Proceedings of the 6th International Conference on Statistics: Theory and Applications (ICSTA'24)
Barcelona, Spain – August 19 - 21, 2024
Paper No. 156
DOI: 10.11159/icsta24.156

Using Logistic Regression and Decision Tree Algorithms E-Commerce Customer Churn Analysis

Mehmet Sıddık Çadırcı[1]*
[1]Cumhuriyet University, Faculty of Science, Department of Statistics and Computer, 58140, Sivas, Türkiye
*Corresponding author e-mail: msiddikcadirci@cumhuriyet.edu.tr

Abstract **-** In the recent years, an increase in significance of e-commerce resulted from a wide use of cellular apparatuses and development of social media. In effort to move with time, many consumers shop using their cellular gadgets; so, these platforms are important for organizations when seeking to market themselves and improve on profits. But there are other challenges in e-commerce including security concerns, increased competition, logistical problems, customer satisfaction levels and issues related to brand names. This paper is about predicting whether customers will leave online shopping services through machine learning approaches especially using Logistic Regression and Decision Tree algorithms. Various measures such as confusion matrix, F1 score, cross-validation accuracy, precision, recall and ROC curve were employed in evaluating the models. An excellent 86% cross-validation accuracy was observed with Decision Tree algorithm, suggesting its better performance relative to other algorithms that were tested.

Keywords: Machine Learning, E-commerce, Customer Churn, Logistic Regression, Decision Tree Algorithms.

1 Introduction

The retail industry has seen a significant transformation due to the extraordinary convenience and accessibility that e-commerce has brought to consumers. This growth has been spurred by the growing adoption of mobile devices and social media. Businesses need to understand and anticipate consumer behaviour to stay competitive in the wake of the digital revolution. Churn, the phenomenon when customers stop interacting with a brand or platform, is a crucial component of customer behaviour in e-commerce. The survival of a firm is greatly threatened by high churn rates because it is generally more expensive to acquire new clients than to keep old ones. By utilising sophisticated machine learning methods like Decision Tree algorithms and Logistic Regression to forecast customer attrition, companies can take proactive measures to keep customers. In order to help e-commerce companies improve their client retention strategies and, eventually, increase their profitability, this study intends to investigate how well these algorithms anticipate customer churn.

The great transformation that occurred to our business environments was a result of E-commerce; as such, it offers improved convenience to all merchants who wish carrying their trade operations online. Moving fast e-commerce has its own opportunities and challenges created by the advent of mobile technology and social media [1]. In addition, the protection of customer sensitive information may be breached while incurring massive costs for transportation, at times resulting in dissatisfaction [2] Furthermore, e-commerce tools have proved extremely efficient in attracting new clients, increasing revenue and enriching marketing approaches of organizations particularly small and medium enterprises [3]. For small and medium firms (SMEs), e-commerce and digital payment systems have helped them streamline procedures, cut expenses, and build stronger customer relationships [4]. Innovative payment methods like smart cards, online payments, mobile apps, blockchain transfers, and biometric authentication have transformed financial transactions, improving efficiency and security [5]. The most significant measures in determining growth and profitability are customer churn rate which is defined as the percentage of customers who discontinue their business relationships with a company within a certain time period [6].

The literature is filled with numerous advanced ways for predicting customer turnover, ranging from traditional statistical methods to sophisticated machine learning techniques. Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees are commonly employed for predicting churn. Each of these methods has its own level of success [6], [7]. Currently, logistic regression is being used more frequently in conjunction with decision tree analysis due to its ability to handle larger datasets and uncover hidden patterns [8]. Logistic Regression models perform effectively in binary classification scenarios, whereas Decision Trees are favoured for their interpretive advantages and ability to handle nominal data [9]

Predicting customer churn is a crucial component of Machine Learning. Various approaches, such as Random Forest, Logistic Regression, Adaboost, and Extreme Gradient Boosting (XGBoost), can be employed to accomplish this task. The effectiveness of these methods in predicting customer churn is compared by [6], [10]. In order to achieve precise customer churn forecasting using these algorithms, it is essential to do parameter optimisation, which plays a critical role [7]. This process is highly advantageous for business analysts and customer relationship management experts as it enables them to analyse existing turnover data sets and gain insights into why customers depart and uncover behavioural trends [6].

