444.0	312.0	1407.0
Good	775.0	693.0	547.0	2844.0
Excellent	1710.0	1857.0	1602.0	9473.0

(b3)



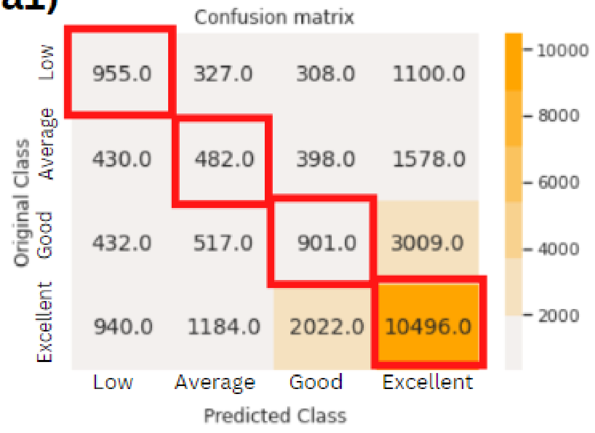
```
              precision    recall  f1-score   support

         Low       0.26      0.43      0.33      2690
     Average       0.13      0.15      0.14      2888
        Good       0.20      0.11      0.15      4859
   Excellent       0.64      0.65      0.65     14642


   macro avg       0.31      0.34      0.32     25079
```

Figure 33 Result of confusion matrix; (a1) RF (a2) K-NN (a3) LR and classification report; (b1) RF (b2) KNN (b3) LR by using RandomUnderSampler and default setting

4.1.3 Over-sampling by SMOTE technique

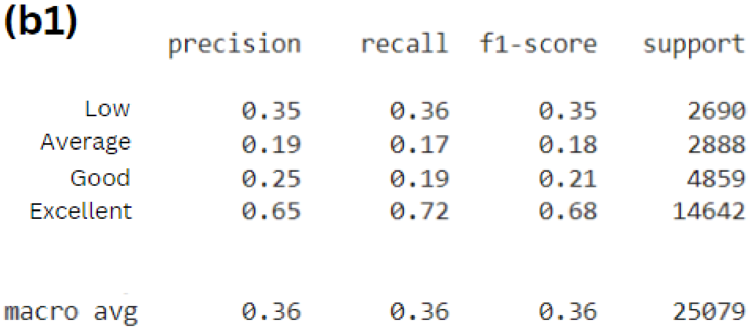
Data after treatment by SMOTE was trained by RF, K-NN, and LR models and report by confusion matrix and all above metrices and performance is shown in Figure 34.

(a1)



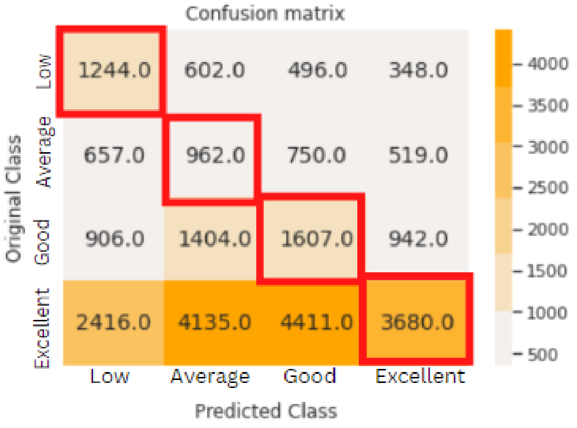
Original Class \ Predicted Class	Low	Average	Good	Excellent
Low	955.0	327.0	308.0	1100.0
Average	430.0	482.0	398.0	1578.0
Good	432.0	517.0	901.0	3009.0
Excellent	940.0	1184.0	2022.0	10496.0

(b1)



