

Customer data prediction and analysis in e-commerce using machine learning

Md Abdullah Al Rahib, Nirjhor Saha, Raju Mia, Abdus Sattar

Department of Computer Science and Engineering, Faculty of Science and Information Technology, Daffodil International University, Dhaka, Bangladesh

Article Info

Article history:

Received Apr 7, 2023

Revised Oct 9, 2023

Accepted Feb 12, 2024

Keywords:

Customer annual spending

Customer churn

Data analysis

E-commerce

Machine learning

Performance metrics

Product on-time delivery

ABSTRACT

Customer churn is a major challenge faced by e-commerce companies, as it leads to loss of revenue and decreased customer loyalty. In recent years, for predicting and reducing client churn machine learning techniques are powerful tools. This research aims to explore the use of machine learning algorithms for predicting customer churn, annual spending, and product on-time delivery in e-commerce. The study first conducted a comprehensive review of the literature on customer churn in machine learning. The literature showed that customer churn has been predicted successfully using a variety of machine learning algorithms, including support vector machine (SVM), random forest, and decision tree in various industries. To address this gap in the literature, the study conducted an empirical analysis of customer churn in e-commerce using machine learning algorithms. The data were then pre-processed and analyzed utilizing machine learning techniques for prediction. According to the study's findings, machine learning algorithms are effective in predicting customer churn, and product on-time delivery in e-commerce. The best-performing algorithm SVM achieved an accuracy of 83.45% in predicting customer churn and 68.42% for product on-time delivery prediction.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Md Abdullah Al Rahib

Department of Computer Science and Engineering, Faculty of Science and Information Technology

Daffodil International University

Dhaka, Bangladesh

Email: abdullah15-12247@diu.edu.bd

1. INTRODUCTION

Now e-commerce industry has seen tremendous growth in recent years, with millions of consumers flocking to the internet to make purchases. However, retaining customers and preventing them from “churning” or switching to competitors is a major challenge for e-commerce businesses. Acquiring new customer is frequently more costly than keeping current ones, customer churn can significantly affect a company's income. Several research papers have been instrumental in advancing our understanding of customer churn prediction in e-commerce.

For instance, Matuszelański and Kopczewska [1] uses extreme gradient boosting and logistic regression models to predict churn by leveraging socio-geo-demographic data from census records. Xiahou and Harada [2] employs k-means clustering, logistic regression, and support vector machines (SVM) to address churn prediction, emphasizing the significance of data balancing and ensemble learning techniques, Pondel *et al.* [3] delves into the realm of big data, employing decision support and neural networks to tackle churn prediction for one-off customers. Also, several studies used datasets from Alibaba Cloud Tianchi platform [4], IBM Telco Customer Churn dataset, whereas others used data from e-commerce marketplaces

such as Kaggle. One study focused on decision support for predicting customer churn in the big data domain, while others focused on using ensemble learning techniques to deal with non-contractual customer data [5], and high-dimensional, unbalanced data. Additionally, a study was dedicated to predictive models using k-nearest neighbor (KNN), decision tree [6], [7], Naive Bayes, and logistic regression for predicting customer churn on e-commerce mobile customers. Also, some studies implemented deep learning techniques such as multilayer perceptron (MPL), recurrent layer recurrent neural network (RNN), artificial neural networks (ANN) [8]. Gan [9] focus on predicting customer churn using XGBoost and an innovative approach based on the recency, frequency, and monetary (RFM) model. Their work addresses data imbalance through methods like imbalance, random up sampling, synthetic minority oversampling technique (SMOTE), and SMOTE Tomek-link. Similarly, Li *et al.* [10] utilize the BG/NBD model in machine learning to predict customer churn, engaging in data pre-processing, parameter estimation, and individual customer forecast. Kędziora and Maksim [11] employ a combination of SVM and ANN to predict customer churn within the context of an insurance company. Chen *et al.* [12] delve into the realm of prediction models in machine learning, utilizing data preprocessing, customer value analysis, and performance measures to anticipate churn among active and lost customers were 67,849 and 1,321 respectively. Furthermore, Sharma *et al.* [13] explore predicting customer spending scores by employing data grouping, piecewise linear analysis, and adaptive spline techniques. In the realm of repurchase customer prediction, Liu *et al.* [14] incorporate logistic regression and decision tree based XG-Boost models to forecast repurchase behavior. Their paper introduces the principle of the XGBoost algorithm and employs stable volatility models to achieve predictive accuracy. With a dataset comprising multiple sample subsets totaling 70,000, they also explore a model fusion algorithm to combine individual model results effectively. Karim *et al.* [15] explore the domain of on-time delivery within e-commerce. They utilize basic and transactional data, business process analysis, and quality assessments to develop an on-time delivery improvement model. Li *et al.* [16] elucidate sales prediction for inventory optimization, employing machine learning algorithms to incorporate sales forecasting, sales volume analysis, and inventory optimization modules. Customer behavior prediction in e-commerce [17], who draw from diverse datasets, including online shopper intentions, Ta-feng, foursquare check-ins, JD, and apparel industry, to identify effective methods for predicting customer behavior. A comparative study by Yu *et al.* [18] who investigates customer churn prediction based on SVM forecasting, showcasing extended support vector machines (ESVM's) superior performance among different models (ESVM, ANN, decision tree, SVM), utilizing a substantial dataset of 50,000 samples and 100,000 test data sets, along with raw data, extract, load and transform (ELT), and data warehouse techniques. Churn prediction via the length, recency, frequency, and monetary (LRFM) model, as demonstrated by [19], underscores customer profiling, partial and total churn analysis techniques, and the efficacy of decision tree ensembles. Leveraging real data from an online store, their study accentuates the efficiency of decision tree ensembles over other algorithms. Employing gradient boost trees, Raeisi and Sajedi [20] enhance churn prediction accuracy within the context of an online food ordering service in Tehran. Further contributions address customer churn through data mining techniques [21], on-time delivery prediction using autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) models [22], e-commerce sales prediction using machine learning [23], e-commerce customer segmentation [24], and customer churn prediction via improved SMOTE and Adaboost algorithms [25].