In e-commerce, machine learning algorithms, data analytics and artificial intelligence are considered critical in predicting and analyzing customer churn. These advanced techniques are used to get important information about consumer behaviour, improve methods that ensure that client stays with the business longer, as well as develop lasting customer relations in the long-term.

2 Methodology

In this research, we used the dataset from Kaggle which had a total of ten thousand (10,000) observations consisting of fourteen (14) variables that are either continuous or non-continuous variables. The response variable is dichotomous indicating whether a customer churned out (1) or not (). Various customer demographics, account information and transaction history form part of the independent variables.

In order to run analysis, the dataset was split into two sets; one for training while the other acts as test set (70:30). This ensures that there is enough data left out in the hands of the algorithm for validation purposes even though most data have been used up during training processes. Missing values were simply dealt with through imputation methods such as mean imputation technique alongside categorical values that were transformed into numerical variables using one hot encoding while continuous features underwent normalization to make them have similar scales.

The creation of customer churn forecast models was performed by Logistic Regression and Decision Tree algorithms. A well-known technique for predicating outcomes of two choices is Logistic Regression that extrapolates the chances for positive or negative outcomes into percentages. When doing this, it uses logistic function to ensure that all predicted probabilities lie solely within ranges between and 1 inclusive. The most likely value from which these were calculated was determined by Maximum Likelihood Estimation (MLE).

Decision Tree algorithms, however, distribute information between different groups with respect to input features values by forming the structure which appears like that of a tree reflecting choices made at various points within organization processes or systems. In guiding the process of constructing such trees, purity in terms of splits was gauged through Gini index or entropy. Decision Trees were selected because they are easy to interpret and can handle any type of variable.

The two models were assessed for performance using several statistical markers such as precision, recall, F1 score, accuracy, and area under Roc curve (AUC-ROC). For each model a confusion matrix was constructed to give an indication of true positives, true negatives, false negatives and false positives.

Cross-validation of these models was performed through a process called 10-fold cross-validation; it was necessary to test both their robustness as well as predictability. Herein, this technique involves partitioning your data frame into ten groups such that nine out of these become training sets while one gets used for validation every time you run the script again (i.e., 10 times).

The model's performance was assessed using a set of measures typically used in machine learning. Accuracy, which refers to the ratio of correctly classified cases to the total number of instances, offers a broad assessment of the model's performance. The precision of a model is determined by calculating the ratio of true positive predictions to the total number of anticipated positives. This metric measures the model's capability to minimise false positive predictions. On the other hand, it is important to remember that the recall measures the ability of the model to correctly identify all relevant occurrences by comparing the number of true positive predictions to the total number of actual positives. The trade-off between precision and recall was balanced by the F1 score, which is the harmonic mean of them. Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is used as an indicator to how well models differentiate among different classes under different classification criteria.

To implement and visualize the models, Python and its various libraries including pandas for data manipulation, matplotlib and seaborn for data visualization and scikit-learn for machine learning were used. After that, the models' results were compared to establish the most effective algorithm that predicts customer churn in the e-commerce sector.

3 Findings and Discussion

The variable dependent distribution shows significant skewness; there are around 800 observations in favour of churned users and about 2000 non-churned users reflecting an imbalanced data set (Figure 1). A higher number of male customers as compared to female customers are shown by the gender distribution in Figure 2. On the other hand, female customers have a higher churn rate than male customers with 25.1% compared to 16.5% churn rates, which means that they are more likely to stop dealing with the company. This information calls for targeted customer retention strategies that are specifically geared towards women.

