Machine Learning Approach to Predict E-commerce Customer Satisfaction Score

Pongthanin Wangkiat
Dept. of Computer Science
Srinakharinwirot University
Bangkok, Thailand
pongthanin.wk@g.swu.ac.th

Chantri Polprasert
Dept. of Computer Science
Srinakharinwirot University
Bangkok, Thailand
chantri@g.swu.ac.th

***Abstract*—In this paper, we investigate the performance of machine learning (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 fact that delivery duration and product rating score obtained from other customers' purchase are one of the main factors that influence customer's satisfaction score, we propose a feature engineering method that creates delivery duration and the average product rating score which are used as the main features in the ML model. We employ Random Forest (RF), Logistic Regression (LR), and K-Nearest Neighbor (K-NN) to predict customers' satisfaction score and their performance are compared with the baseline model which predicts the customer satisfaction score using the average product rating score. Results show that 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.**

***Keywords*—customer satisfaction, e-commerce, classification, machine learning, rating prediction**

I. Introduction

Nowadays, 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 [1], global retail e-commerce sales is $5,211 billion in 2021. The increase in players and the diversity of clients drive intense competition among businesses, which influences customer retention.

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 [2]. 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 [3]. Customer satisfaction depends on several factors, including the purchase, repeat purchase, product return, product attributes, inventory, logistics, and customer support [4]. Many companies should concentrate on improving their operational performance because negative customer feedback could possibly impact sales and brand image. The big challenge for companies is identifying the critical factors influencing customer satisfaction leading to customer retention and enabling more productive use of organizational resources and enhanced operational performance [5].

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 [6]. 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], [10-11]. 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. Section II discusses literature review on research related to online customer satisfaction score prediction using ML method. Methodology is explained in Section III. Experimental results, discussions and conclusions are presented in Section IV and V, respectively.

II. Literature review

There are various pieces of research on online customer prediction using ML models. Researchers from APU [4] apply Decision Tree (DT), Artificial Neural Network (ANN), Support Vector Machine (SVM), and RF models to predict the key drivers that influence the customer satisfaction score of Brazil's e-commerce. The results indicate that RF achieves the best result with the highest accuracy. Moreover, they show that the estimated delivery date and the number of days to deliver goods are the top two important factors that impact customer satisfaction. Another researcher [7] shows that Naive Bayes (NB) classifier yields the best result with 88% accuracy. Article [8] studied ML models to predict product ratings from "GrammarandProduct Reviews" using RF, LR, and XGBoost (XgB). They found that RF yields the best performance in terms of accuracy, precision, recall, and F1-scores of 94%. In addition, researchers [9] employ RF, LR, NB, SVM, DT, and Neutral Networks (NN) to predict online customer reviews from a hotel website "Agoda.com" in Vietnam. They labeled the dataset and classified review scores on two levels; less than 7.0 are negative, and more significant than 7.0 are positive. Their model's performance is evaluated based on accuracy, precision, recall, and F1 score. The results demonstrate that LR achieves the best performance. For the hybrid model of K-NN and DT to predict products rating of Amazon e-commerce datasets, researchers [10] provide the results that the hybrid model is the suitable method and uses significantly an accuracy of 90.75% with higher and faster execution compared to the NB. Reduced Error Pruning (REPTree) and SVM were used as a classifier in the research [11]. Researchers compare the models together in terms of accuracy, recall, and precision and they found that REPTree is the best method. As last, RF and LR are used as a model in the study of [12] to predict customer propensity to buy the product from web browsing. Their result shows that RF is the best fit in terms of accuracy, Recall, precision, and F1-score.

III. Methodology

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, as shown in Fig.1.

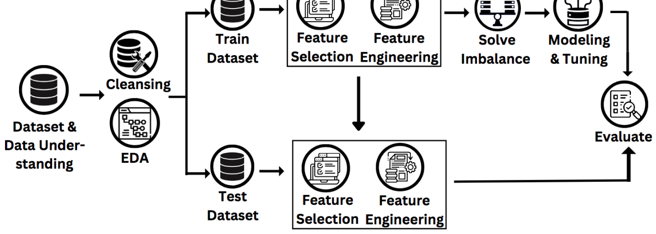


Fig. 1. The prediction process schematic

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. 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 purchase.