Most research shows the customer spending score at shopping malls and customer lifetime value prediction in e-commerce [13], [26]. In this paper, we worked on customer annual spending prediction and analysis in e-commerce. On the other hand, there is research on on-time delivery in manufacturing companies, but in this research, we predict and analyze product on-time delivery in e-commerce [15]. Customer churn research is mostly focused on the telecommunication industry and some are on e-commerce. But in this research, we predict and analyze customer churn based on different parameters, and get good model accuracy where we got minimal difference of train data accuracy and test data accuracy to ensure no over-fitted model. In this research for customer churn, we achieved almost the same train data and test data accuracy.

This research study aims to explore these factors in the context of e-commerce and identify strategies for improving customer retention, on time delivery, and revenue prediction. To achieve this goal, the study will use a combination of quantitative and qualitative research methods, including data analysis with good accuracy. The study's findings will contribute to the development of effective strategies for reducing customer churn, improving on-time delivery, and predicting annual spending in e-commerce. The research will provide valuable insights into customer behavior and help e-commerce companies to tailor their marketing and sales strategies to meet their customer's needs and expectations. Also, this research will have significant implications for the e-commerce industry, as it will help companies to increase customer satisfaction and retention, improve delivery times, and boost revenue. Ultimately, this study will contribute to the growth and development of thee-commerce industry, which has become an integral part of our daily lives.

2. RESEARCH METHOD

The proposed method flowchart for customer data analysis in e-commerce sites which are customer churn, customer annual spending, and product on-time delivery are shown in Figure 1, which contains some steps. First of all, we collected three datasets. Then applied some preprocessing techniques on datasets. Then classified the customer churn dataset which contains churn samples or not churn samples, and also classified the product on-time delivery datasets which contains on-time delivery samples or on-time, not delivery samples. After that, using some machine learning algorithms for customer churn and on-time delivery prediction. After preprocessing the customer annual spending dataset, also use a machine learning algorithm which is linear regression for prediction. In the last step, used some classification metrics to compare the algorithm performance for all datasets and analyze the test data based on the dataset's parameters. Describe those steps in detail in the methodology subsections.

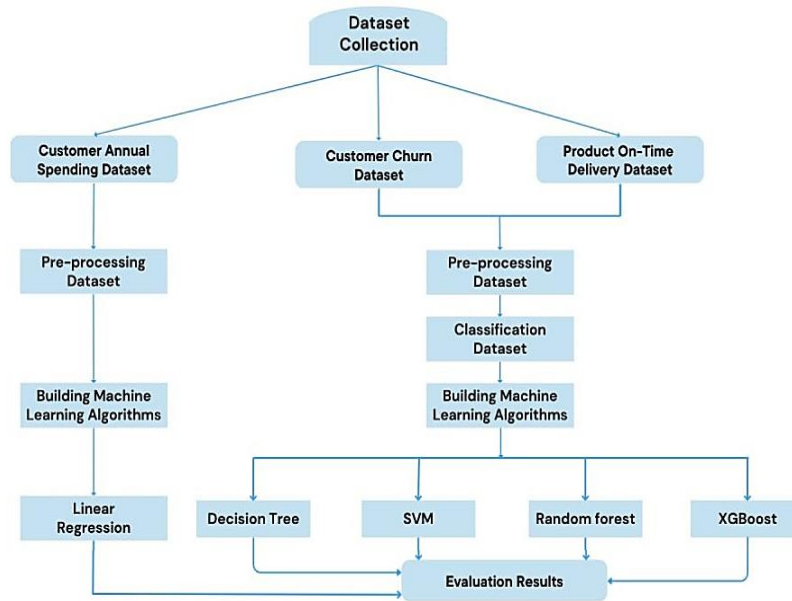


Figure 1. Proposed methodology

2.1. Data collection

In this research, we used three datasets which are collected from the Kaggle website to predict and analyze customer data. Those datasets are the customer churn dataset, customer annual spending based on device usage dataset, and product on-time delivery dataset which are contains many parameters, as shown in Tables 1 to 3.