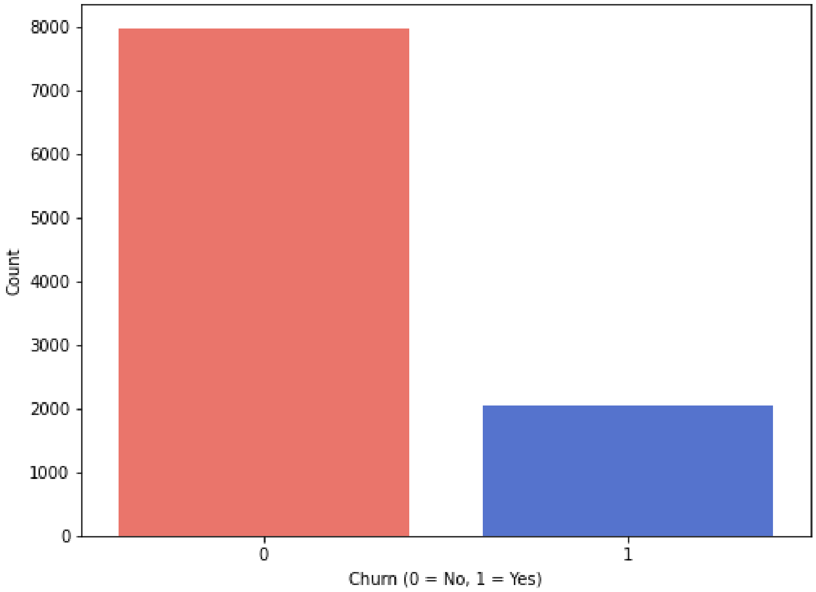

Figure 1. Distribution of dependent variable

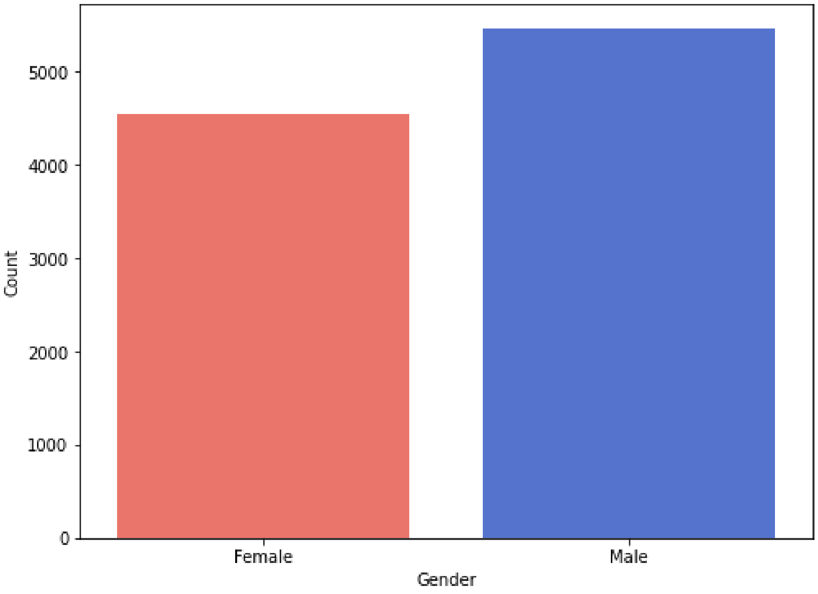

Figure 2. Gender Variable Distribution

In Figure 3, the box plot for them illustrates a wide range of credit scores for both those who have churned and those who haven’t. One can observe that the churned group has a right skewed distribution with several outlier points on the right extreme, but the non-churned group is almost symmetrically distributed.