A. Dataset & 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 [13].

Datasets contain approximately 34 columns and 102,710 rows. In addition, it includes 8 tables of data which list as follows:

- payments_dataset
- products_dataset
- product_category_name_translation
- order_reviews_dataset
- orders_dataset
- order_items_dataset
- sellers_dataset
- order_customers_dataset
- geolocation_dataset

The data is separated into different datasets to facilitate comprehension which explains and shows the relationship between each table. The 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.

B. 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.

1) **Customers Table:** customer_unique_id has been removed. Using a customer_id instead.
2) **Geo-location Table:** this table has been removed
3) **Order Items Table**: grouping summarize is used to sum overall freight expense and item price of each order.
4) **Payments Table:** summarize order payment values in each payment type and overall transaction value.
5) **Order Reviews Table:** review_comment isn't used, and it's removed from the table.
6) **Order Table:** all features are needed.
7) **Products Table:** merging with product category name translation table to translate product category name from Portuguese to English.
8) **Sellers Table:** all features are needed.

C. Exploratory data analysis (EDA) and visualization

Over 75% of customers use credit card to purchase goods as shown in Fig. 2.

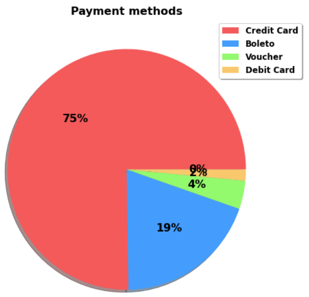


Fig. 2. Distribution of payment method

Fig.3 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 42,000 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.

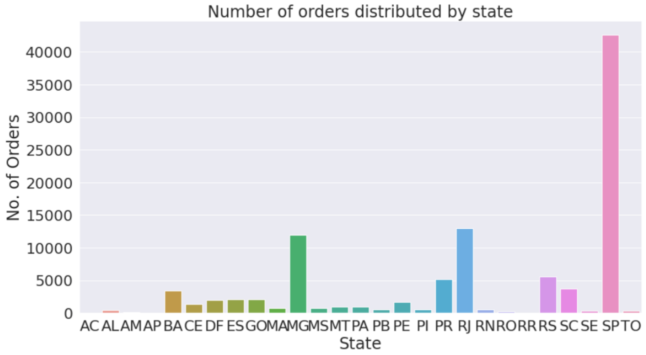


Fig. 3. Number of orders distributed by the state

Customer satisfaction score is grouped into 4 categories: Low, Average, Good, and Excellent. A number of reviews based on each category is presented in Fig.4 . From the picture, most orders receive Excellent review scores, followed by Good, Average and Low.

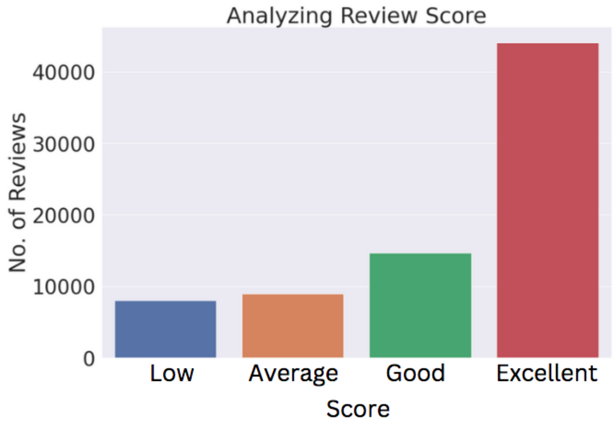


Fig. 4. Analyzing Review Score after grouping in train dataset

D. Split Data

Dataset is split into train and test data sets in a ratio of 75:25 where 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.

E. Feature engineering

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

1) Process duration

As mentioned in [2], 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.