Table 1. Customer churn dataset parameters, description, and the data types

Parameter	Description	Data type
Customer ID	Unique customer ID	Integer
Gender	Gender of customer	Object
Preferred order cat	Preferred order category of customer in last month	Object
Marital status	Marital status of customer	Object
Number of address	Total number of added addresses on particular customer	Integer float integer
Order count	Total number of orders has been places in last month	
Churn	Churn flag	

Table 2. Customer annual spending dataset parameters, description, and the data types

Parameter	Description	Data type
Avg. session length	Average usage session length of the customer	Float
Time on app	Daily spend time on this e-commerce mobile application	Float
Time on website	Daily spend time on this e-commerce website	Float
Length of membership	Membership length of the customer	Float
Yearly amount spent	Yearly amount spending money on this e-commerce	Float

Table 3. Product on-time delivery dataset parameters, description, and the data types

Parameter	Description	Data type
ID	Unique customer ID	Integer
Customer_care_calls	The number of calls made from enquiry of the shipment	Integer
Customer_rating	The company has rated from every customer	Integer integer
Cost_of_product	Cost of the product in US Dollars	Integer object
Prior_purchases	The number of prior purchase	Integer integer
Product_importance	Categorized the product in the various parameters	Integer
Discount_offered	Discount offered on that specific product	
Weight_in_gms	Weight of the product	
Reached.on.Tine_Y.N	Product reached on time or not	

The customer churn dataset contains 5630 samples and 5 parameters to classify those churned samples or not churned samples, as shown in Figure 2. The original churn dataset which is collected from the Kaggle website contains 18 parameters but, in this paper, we used 5 parameters to get good output using the minimal parameters. The annual spending dataset contains 448 samples and 4 parameters and the product on-time delivery dataset contains 10,999 samples and 7 parameters to classify those delivered samples or not delivered samples, as shown in Figure 3.

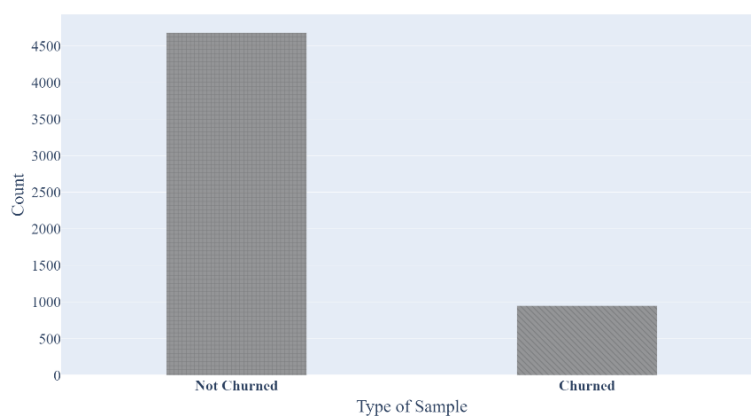


Figure 2. Customer churn dataset

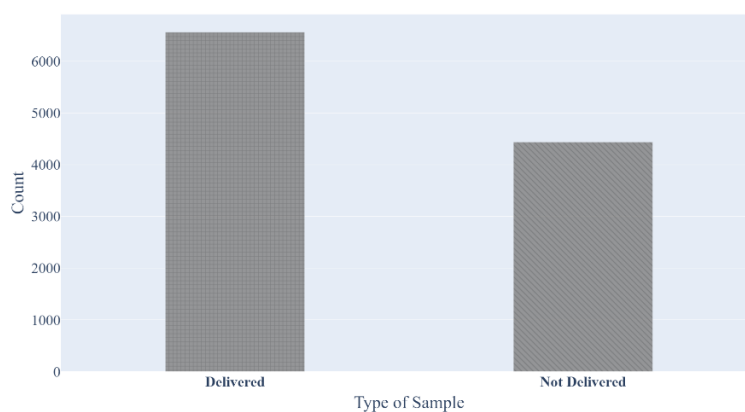


Figure 3. Product on-time delivery dataset

2.2. Pre-processing dataset

In this research, data pre-processing is very essential to make those datasets more reliable and consistent. Applied some data pre-processing techniques. To handle null or missing values, we used median value for those columns data where values are missing or null, and machine learning algorithms do not work with categorical features; it works only with numerical features. For this reason, we used one-hot encoding and label encoding to convert numerical data from categorical data on those datasets.