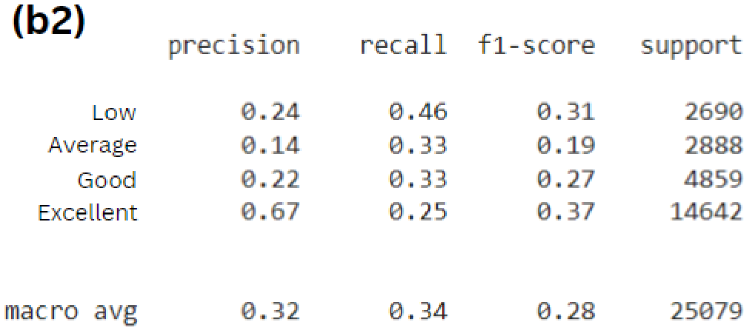
```
              precision    recall  f1-score   support

         Low       0.35      0.36      0.35      2690
     Average       0.19      0.17      0.18      2888
        Good       0.25      0.19      0.21      4859
   Excellent       0.65      0.72      0.68     14642


   macro avg       0.36      0.36      0.36     25079
```

(a2)

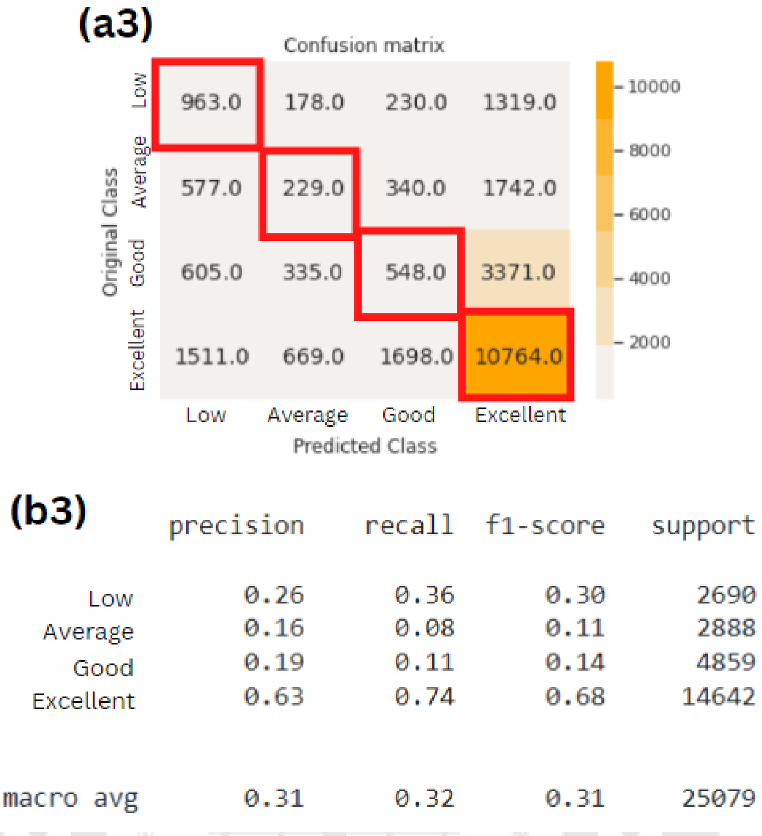


Original Class \ Predicted Class	Low	Average	Good	Excellent
Low	1244.0	602.0	496.0	348.0
Average	657.0	962.0	750.0	519.0
Good	906.0	1404.0	1607.0	942.0
Excellent	2416.0	4135.0	4411.0	3680.0

(b2)



```
              precision    recall  f1-score   support

         Low       0.24      0.46      0.31      2690
     Average       0.14      0.33      0.19      2888
        Good       0.22      0.33      0.27      4859
   Excellent       0.67      0.25      0.37     14642


   macro avg       0.32      0.34      0.28     25079
```



	Low	Average	Good	Excellent
Low	963.0	178.0	230.0	1319.0
Average	577.0	229.0	340.0	1742.0
Good	605.0	335.0	548.0	3371.0
Excellent	1511.0	669.0	1698.0	10764.0

```
              precision    recall  f1-score   support

         Low       0.26      0.36      0.30      2690
     Average       0.16      0.08      0.11      2888
        Good       0.19      0.11      0.14      4859
   Excellent       0.63      0.74      0.68     14642


   macro avg       0.31      0.32      0.31     25079
```

Figure 34 Result of confusion matrix; (a1) RF (a2) K-NN (a3) LR and classification report; (b1) RF (b2) KNN (b3) LR by using SMOTE and default setting.