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.

Fig. 5 shows a heat map of the correlation matrix of every feature used in the ML model.

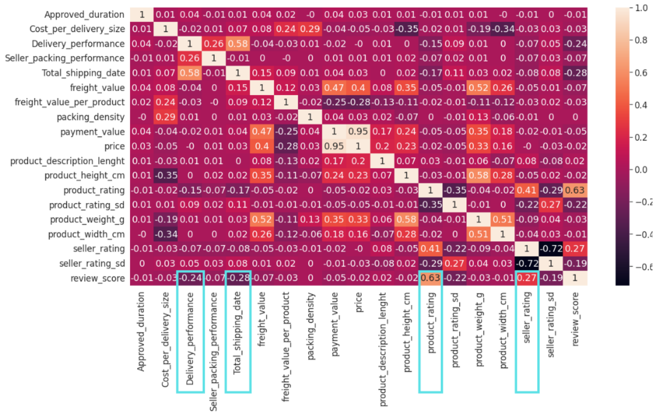


Fig. 5. Correlation heatmap of features

From Fig. 5, 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. Fig. 6a-6d display box plots of product_rating, seller_rating, total_shipping_date and delivery_performance", respectively.

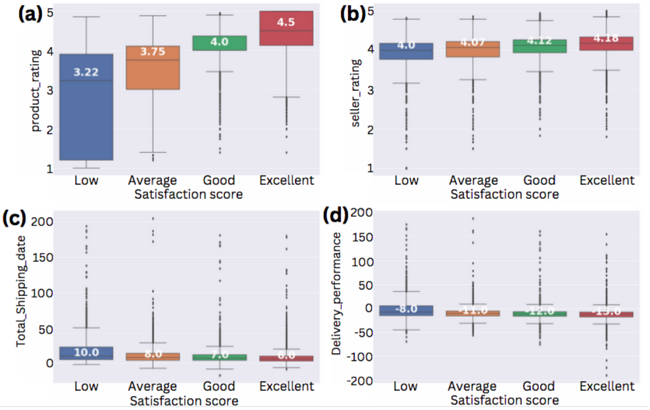


Fig. 6. Box plot of a new feature with the satisfaction score (a) product rating (b) seller rating (c) total shipping date (d) delivery performance

As presented in Fig. 6a, product_rating exhibits a strong tendency to differentiate 4 classes of review score with some overlapped duration between Low and Average classes. From Fig. 6c and 6d, total_shipping_date and deliver_performance are capable of differentiating parts of Low class from others. seller_rating in Fig. 6b doesn't exhibit any obvious capability to differentiate any classes.

F. 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

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

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 Fig 7.