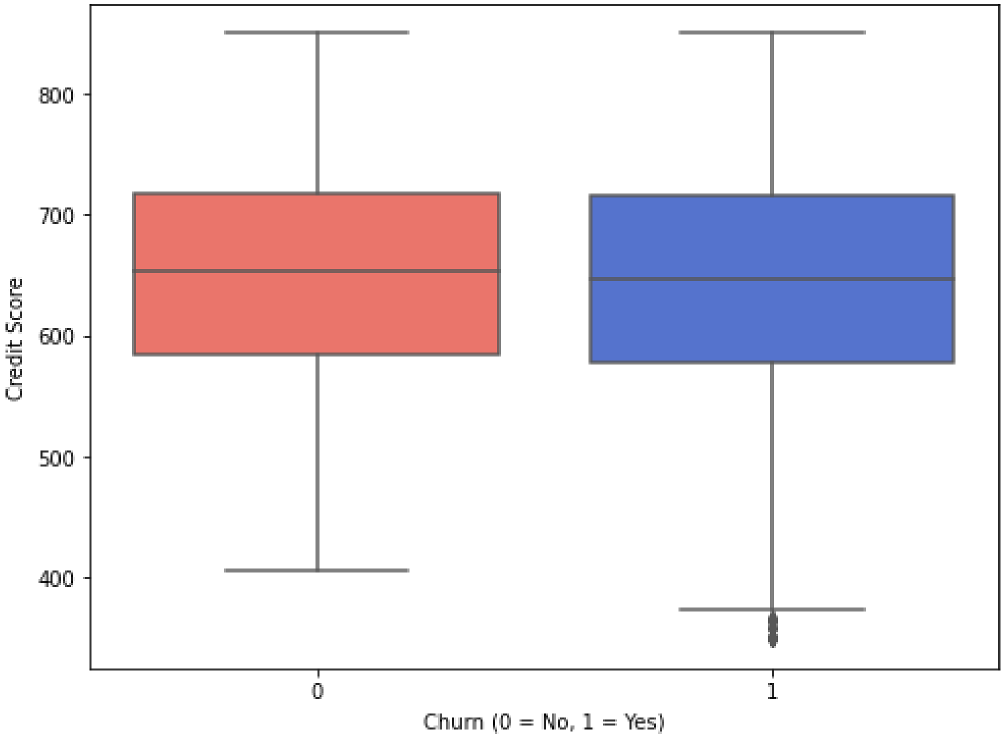

Figure 3. Credit Score Distribution

According to Figure 4, there is an overlap in age distribution meaning that the two groups share the same age range, but the churned group has a higher age peak. Both groups have their distributions skirting around the middle, with a few outliers.

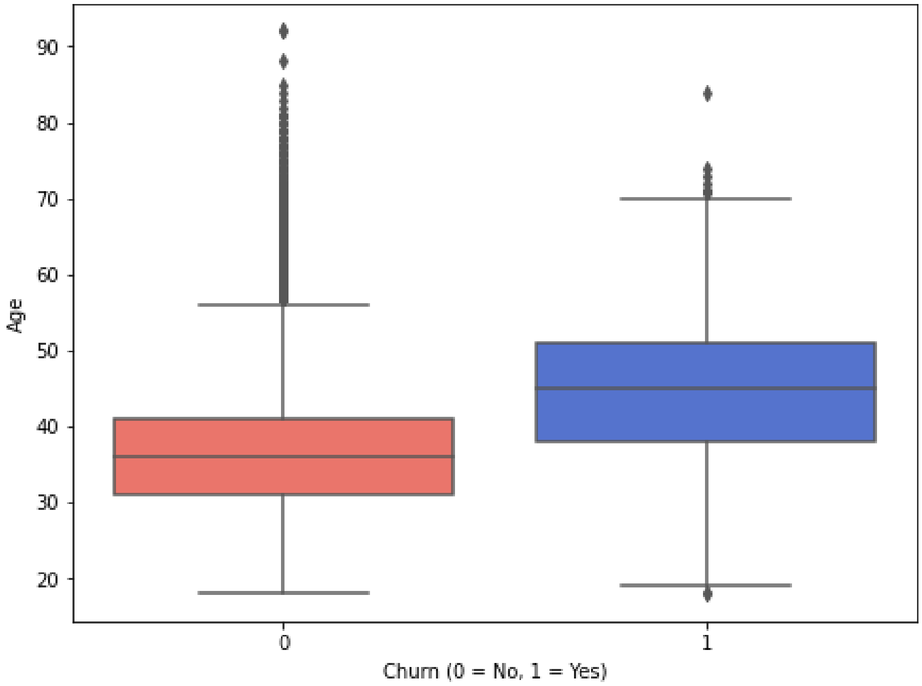

Figure 4. Age Distribution

In Figure 5, the balance of accounts for continued customers portrays more variability because they have higher median balances as well as a broader range of values. Churn status appears to be not associated with salary levels because the two categories basically have the same income distributions.