Result comparison in every metric of all model represent in Table 30 and 31, show RF is the best model in both of imbalance treatment (RandomUnerSampler and SMOTE).

Table 30 Summaries result after imbalance treatment in every model.

Imbalance treatment	Model	Precision	Recall	F1-Score
Random UnderSampler	RF	0.34	0.36	0.32

Table 31 (Continue)

Random UnderSampler	K-NN	0.30	0.32	0.26
	LR	0.31	0.34	0.32
SMOTE	RF	0.36	0.36	0.36
	K-NN	0.32	0.34	0.28
	LR	0.31	0.32	0.31

4.2 Result of hyper-parameter tuning

After hyper-parameter tuning for maximize model performance by using grid search technique, Table 32 and 33 shows the best imbalance treatment method with the best hyperparameter tuning to archive both precision and recall.

Table 32 The Best Results of Parameter Adjustment and Imbalance Treatment Process.

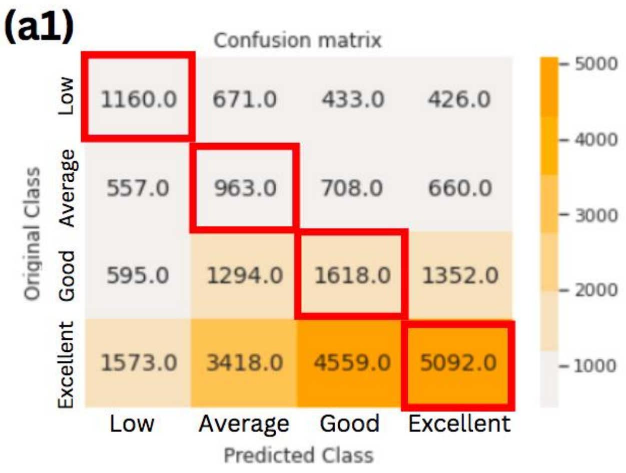
Model	Imbalance treatment	Parameter name	Parameter value
RF	Random UnderSampler	min_samples_split	10
		n_estimators	150

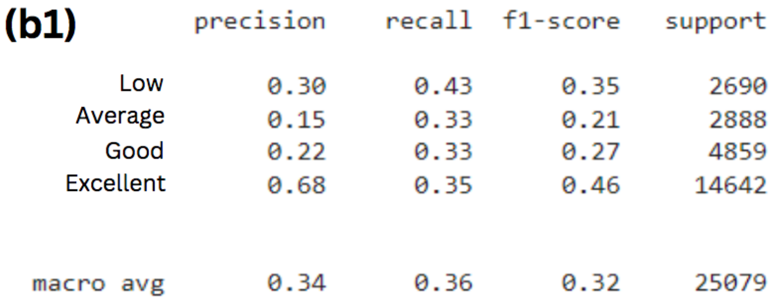
Table 33 (Continue)

K-NN	Random UnderSampler	n_neighbors	27
LR	Random UnderSampler	C	0.1

4.3 Result of model evaluation

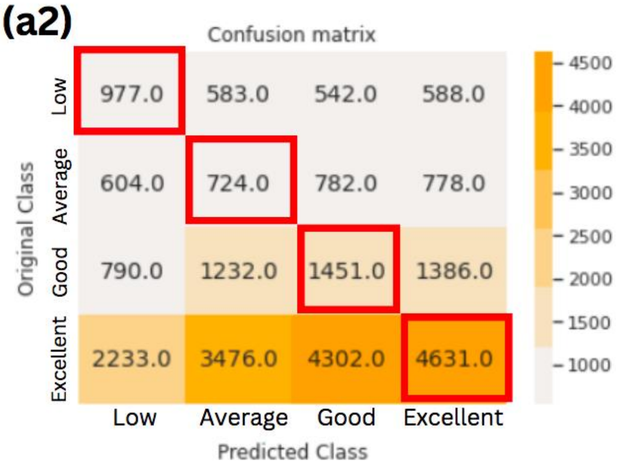
Using hyper-parameters and imbalance treatment by using random under-sampling from Table 32 and 33, performance of all ML models in terms of confusion matrix, precision, recall and F1 score of every class are presented in Figure 35. Figure 35a1, 35a2 and 35a3 show a confusion matrix of the RF, K-NN and LR models, respectively. Figure 35b1, 35b2 and 35b3 show precision, recall and F1 score of each class of RF, K-NN and LR models, respectively.