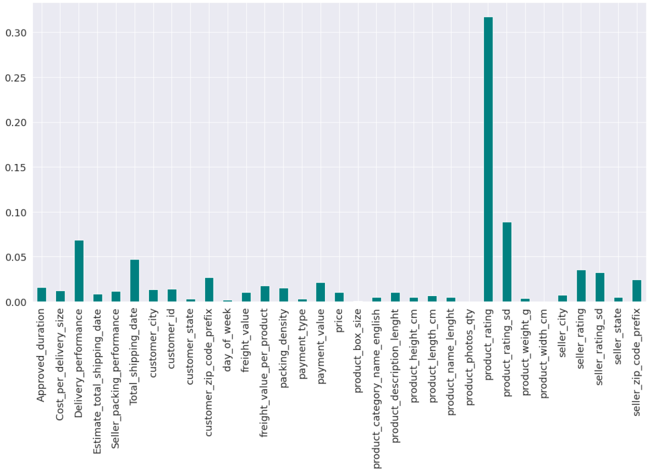


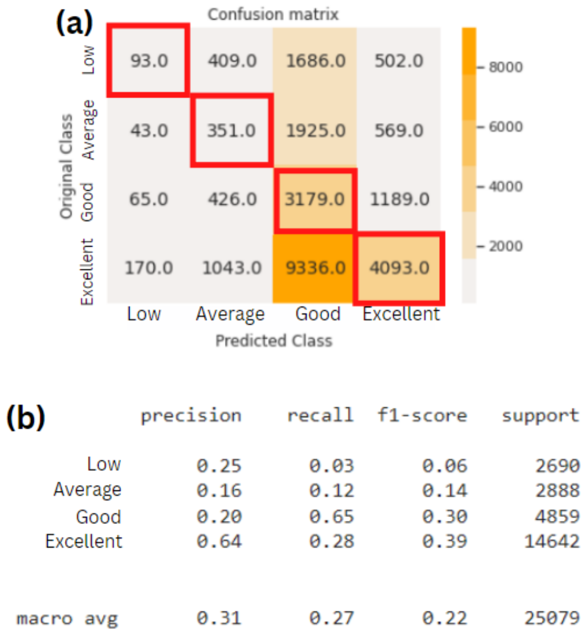
Fig. 7. Feature important by mutual information

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

G. Modeling

Due to their simplicity and strong performance, RF, K-NN, and LR are three ML models that are used to predict customer satisfaction scores Due to their simplicity and strong performance, RF, K-NN, and LR are three ML models that are used to predict customer satisfaction scores because, in related research of this study, they are used in terms of both regression and classification which represent these 3 models is the group of best in performance for e-commerce satisfaction score prediction.

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 Fig 9. 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
```

Fig. 9. (a) Confusion matrix & (b) Classification report of the average score method (baseline)

H. 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.

1) Imbalanced data set treatment: By bringing the train data to test into 2 parts of the data that has been added (Over-Sampling) with SMOTE technique, and data reduction (Under-Sampling) with RandomUnder technique.

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.

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.

IV. Experimental Result

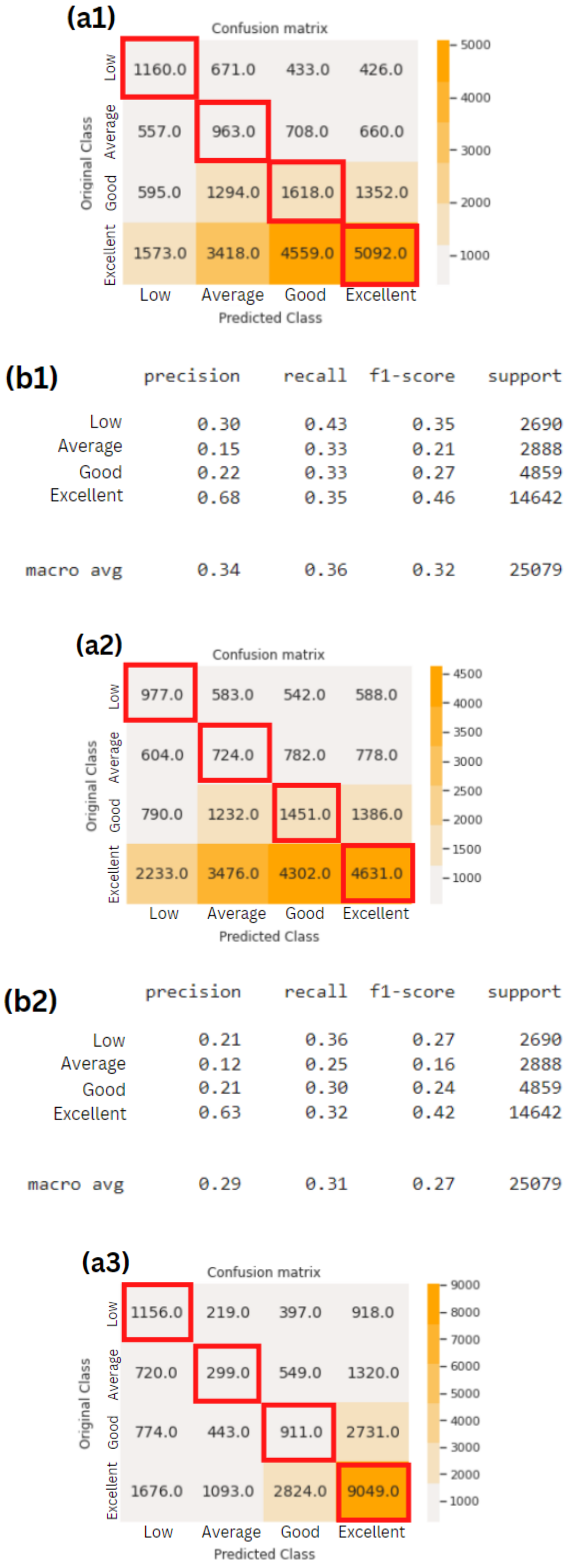
For data imbalance, undersampling (RandomUnder Sampler) and oversampling (SMOTE) are used to treat data and increase more in terms of recall before running through hyperparameter tuning to find the best parameter in each model.

After tuning, Table I shows the best imbalance treatment method with the best hyperparameter tuning to archive both precision and recall.

TABLE I. 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
K-NN	Random UnderSampler	n_neighbors	27
LR	Random UnderSampler	C	0.1

Using hyper-parameters from Table I, performance of all ML models in terms of confusion matrix, precision, recall and F1 score of every class are presented in Fig. 10. Fig. a1, a2 and a3 show a confusion matrix of the RF, K-NN and LR models, respectively. Fig. b1, b2 and b3 show precision, recall and F1 score of each class of RF, K-NN and LR models, respectively.

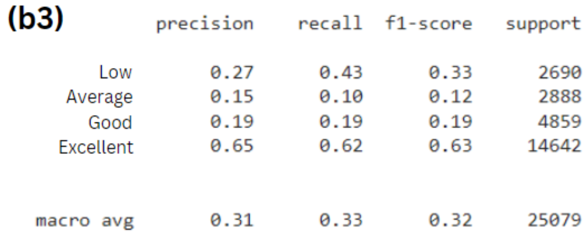


(b1)