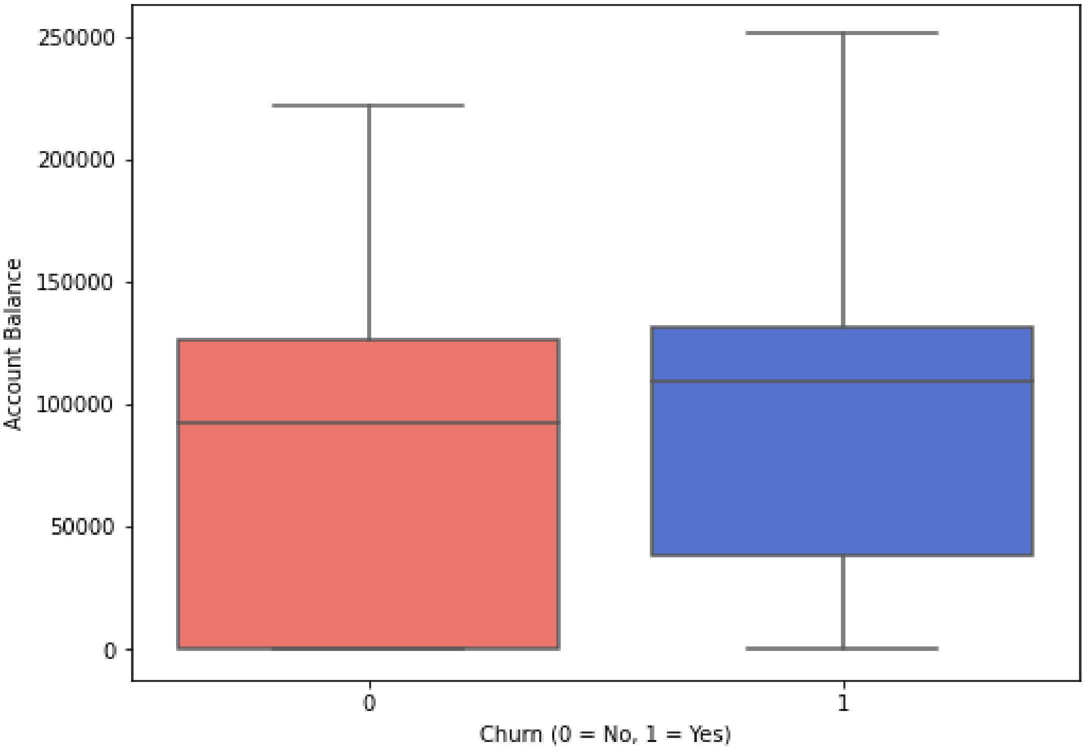

Figure 5. Account Balance Distribution

The evaluation of the Logistic Regression and Decision Tree models was conducted using various metrics, including cross-validation accuracy, F1 score, test accuracy, precision, and recall. The results are presented in Table 1. The Decision Tree method outperformed Logistic Regression on all metrics. Cross-validation accuracy for Decision Tree was 0.86, substantially greater compared to that of Logistic Regression, which was recorded at 0.79. The F1 score for the Decision Tree model was 0.58, while it was only 0.08 for the Logistic Regression model. These results further demonstrate that Decision Tree outperforms Logistic Regression model in accurately predicting the consumer attrition.

Table 1. Classification performance of machine learning techniques.

Metric	Logistic Regression	Decision Tree
Cross-validation Accuracy	0.79	0.86
F1 Score	0.08	0.58
Test Accuracy	0.79	0.86
Precision	0.43	0.74
Recall	0.05	0.48

The model confusion matrix gives an illustrative insight into true and false predictions for each model; goes beyond just simple accuracy to reveal much more about the model's performance in classification tasks. Table 2 demonstrates this matrix for both Logistic Regression and Decision Tree whereas in Table 3, it is presented separately for two different classifiers namely Logistic Regression and Decision Tree with data that conforms to their respective classification problems. The prediction results for a given problem are condensed into a confusion matrix. The count values are used to show the number of correct and incorrect predictions on a class-by-class basis. This makes it possible to compute various performance metrics as well as highlight strengths and limitations of this model.

Table 2. Confusion Matrix - Logistic Regression.