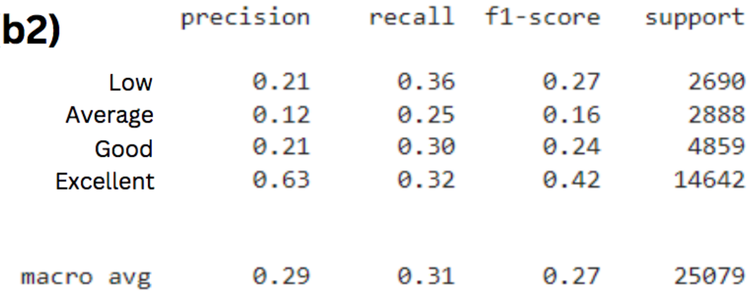


```
              precision    recall  f1-score   support

         Low       0.30      0.43      0.35      2690
     Average       0.15      0.33      0.21      2888
        Good       0.22      0.33      0.27      4859
   Excellent       0.68      0.35      0.46     14642


   macro avg       0.34      0.36      0.32     25079
```

(a2)

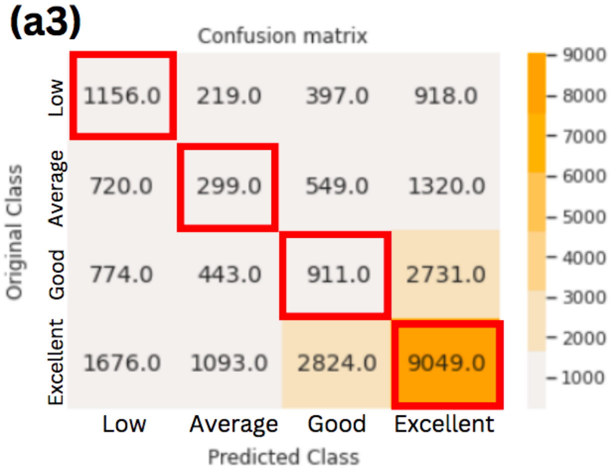


(b2)

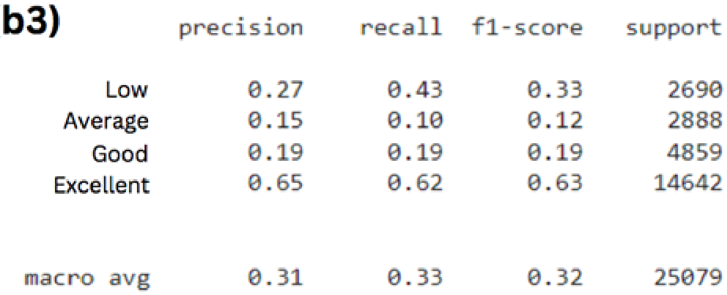


```
              precision    recall  f1-score   support

         Low       0.21      0.36      0.27      2690
     Average       0.12      0.25      0.16      2888
        Good       0.21      0.30      0.24      4859
   Excellent       0.63      0.32      0.42     14642


   macro avg       0.29      0.31      0.27     25079
```

(a3)



(b3)



```
              precision    recall  f1-score   support

         Low       0.27      0.43      0.33      2690
     Average       0.15      0.10      0.12      2888
        Good       0.19      0.19      0.19      4859
   Excellent       0.65      0.62      0.63     14642


   macro avg       0.31      0.33      0.32     25079
```

Figure 35 Result of confusion matrix; (a1) RF (a2) K-NN (a3) LR and classification report; (b1) RF (b2) KNN (b3) LR after fine-tuning

