Optimising e-commerce customer satisfaction with machine learning

Ann-Nee Wong* and Booma Pooalan Marikannan
School of Computing, Asia Pacific University of Technology & Innovation

*Corresponding author e-mail: wongannnee88@gmail.com

Abstract. Customer insights is the key to the success of e-commerce. Therefore, factors affecting customer satisfaction leading to product purchase and re-purchase should be studied extensively. This study intends to identify the key drivers that influence the satisfaction and the model which can predict the likelihood of customer satisfaction. The outcome would provide insights to prioritise factors that are significant, as well as to provide advice to a wide range of sellers. Four classification machine learning algorithms decision tree, random forest, artificial neural network and support vector machine are evaluated to classify customer satisfaction based on a 3-year historical data from an e-commerce retailer. There were a few challenges with the dataset, such as imbalanced, skewed and missing. Data pre-processing was conducted, and different techniques were evaluated. Of the algorithms evaluated, the best result is achieved by Random Forest with the highest accuracy and reasonable processing time. Meeting the estimated delivery date and the number of days taken to deliver an order is found to be the top two important factors affecting customer satisfaction.

Index Terms. E-commerce, Machine Learning, Customer Satisfaction, Predictive Modelling

1. Introduction
Retail e-commerce sales has steadily contributed to the global retail scene. This area continues be attractive with strong projected growth worldwide, driven by technological connectivity and maturity in the consumer behaviour. In 2020, e-commerce accounts for 15.5% of the global retail sales [1]. As the emerging markets grow in its global economic importance, the North America’s and Europe’s share of global e-commerce sales are decreasing [1]. Retailers have started reaching out to consumers in emerging markets like China, Russia and Brazil which hold the top ten positions in terms of projected sales of billions of USD in 2019 [2].

Consumer behaviour and expectations vary across geography as multiple factors affect the purchase, re-purchase or return of a product, ranging from the product features, inventory, logistics and customer support [3, 4, 5]. In Brazil, the e-commerce market is challenging due to customer uncertainty in the security of payments, fulfilment of deliveries, and high cross border taxes, all providing the advantage to the local retailers [6, 7, 8]. Therefore, the application of machine learning will enable retailers to overcome the challenges by learning more about the customers, listening to
what customer has to say, improving product recommendations, price and demand forecasting, and enhancing customer services [9, 10, 11]. This study intends to identify the key drivers that influence the satisfaction and predict the likelihood of the e-commerce customer satisfaction in Brazil using machine learning algorithms.

2. Related works
Classification is a popular two-step machine learning process which starts from training a model with labelled target variable in a historical dataset and predicting the target variable of a given dataset using the model. A study on the applications of classification algorithms in e-commerce shows that it can be applied in a wide range of predictions.

eBay researched algorithms from Naïve Bayes (NV), to Logistic Regression (LR), Decision Tree (DT), Random Forest (RF) and Gradient Boosting (GB) to predict the user’s click and purchase propensity using product features like price, conduct, format, title, and popularity [12]. They found that GB is able to closely predict the top 5 items that has the highest user click and purchase measured by the Area Under the Curve (AUC) and Normalised Discounted Cumulative Gain (NDCG) metrics, both measurements normally used for recommendation systems.

Many studies found that RF performs the best amongst other algorithms. Gender classification on micro-blogging sites were studied by classifying the emoticons, textual information using natural language processing, and emotional punctuations [13]. In this scenario, the RF outperforms NV, AdaBoost (AB) and Support Vector Machine (SVM) with the highest F1-score of the prediction. Classification algorithms were used to predict the shopping platform which users will use the next time they make a purchase by analysing temporal, user profile, demographics and loyalty features using RF, NV, SVM, and Long Short-Term Memory Network (LSTM) [14]. Again, RF obtained the highest accuracy, precision, recall, and F1-score. A study of repeat buyer prediction to identify buyers with the potential to purchase more products was carried using GB, RF, and XGBoost using transaction data, transaction history and sample promotion information [15]. RF showed the highest AUC score.