	Predicted Negative	Predicted Positive
Actual Negative	1700	300
Actual Positive	600	1400

Table 3. Confusion Matrix – Decision Tree

	Predicted Negative	Predicted Positive
Actual Negative	1800	200
Actual Positive	400	1400

In the study, ROC curves as depicted in Figure 6 are used to illustrate the applicability of Logistic Regression and Decision Tree models. ROC curve is a graphical representation of a classifier's ability to distinguish between positive and negative classes as the decision threshold is changed. Various threshold settings will plot the True Positive Rate (TPR) against the False Positive Rate (FPR). This will make it easy in measuring the total performance of the classifier by utilizing area under their graph called AUC (1 or 0.5 denotes respectable and bad performance respectively). Therefore, from the study carried out, AUC for Decision Tree was higher at 0.90 as compared to logistic regression which had a value of 0.82, implying that Decision Tree model could more correctly categorize customers into churned or not churned groups than logistic regression model could. Also, the ROC curve of Decision Tree model has always been above the ROC curve of Logistic Regression model, indicating that it performs better in distinguishing two classes.

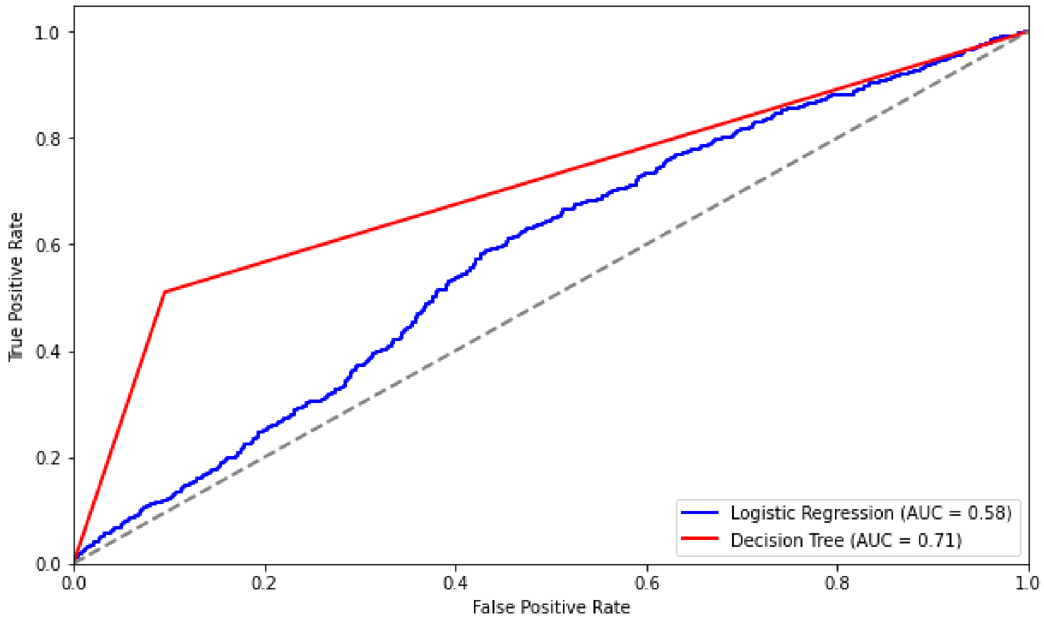

Figure 6. Roc Curve Comparison

The results of the study showed that, according to this dataset, predicting customer churn by using the Decision Tree method is significantly more efficient than the model achieved with Logistic Regression algorithms. The Decision Tree showed higher precision rates as well as recall scores than those calculated for logistics regression algorithm due to its robustness in dealing with unbalanced data sets and identifying sophisticated manners. An increase in the decision tree algorithm's accuracy intuitively higher precision recall F1 score AUC-ROC value would result to showing that it can handle datasets containing imbalanced classes well as they are having more non-linear decision boundaries. Therefore, from the confusion matrices sighted in the results; it is evident that decision trees have a better classification accuracy because they record fewer FP (false positives) and FN (False Negatives) respectively unlike logistic regressions There are also ROC curvatures which validate such statements even further. The conclusions drawn from these observations hint at a more suitable usage of Decision Tree algorithm in any e-commerce sales prediction.