From Figure 35, RF yields the best precision, recall and F1 score of every class with the average precision, recall and F1 score equal to 0.34, 0.36 and 0.32, respectively. LR exhibits lower performance with the average precision, recall, and F1 score equal to 0.31, 0.33, and 0.32, respectively. Among the ML modes, K-NN achieves the lowest performance with average precision, recall, and F1 score equal to 0.29, 0.31, and 0.27, respectively. The baseline model displays high precision score for the Excellent class and high recall score for the Good class but achieves low score for other classes, resulting in the average precision, recall, and F1 score equal to 0.31, 0.27 and 0.22, respectively. In addition, using the extracted features enhances RF recall performance to identify Low, Average and Good classes as good as the Excellent one as presented in Figure 35a1 and 35b1. This is confirmed with the box plot of product_rating feature displayed in Figure 27a. With the RF model, top four features are "product rating", "product rating sd", "delivery performance", and "total shipping date" whose feature importance are equal to 0.312, 0.086, 0.065 and 0.050, respectively.

4.4 Recheck feature important from best model

Based on the evaluation matrix for every class, the RF with Random undersampling treatment and hyperparameter tuning with a minimum split of 10 and an n estimator of 100 represent the best performance. To reconfirm that the most affected features show the same result as the feature selection process, feature importance will be used.

Figure 36 shows the feature importance of the RF model using the MDI result, which indicates that "Product rating" and "Delivery performance" are the two best features with the highest MDI scores, consistent with the feature selection step.

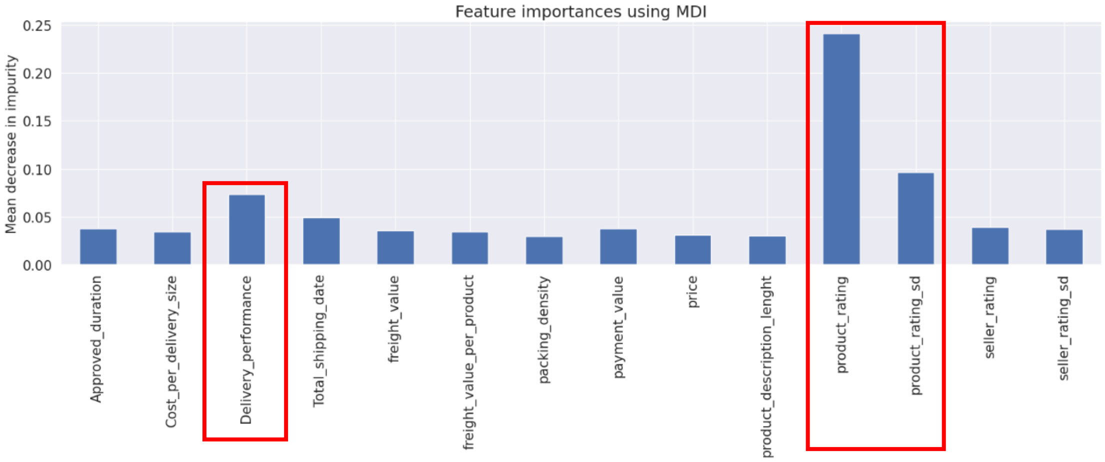


Figure 36 Feature importance using MDI in RF best tuning hyperparameter and imbalance treatment

Chapter 5
Discussion and Conclusion

In researching the prediction of customer satisfaction score and understanding the most affected feature that reflects customer satisfaction score, using the Brazilian e-commerce company Olist Shop's historical data set and machine learning, the researcher measured the performance of each model to compare and summarize the results. The topics for summarizing the results can be divided as follows:

1. Conclusion and discussion
2. Expected Benefits
3. Suggestion

5.1 Conclusion and discussion

The primary objective of this work is to develop an ML model that can accurately predict customer satisfaction scores and identify the features that have the greatest impact on customer satisfaction. Customer satisfaction is a crucial factor in the success of online retail stores and e-commerce businesses. By understanding which features are most important to customers, these businesses can improve their strategies, services, logistics, and customer support to better meet the expectations of their customers.