In a different scenario, machine learning is used to improve the effectiveness of promotion campaigns by identifying customers who will purchase a product after receiving the free samples. Various machine learning algorithms were evaluated, from LR to DT, SVM, multiple discriminant analysis (MDA) and Neural Network (NN). SVM showed the highest accuracy [16]. Classification of the e-commerce merchants was studied using their websites information by mining the text available in the homepage, first level and all pages with different natural language pre-processing methods were studied. Between six different algorithms of DT, NV, LR, SVM, k-Nearest Neighbour (kNN) and Multilayer Perceptron, SVM showed the highest F1 measure as compared to the rest [17]. Different algorithms showed superior accuracy, precision, recall or F1 measures in different applications, and therefore, it is necessary to select the best algorithm to predict customer satisfaction in an e-commerce scenario.

3. Materials and Methods

3.1. Dataset
The study is conducted using a 3-year data from the “Brazilian E-Commerce Public Dataset by Olist” with 112,000 orders [18]. The dataset was contained in 8 tables containing information on the order, delivery, customer, seller, payment, product, language translation, and order review. Six (6) unique identifiers were used to merge the tables into a data frame (Table 1). Two new features defining the efficiency of delivery were created. The target
variable is the customer’s review_score for each order_id in a 5-point likert scale that is transformed into a 2-level satisfaction feature of “yes” representing the rating of 3,4,5 and “no” representing rating of 1 and 2.

Table 1. Variables and descriptions

| Variable                  | Description                                      |
|---------------------------|--------------------------------------------------|
| customer_id               | id of the customer                               |
| order_id                  | id of the order                                  |
| seller_id                 | id of the seller                                 |
| product_id                | id of the product                                |
| order_qty                 | qty of the order                                 |
| shipping_limit_date       | date of shipping limit                           |
| price                     | price of the product                             |
| freight_value             | value of the freight                             |
| order_status              | status of the order                              |
| delivery_performance      | date_delivered - date_estimated                  |
| purchase_delivery_days    | date_delivered - date_purchased                  |
| product_name_length       | length of the product_name                       |
| product_description_length| website product description length               |
| product_photos_qty        | website product photo quantity                   |
| product_weight_g          | product size weight                              |
| product_length_cm         | product size length                              |
| product_height_cm         | product size height                              |
| product_width_cm          | product size width                               |
| seller_zip_code_prefix    | seller address zip code                          |
| seller_city               | seller address city                              |
| seller_state              | seller address state                             |
| satisfaction              | customer satisfaction                            |
| count_pay_sequence        | number of payments                               |
| mode_pay_type             | mode of payment                                  |
| sum_pay_inst              | total installment                                |
| sum_pay_value             | total value paid                                 |
| customer_unique_id        | id of customer by order                          |
| customer_zip_code_prefix  | customer address zip code                        |
| customer_city             | customer address city                            |
| customer_state            | customer address state                           |
| product_category_combine  | combination of the product_category              |

3.2. Methodology

The data mining methodology conducted in R programming is shown in Figure 1. The model was trained using the Decision Tree, Random Forest, Support Vector Machine and Artificial Neural Network algorithms. The algorithms were compared in terms of accuracy, sensitivity, specificity, F1-score and computation time. A comparison of the effect of feature selection, imbalanced data treatment and skewed data treatment was conducted. 50% of the dataset was used to train the model to reduce the computational time. The dataset then was split into 70:30 training and test data. All studies were performed on an Intel Core i5 CPU (2.3 GHz) with an 8 GB memory.
3.3. Model training
Machine learning methods were applied to train models to predict customer satisfaction. The following is a brief description of the algorithms used and the optimisation that followed. Decision Tree is similar to a tree-like flow-chart that starts from a root-node, then decision nodes that require choices to be made based on an attribute [19, 20]. In this study, the decision tree classification is conducted using the ‘rpart’ package. Random Forest builds multiple decision trees that are trained using bagging method and merges them together to obtain a more accurate and stable prediction [20]. Random Forest classification is conducted using the ‘randomForest’ package with a default of 5 times cross-validation, and an evaluation of the effect to computational time and accuracy when the cross-validation is reduced to 1 time.

The Support vector machines (SVM) creates a boundary, known as hyperplane, to partition data into groups of similar class [19, 20]. SVM classification is conducted using the ‘e1071’ package. During the training stage, the SVM parameters were tuned using the ‘tune.svm’ function in the ‘e1074’ package. It identifies the best parameter by optimizing the model over a specified range. Artificial neural network (NN) models the relationship between the input variables in the input layer and the target variables in the output layer by assigning weights to each input variable that contributes to activation functions f(x) in the hidden layer [20]. Artificial neural network classification is conducted using the ‘neuralnet’ package. NN requires the variables to be normalised to increase the computation speed. The min-max normalisation technique was applied using the base function ‘apply’. The algorithm was optimised by changing the stepmax and threshold.