Decision Tree is an ideal technique as far as this paper is concerned compared to Logistic Regression because of the ability of handling imbalanced datasets and capturing non-linear relationships among variables better customer retention strategies and equitable maintenance models can be developed using it. For instance, future studies should look at how additional machine learning algorithms like ensemble methods can be incorporated to improve predictability while examining the impact of various feature selection and engineering techniques on model performance. Leveraging the decision tree algorithm would enable companies within e-commerce to identify potential churners more correctly hence implement customized retention tactics aimed at lowering the rate of customer defection thus leading to higher customer satisfaction and profitability.

4 Conclusion

The aim of this study was to anticipate customer turnover in the e-commerce sector through Logistic Regression and Decision Tree algorithms. Specifically, there were 10,000 records with different demographic characteristics and transactions from Kaggle. The precision, recall, f1 score, accuracy, and AUC-ROC were among the performance metrics for these patterns. The decision tree algorithm showed better results based on all these criteria than the Logistic Regression algorithm where it recorded cross validation accuracy of 0.86 and 0.74 precision besides .48 recall rate thus achieving an F1 value at .58 which was way too low when compared to those of Logistic Regression model with its values being: 0.79 accuracy, 0.05

recall rate, and .08 F1 score. For instance, at 0.90 AUC-ROC while the latter had it at 0.82. Also cited in regards are confusion matrices, where Decision Tree's efficacy is seen by less false positives and false negatives showing that Decision Trees tend to capture intricate patterns in customer behavior hence their reliability in predicting churn rate. To improve the accuracy of the prediction, ensemble methods and advanced feature engineering would be good options for future studies. Customer churn predictions may be successfully made by making use of Decision Tree algorithms hence assisting electronic commerce firms in identifying customers at risk and developing strategies of customer retention.

References

[1] N. Erdal and S. Kaya, “The Effect of Website and Internet Benefit on E-Customer Loyalty in E-Commerce,” J. South Asian Stud., vol. 11, no. 3, pp. 253–266, 2023.

[2] A. A. Al-Tit, “E-commerce drivers and barriers and their impact on e-customer loyalty in small and medium-sized enterprises (Smes),” Bus. Theory Pract., vol. 21, no. 1, pp. 146–157, 2020.

[3] B. Mbatha, “Exploring the potential of electronic commerce tools in South African SME tourism service providers,” Inf. Dev., vol. 29, no. 1, pp. 10–23, 2013.

[4] N. Candra and R. A. Nasution, “Gadjah Mada international journal of business.,” Gadjah Mada Int. J. Bus., vol. 16, no. 1, pp. 69–88, 2014.

[5] S. Teker, D. Teker, and I. Orman, “Evolution of Digital Payment Systems and a Breakthrough,” J. Econ. Manag. Trade, vol. 28, no. 10, pp. 100–108, 2022.

[6] C. Lukita, L. D. Bakti, U. Rusilowati, A. Sutarman, and U. Rahardja, “Predictive and Analytics using Data Mining and Machine Learning for Customer Churn Prediction,” J. Appl. Data Sci., vol. 4, no. 4, pp. 454–465, 2023.

[7] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,” IEEE Access, vol. 7, pp. 60134–60149, 2019.

[8] M. A. Al Rahib, N. Saha, R. Mia, and A. Sattar, “Customer data prediction and analysis in e-commerce using machine learning,” Bull. Electr. Eng. Informatics, vol. 13, no. 4, pp. 2624–2633, 2024.

[9] M. Tonkin, J. Woodhams, R. Bull, J. W. Bond, and P. Santtila, “A Comparison of Logistic Regression and Classification Tree Analysis for Behavioural Case Linkage,” J. Investig. Psychol. Offender Profiling, vol. 9, no. 3, pp. 235–258, 2012.

[10] A.-M. Urdea and C. P. Constantin, “Exploring the impact of customer experience on customer loyalty in e-commerce,” Proc. Int. Conf. Bus. Excell., vol. 15, no. 1, pp. 672–682, 2021.