To achieve this goal, new features that may be relevant to customer satisfaction, such as shipping time, product quality, and customer support response time, have been constructed and fed into three different machine learning models, including RF, K-NN, and LR models. By comparing the performance of these models with a baseline average score, the most effective model for predicting customer satisfaction can be determined.

In addition to predicting customer satisfaction scores, this research also aims to understand which features have the greatest impact on customer satisfaction. By identifying these features, businesses can focus their efforts on improving the areas that matter most to their customers. To this end, the researcher has conducted a feature importance analysis on each of the models. This analysis allows for the determination of which features are the most important for predicting customer satisfaction, as well as how these features interact with each other.

Overall, the insights gained from this research have the potential to help online retail stores and e-commerce businesses optimize their operations and provide better customer experiences. By understanding the factors that drive customer satisfaction, these businesses can improve their services and strategies to better meet the needs and expectations of their customers.

The e-commerce sales data was analyzed using three different machine learning models: RF, K-NN, and LR. The experimental results show that all three models yield superior prediction performance compared to the average score in almost every class, indicating that machine learning models can be effective in predicting customer satisfaction scores for e-commerce businesses.

Of the three models in table 34, In every model show the greater performance compared to baseline model but RF exhibits the best performance in terms of confusion matrix, precision, recall, and F1-score in every class by using the undersampling method. This indicates that RF is particularly effective at predicting customer satisfaction for these three classes.

Table 34 Summaries result after imbalance treatment and fine-tuning in every models compare with baseline model

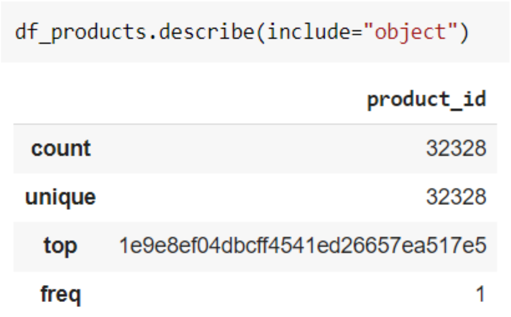
Model	Precision	Recall	F1-Score
Baseline model	0.31	0.27	0.21
RF	0.34	0.36	0.32
K-NN	0.29	0.31	0.27
LR	0.31	0.33	0.32

Further analysis was conducted to determine which features had the highest importance in predicting customer satisfaction. Using the undersampling method, it was found that the mean and standard deviation of the product rating were the two features with the highest importance, with importance scores of 0.313 and 0.087, respectively. This suggests that the quality of the product is a crucial factor in determining customer satisfaction.

The analysis also revealed that customer satisfaction is primarily influenced by two main types of features: product rating and delivery performance. This was confirmed by the correlation matrix presented in Figure 26 and the box plot in Figure 27 then reconfirmation again in Figure 37 by using the MDI score or feature important inside the RF library. Therefore, by focusing on improving the quality of the products and optimizing delivery performance, businesses can improve customer satisfaction and increase their chances of success in the e-commerce industry.

However, when the results of the three metrics used in this study are compared to related research, Insightly's performance is still quite low. The research shows that there are several topics that have a significant effect on the overall performance of the three models.

The first topic is the order average per product ID contribution. The overall data contains 112,000 orders with a total of 32,328 registered products in their database, as shown in Figure 37.



```
df_products.describe(include="object")
```

	product_id
count	32328
unique	32328
top	1e9e8ef04dbcff4541ed26657ea517e5
freq	1

Figure 37 Amount of product in this data set

However, the average number of orders per product ID is only 3-4, which is not enough to train the model to accurately predict product ratings. This means that the data is insufficient to generate reliable insights into customer satisfaction scores.

The second topic is the big gap difference in each class. According to Figure 23, the majority of product ratings are located in the "Excellent" class, indicating that performance in this class is significantly higher than in other classes. This means that the model is biased towards predicting higher ratings and is not able to accurately predict lower ratings.