```
           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
```

(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
```



```
              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
```

Fig. 10. Result of confusion matrix; (a1) RF (a2) K-NN (a3) LR and classification report; (b1) RF (b2) KNN (b3) LR after fine-tuning

From Fig. 10, 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 Fig. 10a1 and 10b1. This is confirmed with the box plot of product_rating feature displayed in Fig. 6a. 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.

V. Discussion and conclusion

This work aims to predict customer satisfaction score and understand the most affected feature that reflects customer satisfaction score. For better prediction, showing high relationship between features and customer satisfaction can help online retail stores or e-commerce improve their strategies, services, logistics, and customer support to meet customer expectations before and after ordering their products. In this research, we construct new features and feed them to all three different ML models in order to predict customer satisfaction score. Performance of all ML models are compared with the average score (baseline). E-commerce sales data is analyzed using RF, K-NN, and LR models, and the experimental results show all models yield superior prediction performance compared to the average score in almost every class. In Class ‘bad, ‘average’ and ‘Good’ yield average results due to their small sample size compared to ‘Excellent’ class. However, with our proposed features that are derived from the dataset, the RF model shows the best recall performance compared to others. In addition, RF exhibits the best performance in terms of confusion matrix, precision, recall, and F1-score in every class by using the undersampling method. Moreover, customer satisfaction is primarily influenced by two main types of features: product rating and delivery performance. This is confirmed by the correlation matrix as presented in Fig. 5 and confirmed by the box plot in Fig. 6. Future works will focus more on taking user comments as another type of features that the ML model will use to predict customer satisfaction score.

Acknowledgment

This work was partially supported by a research grant from the Graduate School, Srinakharinwirot University.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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 Aug, 2022).

[7] 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.

[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] 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.

[11] 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.

[12] 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.

[13] "Brazilian E-Commerce Public Dataset by Olist", Kaggle.com, 2020. [Online].Available:https://www.kaggle.com/olistbr/brazilian-ecommerce. (accessed Aug 2022).