4. Results and Discussion
Four algorithms of Decision Tree (DT), Random Forest (RF), Artificial neural network (NN) and Support Vector Machine (SVM) with different feature selections, skewed data treatment and imbalanced data treatment were evaluated. During the model training, factors that affect the customer satisfaction is extracted from the model training information and discussed.
4.1. Feature creation, selection and its effect

Two new features were created to represent the efficiency of delivery that was identified as one of the key challenges of the e-commerce industry in Brazil:

i. Delivery_performance which is the number of days the actual delivery date exceeded the estimated delivery date

ii. Purchase_delivery_days which is the number of days taken for the actual delivery from the date of purchase

The variable importance tested using the decision tree algorithm (Table 2) showed that both features created were amongst the top 5 important features, while the rest of the features showed weaker importance.

| No | Feature                   | Importance |
|----|---------------------------|------------|
| 1  | delivery_performance      | 1408.5     |
| 2  | purchase_delivery_days    | 904.9      |
| 3  | order_qty                 | 192.2      |
| 4  | sum_pay_value             | 98.1       |
| 5  | customer_state            | 72.8       |

Four algorithms DT, RF, NN and SVM were trained using all 20 features, and compared to training using the top 5 important features. The performance of accuracy, sensitivity, specificity, F1-score and computational time were evaluated (Table 3). In terms of accuracy, sensitivity, specificity, F1 score, algorithms trained with the 5 important features performed similar to the algorithms trained with 20 features. However, the computation time has reduced significantly with less features. The training of the NN and SVM algorithms with 20 features were unable to be executed within 12 hours. This confirms that it is possible to maintain the accuracy of the model while making the model less costly computationally with the right feature selection.

| Algorithm | Treat imbalanced dataset | No. of features | Training computation time (s) | Predicting computation time (s) | Accuracy (%) | Sensitivity (%) | Specificity (%) | F1 score |
|-----------|--------------------------|-----------------|-------------------------------|-------------------------------|---------------|----------------|----------------|----------|
| DT        | None                     | 20              | <1                            | 87.2                          | 98.3          | 25.3           | 0.93           |
| DT        | None                     | 5               | 1                             | 87.2                          | 98.3          | 25.3           | 0.93           |
| RF        | None                     | 20              | 1875                         | 87.5                          | 97.9          | 29.2           | 0.93           |
| RF        | None                     | 5               | 1772                         | 87.5                          | 97.9          | 29.6           | 0.93           |
| NN        | None                     | 20              | >43200                       | -                             | -             | -              | -              |
| NN        | None                     | 4               | 550                          | 87.3                          | 98.4          | 25.1           | 0.93           |
| SVM       | None                     | 20              | >43200                       | -                             | -             | -              | -              |
| SVM       | None                     | 5               | 402                          | 87.1                          | 98.7          | 22.5           | 0.93           |

4.2. Feature Transformation and its effect

A majority of the top 5 features have a positive skewed distribution. Normalising the dataset tries to give all the variables an equal weight and is known to help increase the computational speed of the training phase [19]. Therefore, the ‘bestNormalize’ package was used to calculate and perform the skewed data treatment. The function attempts a variety of normalising transformations for example log, square root, exponential, Box-Cox, Yeo-Johnson, and ordered quantile normalization to find the best technique with the lowest Pearson P test for
normality. Two of the variables were then normalized with log and ordered quantile normalization treatment and its effect on the performance of the algorithms evaluated. Models trained with and without data normalization has a similar accuracy, sensitivity, specificity and F1 score in this scenario (Table 4). However, the normalizing treatment was not effective in improving the computational time in this case.

### 4.3. Effect of imbalanced data treatment techniques

Imbalanced data is a common problem associated with classification tasks when the classes of the target variables are not of an equal number. When the dataset is under-represented, the class distribution is skewed. Therefore, when a model is trained with this imbalanced dataset, traditional classification algorithms is usually unable to accurately identify the minority class, represented by the specificity and F1 score. There are four common unbalanced data treatment techniques, under-sampling, over-sampling, Synthetic Minority Over-Sampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE). Studies has shown that SMOTE is generally a better technique as compared to other techniques [21, 22, 23].