Despite these limitations, the findings of this study demonstrate the potential of machine learning models to accurately predict customer satisfaction scores and identify the most important features in determining customer satisfaction for e-commerce businesses. By understanding which features are most influential in driving customer satisfaction, businesses can focus their efforts on improving those areas to provide better experiences for their customers. Therefore, it is crucial for companies to consider the limitations of their data and model, and to continually improve and refine their approaches to better predict customer satisfaction scores.

5.2 Expected Benefits

The main purpose for finding the important features that affect the satisfaction score. This information enabled companies to adjust marketing and operation strategies to address the issues more effectively. For example, if the store has low delivery performance, they can address this by changing to a more suitable delivery service.

5.3 Suggestion

5.3 .1 Future work in this area could focus on incorporating user comments as another type of feature to improve the accuracy of the machine learning model in predicting customer satisfaction scores.

5.3.2 Need to improve the responsiveness of a low customer satisfaction review because most buyers do not respond after receiving their package, in order to collect more data and increase the accuracy of the average rating.

5. 3 .3 Future work can modify and tested with other ML models to see the possibility of improvement in terms of performance.

REFERENCES

[1] W. C. Ben Lutkevich, Brian Holak. "e-commerce." https://www.techtarget.com/searchcio/definition/e-commerce (accessed Aug, 2022).

[2] S. Chevalier. "Global retail e-commerce sales 2014-2026." https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/ (accessed Aug, 2022).

[3] A. Griva, ""I can get no e-satisfaction". What analytics say? Evidence using satisfaction data from e-commerce," *Journal of Retailing and Consumer Services,* vol. 66, p. 102954, 2022/05/01/ 2022, doi: https://doi.org/10.1016/j.jretconser.2022.102954.

[4] R. Chinomona, G. Masinge, and M. Sandada, "The Influence of E-Service Quality on Customer Perceived Value, Customer Satisfaction and Loyalty in South Africa," *Mediterranean Journal of Social Sciences,* vol. 5, pp. 331-341, 05/01 2014, doi: 10.5901/mjss.2014.v5n9p331.

[5] A.-N. Wong and B. Poolan Marikannan, *Optimising e-commerce customer satisfaction with machine learning*. 2020.

[6] T. Wu and X. Liu, "A dynamic interval type-2 fuzzy customer segmentation model and its application in E-commerce," *Applied Soft Computing,* vol. 94, p. 106366, 2020/09/01/ 2020, doi: https://doi.org/10.1016/j.asoc.2020.106366.

[7] B. Marr. "How To Understand Your Customers And Their Needs With The Right Data." https://www.forbes.com/sites/bernardmarr/2022/02/03/how-to-understand-your-customers-and-their-needs-with-the-right-data/?sh=68e674492f68 (accessed.

[8] M. I. Hossain, M. Rahman, T. Ahmed, and A. Z. M. T. Islam, "Forecast the Rating of Online Products from Customer Text Review based on Machine Learning Algorithms," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 27-28 Feb. 2021 2021, pp. 6-10, doi: 10.1109/ICICT4SD50815.2021.9396822.

[9] T. K. Phung, N. A. Te, and T. T. T. Ha, "A machine learning approach for opinion mining online customer reviews," in *2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter)*, 28-30 Jan. 2021 2021, pp. 243-246, doi: 10.1109/SNPDWinter52325.2021.00059.

[10] V. Jain, B. Malviya, and S. Arya, "An Overview of Electronic Commerce (e-Commerce)," *Journal of Contemporary Issues in Business and Government,* vol. 27, pp. 665-670, 05/22 2021, doi: 10.47750/cibg.2021.27.03.090.

[11] D. Clarke and R. Kinghorn, "Experience is everything: Here's how to get it right," p. 5. [Online]. Available: pwc.com/future-of-cx

[12] T. Koivumäki, "Customer Satisfaction and Purchasing Behaviour in a Web-based Shopping Environment," *Electronic Markets,* vol. 11, 07/01 2001, doi: 10.1080/101967801681008022.

[13] M. Kotsokechagia, "Predictive model for customer satisfaction in

e-commerce," Master of Science, e-Business and Digital Marketing, UNIVERSITY CENTER OF INTERNATIONAL PROGRAMMES OF STUDIES SCHOOL OF SCIENCE AND TECHNOLOGY 2021. [Online]. Available: http://hdl.handle.net/11544/29876

[14] I. C. Education. "Machine Learning." July 2020. https://www.ibm.com/cloud/learn/machine-learning#:~:text=Machine%20learning%20is%20a%20branch,learn%2C%20gradually%20improving%20its%20accuracy (accessed Sep, 2022).

[15] javatpoint. "Classification Algorithm in Machine Learning." https://www.javatpoint.com/classification-algorithm-in-machine-learning (accessed Sep, 2022).

[16] I. C. Education. "Random Forest." https://www.ibm.com/cloud/learn/random-forest (accessed Aug, 2022).

[17] IBM. "What is the k-nearest neighbors algorithm?" https://www.ibm.com/th-en/topics/knn (accessed Aug, 2022).

[18] javatpoint. "K-Nearest Neighbor(KNN) Algorithm for Machine Learning."

https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (accessed Aug, 2022).

[19] javatpoint. "Logistic Regression in Machine Learning." https://www.javatpoint.com/logistic-regression-in-machine-learning (accessed Dec, 2022).

[20] S. Mazumder. "5 Techniques to Handle Imbalanced Data For a Classification Problem." https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/ (accessed Sep, 2022).

[21] dominodatalab. "Model Evaluation." https://www.dominodatalab.com/data-science-dictionary/model-evaluation#:~:text=Model%20evaluation%20is%20the%20process,a%20role%20in%20model%20monitoring (accessed Aug, 2022).

[22] J. JORDAN. "Evaluating a machine learning model." https://www.jeremyjordan.me/evaluating-a-machine-learning-model/ (accessed Aug, 2022).

[23] S. Narkhede. "Understanding Confusion Matrix." https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (accessed Aug, 2022).

[24] N. N. Moon, I. M. Talha, and I. Salehin, "An advanced intelligence system in customer online shopping behavior and satisfaction analysis," *Current Research in Behavioral Sciences,* vol. 2, p. 100051, 2021/11/01/ 2021, doi: https://doi.org/10.1016/j.crbeha.2021.100051.

[25] Kareena and R. Kumar, "A consumer behavior prediction method for e-commerce application," *International Journal of Recent Technology and Engineering,* vol. 8, no. 2 Special Issue 6, pp. 983-988, 2019/7// 2019, doi: 10.35940/ijrte.B1171.0782S619.

[26] S. A. Alquhtani and A. Muniasamy, "Analytics in Support of E-Commerce Systems Using Machine Learning," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 20-22 July 2022 2022, pp. 1-5, doi:

10.1109/ICECET55527.2022.9872592.

[27] P. Hamsagayathri and K. Rajakumari, "Machine learning algorithms to empower Indian women entrepreneur in E-commerce clothing," in *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 22-24 Jan. 2020 2020, pp. 1-5, doi: 10.1109/ICCCI48352.2020.9104111.

[28] S. R. Siva, S. N. Bushra, B. Maheswari, and M. K. Prabukumar, "Performance Analysis of Different Machine Learning in Customer Prediction," in *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*, 28-30 April 2022 2022, pp. 1425-1430, doi: 10.1109/ICOEI53556.2022.9776714.

[29] Y. Zhang, "Prediction of Customer Propensity Based on Machine Learning," in *2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, 22-24 Jan. 2021 2021, pp. 5-9, doi: 10.1109/ACCTCS52002.2021.00009.

[30] H. Gou, L. Su, G. Zhang, W. Huang, Y. Rao, and Y. Yang, "A XGBoost Method Based on Telecom Customer Satisfaction Enhancement Strategy," in *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, 19-21 Aug. 2022 2022, pp. 209-213, doi: 10.1109/PRAI55851.2022.9904203.

[31] O. A. Sionek. *Brazilian E-Commerce Public Dataset by Olist*, doi: 10.34740/kaggle/dsv/195341.

มหาวิทยาลัยศรีนครินทรวิโรฒ

VITA