The effect of the four unbalanced data treatment techniques, under-sampling, over-sampling, Synthetic Minority Over-Sampling Technique (SMOTE) and Random Over-Sampling Examples (ROSE) were compared using the decision tree algorithm (Table 5).

### Table 4. Comparison of performance with or without feature normalisation, using different algorithms

| Algorithm | Treat skewed dataset | No. of features | Training computation time (s) | Predicting computation time (s) | Accuracy (%) | Sensitivity % Predict Positive | Specificity % Detect Negative | F1 score |
|-----------|----------------------|----------------|-------------------------------|-------------------------------|--------------|-----------------------------|-----------------------------|----------|
| DT        | None                 | 5              | 1                             | <1                            | 87.2         | 98.3                        | 25.3                        | 0.93     |
| DT        | Yes                  | 5              | 1                             | <1                            | 87.2         | 98.3                        | 25.3                        | 0.93     |
| RF        | None                 | 5              | 1772                          | 2                             | 87.5         | 97.9                        | 29.6                        | 0.93     |
| RF        | Yes                  | 5              | 1818                          | 1                             | 87.4         | 97.9                        | 29.7                        | 0.93     |
| NN        | None                 | 4              | 550                           | 1                             | 87.3         | 98.4                        | 25.1                        | 0.93     |
| NN        | Yes                  | 4              | 350                           | 1                             | 87.3         | 98.4                        | 25.1                        | 0.93     |
| SVM       | None                 | 5              | 402                           | 17                            | 87.1         | 98.7                        | 22.5                        | 0.93     |
| SVM       | Yes                  | 5              | 504                           | 18                            | 87.0         | 98.7                        | 21.5                        | 0.93     |

### Table 5. Comparison of performance with or without imbalanced data treatment, using different algorithms

| Algorithm | Treat imbalanced dataset | No. of features | Training computation time (s) | Predicting computation time (s) | Accuracy (%) | Sensitivity % Predict Positive | Specificity % Detect Negative | F1 score |
|-----------|--------------------------|----------------|-------------------------------|-------------------------------|--------------|-----------------------------|-----------------------------|----------|
| DT        | None                     | 5              | 1                             | <1                            | 87.2         | 98.3                        | 25.3                        | 0.93     |
| DT        | SMOTE                    | 5              | 738                           | <1                            | 87.1         | 97.9                        | 26.4                        | 0.93     |
| DT        | Undersampling            | 5              | 20                            | <1                            | 87.1         | 97.9                        | 26.4                        | 0.93     |
| DT        | Oversampling             | 5              | 1074                          | <1                            | 87.1         | 97.9                        | 26.4                        | 0.93     |
| DT        | ROSE                     | 5              | 20                            | <1                            | 87.1         | 97.9                        | 26.4                        | 0.93     |

All four techniques were not able to improve the specificity and F1 score of the model. In fact, with the imbalanced data, sensitivity which the ability to predict positive class is above 97%, while specificity which is the ability to predict the negative class is approximately 25% to 26%. However, computational was higher using imbalanced data treatments, especially the SMOTE and oversampling techniques.

### 4.4. Modelling for Customer satisfaction

Four algorithms, decision tree, random forest, artificial neural network and support vector machine were studied. The performance of these algorithms consistently produced an accuracy of a range of 87.0% to 87.5%, even with various data pre-processing methods and feature engineering (Table 2,3,4). It shows all algorithms perform quite equally in its prediction with the given set of input variables and observations.
Comparing the performance between the four different algorithms, the RF algorithm has the highest accuracy and specificity as compared to DT, SVM and NN (Table 6). However, RF has a long training computation time. On the other hand, DT has the fastest training and prediction computation with reasonable accuracy.

Table 6. Tuning the number of cross-validation in Random Forest

| Algorithm   | Treat imbalanced dataset | No. of features | Training computation time (s) | Predicting computation time (s) | Accuracy (%) | Sensitivity (%) | Specificity (%) | F1 score |
|-------------|--------------------------|-----------------|-------------------------------|--------------------------------|--------------|-----------------|-----------------|---------|
| RF (cv=5)   | None                     | 5               | 1875                         | 2                              | 87.5         | 97.9            | 29.2            | 0.93    |
| RF (cv=4)   | None                     | 5               | 1593                         | 2                              | 87.5         | 97.9            | 29.6            | 0.93    |
| RF (cv=3)   | None                     | 5               | 1038                         | 2                              | 87.5         | 97.9            | 29.4            | 0.93    |
| RF (cv=2)   | None                     | 5               | 702                          | 2                              | 87.6         | 98.0            | 29.5            | 0.93    |
| RF (cv=1)   | None                     | 5               | 2                            | 2                              | 87.6         | 98.0            | 29.5            | 0.93    |

cv = cross validation

To improve the computation time, Random Forest was trained by reducing the numbers of cross-validations from the default of 5 to 1. Though the training computation time for one (1) time cross-validation was reduced significantly to 2 seconds, the accuracy and specificity maintained at above 87.5% and 29% respectively (Table 5). Therefore, decreasing the number of cross-validation does not significantly affect the accuracy and specificity of the trained Random Forest algorithm but reduces the computation time significantly.

4.5. Key drivers of customer satisfaction

In the DT and RF model training stage, variable importance information was obtained. The features and its importance were consistent for both algorithms (Table 1). Delivery_performance is the most important factor, and purchase_delivery_days came in the second. A scatter plot of delivery_performance vs purchase_delivery_day by satisfaction (Figure 2) clearly showed that Brazilian customers are not satisfied when the delivery was later than the estimated delivery date, represented by a positive value in the delivery_performance.

Figure 2. Scatter plot of two key numerical variables by satisfaction

Support Vector Machine and Artificial neural network are known as black box algorithms where the mechanism that transforms the input into the output is computed in an imaginary box without any intervention from the user (Lantz, 2015). However, based on the Artificial neural network weight for each factor (Figure 3), delivery_performance is weighted the highest.
In conclusion, the delivery performance has been identified by all algorithms used as the key factor in the Brazilian e-commerce setting and should be given a priority.

5. Conclusions
Based on the multiple attempts to treat the imbalanced and skewed data and feature selection based on variable importance, in general, the accuracy of the trained algorithms seems to be consistent around 87.0% to 87.5%, and the specificity is around 21.5% to 29.7%. The range of accuracy found in this study is consistent to the range of accuracy found in the related works from 75% to 99%.

Random Forest has the highest accuracy, sensitivity, specificity performance as compared to Decision Tree, Support Vector Machine and Artificial Neural Network, even with low number of cross-validation. Alternatively, Decision Tree is a fast computation algorithm that has slightly lower accuracy but is able to respond in seconds. In the implementation environment where data is big and speed is of essence, the data scientist will have to tune the parameter and balance between accuracy and computation speed.

6. Future works
Further study to improve the specificity is required as the prediction of customer dissatisfaction is an important criterion. One suggestion for further study is to incorporate the non-structured data, for example a customer’s comments on the review message, which could shed light on the customer’s sentiment, score and magnitude. These inputs can be used to build an enhanced classification model.

References
[1] “eCommerce - worldwide | Statista Market Forecast”, Statista, 2020. [Online]. Available: https://www.statista.com/outlook/243/100/e-commerce/worldwide. [Accessed: 8 March 2020].
[2] "Global Ecommerce Statistics and Trends to Launch Your Business Beyond Borders", Enterprise Ecommerce Blog - Enterprise Business Marketing, News, Tips & More, 2020. [Online]. Available: https://www.shopify.com/enterprise/global-ecommerce-statistics. [Accessed: 8 March 2020].
[3] S. Rose, N. Hair and M. Clark, “Online Customer Experience: A Review of the Business-to-Consumer Online Purchase Context”, International Journal of Management Reviews, vol. 13, no. 1, pp. 24-39, 2011, available: 10.1111/j.1468-2370.2010.00280.x.
[4] D. Nguyen, S. de Leeuw and W. Dullaert, "Consumer Behaviour and Order Fulfilment in Online Retailing: A Systematic Review", International Journal of Management Reviews, vol. 20, no. 2, pp. 255-276, 2016, available: 10.1111/ijmr.12129.

[5] H. Ceribeli, H. Tamashiro and E. Merlo, "Online flow and e-satisfaction in high involvement purchasing processes", BASE, vol. 14, no. 1, p. 16-29, 2017, available: 10.4013/base.2017.141.02.

[6] Unido.org, 2020. [Online]. Available: https://www.unido.org/sites/default/files/2017-10/WP_14.pdf. [Accessed: 8 March 2020].

[7] "Brazil – Lucrative but Challenging E-commerce Industry", EOS Intelligence - Powering Informed Decision-Making, 2020. [Online]. Available: https://www.eos-intelligence.com/permissions/consumer-goods-retail/brazil-lucrative-but-challenging-e-commerce-industry. [Accessed: 8 March 2020].

[8] "Brazil Commercial Guide | International Trade Administration", Export.gov, 2020. [Online]. Available: https://www.export.gov/article?id=Brazil-e-Commerce. [Accessed: 8 March 2020].

[9] A. Eisenberg, "20 Applications for Artificial Intelligence in Ecommerce [2019 Edition] - Ignite Ltd.", Ignite Ltd., 2020. [Online]. Available: https://igniteoutsourcing.com/ecommerce/artificial-intelligence-ecommerce. [Accessed: 8 March 2020].

[10] "How AI is revolutionizing e-commerce | Smart Insights", Smart Insights, 2020. [Online]. Available: https://www.smartinsights.com/ecommerce/ecommerce-strategy/ai-revolutionizing-ecommerce. [Accessed: 8 March 2020].

[11] "Use-cases of Machine Learning in E-Commerce | CloudsLab Blog", CloudsLab Blog, 2020. [Online]. Available: https://cloudslab.com/blog/use-cases-machine-learning-e-commerce/. [Accessed: 8 March 2020].

[12] Y. M. Brovman, M. Jacob, N. Srinivasan, S. Neola, D. Galron, R. Snyder, and P. Wang, "Optimizing Similar Item Recommendations in a Semi-structured Marketplace to Maximize Conversion" in Proc. 10th ACM Conference on Recommender Systems (RecSys ’16), New York, NY, USA, 2016, pp. 199–202, doi:https://doi.org/10.1145/2959100.2959166.

[13] Y. Yu and T. Yao, “Gender Classification of Chinese Weibo Users” in Proc. 2017 International Conference on E-commerce, E-Business and E-Government (ICEEG 2017), New York, NY, USA, 2017, pp. 5–8, doi:https://doi.org/10.1145/3108421.3108423.

[14] H. Huang, B. Zhao, H. Zhao, Z. Zhuang, Z. Wang, X. Yao, X. Wang, H. Jin, and X. Fu, “A Cross-Platform Consumer Behavior Analysis of Large-Scale Mobile Shopping Data” in Proc. 2018 World Wide Web Conference (WWW ’18), Republic and Canton of Geneva, CHE, 2018, pp. 1785–1794, doi:https://doi.org/10.1145/3178876.3186169.

[15] T. Charanasomboon and W. Vivanon. “A Comparative Study of Repeat Buyer Prediction: Kaggle Acquired Value Shopper Case Study” in Proc. 2019 2nd International Conference on Information Science and Systems (ICISS 2019), New York, NY, USA, 2019, pp. 306–310, doi:https://doi.org/10.1145/3322645.3322681.

[16] H.R. Won, M.J. Kim, and H. Ahn, “A Machine Learning-based Customer Classification Model for Effective Online Free Sample Promotions”, The Journal of Information Systems, vol. 27, no. 3, pp. 63–80, Sep. 2018.

[17] G.T. Shahid, R. Mahendra, and I. Budi, “E-Commerce Merchant Classification using Website Information” in Proc. 9th International Conference on Web Intelligence, Mining and Semantics (WIMS2019), New York, NY, USA, 2019, Article No. 5, pp. 1-10, doi:https://doi.org/10.1145/3326467.3326486.

[18] "Brazilian E-commerce Public Dataset by Olist", Kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce. [Accessed: 8 March 2020].

[19] J.W. Han, M. Kamber, and J. Pei, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann Publishers, Waltham, 2011.

[20] B. Lantz, “Machine Learning with R”, Birmingham, Packt Publishing, 2015.

[21] P. Branco, L. Torgo and R. Ribeiro, "A Survey of Predictive Modeling on Imbalanced Domains", ACM Computing Surveys, vol. 49, no. 2, pp. 1-50, 2016, available: 10.1145/2907070.

[22] C. Tantithamthavorn, A. Hassan and K. Matsumoto, "The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models", IEEE Transactions on Software Engineering, pp. 1-1, 2018, available: 10.1109/tse.2018.2876537.

[23] Y. Zhao, Z. S.Y. Wong, and K. L. Tsui, “A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events’ Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection,” Journal of Healthcare Engineering, vol. 2018, pp. 1–11, 2018.