2.3. Machine learning models

Datasets are ready to fit into machine learning algorithms for the prediction and analysis of data after pre-processing those datasets. The train data and test data should be divided into two categories from the three datasets. For the customer churn and customer annual spending datasets split 75% data for training and 25% data for testing. And for product on-time delivery dataset split 70% data for training and 30% data for testing. In Table 4, we can see the sample count for all datasets, training, and testing. We utilized four well-known machine learning algorithms for customer churn and product on-time delivery prediction and analysis: decision tree, SVM, random forest, and XGBoost. We used linear regression algorithm for customer annual spending prediction and analysis. We'll see in the subsections, a general analysis of utilized algorithms and the parameters which we used.

Table 4. Number of samples in three datasets

Dataset	All dataset	Training	Testing
Customer churn	5630	4222	1408
Customer annual spending	448	336	112
Product on-time delivery	10999	7699	3300

2.4. Linear regression

A statistical technique known as linear regression is used to determine the connection between a dependent variable, such as a customer's annual spending on online shopping, and other independent variables, such as the average session length, and time spent on an app. It presumes that the dependent variable and the independent variable have a linear connection, and it seeks to identify the best-fit line that can forecast the dependent variable from the independent variable [13]. In our research, for the customer annual spending dataset where used the following parameters of linear regression: i) copy_X=True, ii) fit_intercept=True, iii) n_jobs=None, and iv) random_state=104.

2.5. Decision tree

Decision tree is utilized for regression and classification in machine learning. This algorithm divides the data into groups a particular characteristic is used to divide each partition, and the procedure is repeated until the subgroups are homogeneous. Decision trees are a common solution for decision-making issues because they are simple to understand and can classify customer churn and on-time product delivery [7]. For the customer churn dataset, we used the following decision tree parameters: i) criterion= 'entropy', ii) max_depth=6, and iii) random_state=104. And for the on-time product delivery dataset, we used the following decision tree parameters in our research: i) criterion= 'entropy', ii) max_depth=5, and iii) random_state=42.

2.6. eXtremegradient boosting

XGBoost is utilized for classification and regression problems which is an advanced and potent machine-learning method. It is a variation of the gradient boosting method that joins several decision trees to produce an effective learner that can forecast outcomes accurately [9]. It has been used effectively in several uses, including customer churn forecasts and on-time delivery prediction. In our experiment, the following parameters of XGBoost are used for customer churn dataset: i) use_label_encode=False, ii) eval_metric='mlogloss', and iii) random_state=104. And the following XGBoost parameters are used for the on-time product delivery dataset: i) criterion= 'entropy', ii) max_depth=5, and iii) random_state=42.

2.7. Support vector machine

SVM technique is utilized for regression or classification jobs. It can accurately manage challenging datasets like customer churn and on-time product delivery [25]. By transforming the incoming data into a higher dimensional space using kernel functions, SVM can also be used for non-linear classification problems [2], [18]. In this research, the following parameters of SVM are utilized for the customer churn dataset: i) kernel='rbf' and ii) random_state=104. And for the on-time product delivery dataset: i) gamma=0.00001, ii) C=10, iii) kernel='rbf', and iv) random_state=42.

2.8. Random forest

An ensemble learning method built on decision trees is called random forest (RF) [2]. The predictions from multiple decision trees are combined to produce a final estimate. It can manage big datasets like the on-time product delivery dataset and feature spaces with many dimensions. Additionally, it can manage incomplete data and resists overfitting. For the customer churn dataset, we utilized the following parameters of

RF: i) $n_estimators=100$ and ii) $random_state=1$. And the following parameters are used for the On-time delivery dataset: i) $n_estimators=100$, ii) $criterion= 'entropy'$, and iii) $random_state=1$.

3. RESULTS AND DISCUSSION

3.1. Machine learning predictive models

In this research, four machine learning algorithms are used on the customer churn dataset with particular parameters which are described in the earlier section, and the performance results of all models is shown in Table 5. To avoid overfitting issues, we will choose the algorithm that get low accuracy difference between train data and test data. If we analyze the performance metrics of the customer churn dataset. The SVM algorithm test data and train data accuracy difference is less than others and get good values of precision, recall, and F1-score. Also, get the highest test data accuracy which is 83.45%. So, the SVM algorithm gives the best output from those four algorithms.

Table 5. Performance metrics of all models on customer churn dataset

Algorithm	Test data				Train data			
	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score
XGBoost	83.38	0.80	0.83	0.77	83.87	0.83	0.84	0.78
DT	83.38	0.80	0.83	0.77	83.87	0.83	0.84	0.78
RF	82.10	0.79	0.82	0.80	87.70	0.87	0.88	0.86
SVM	83.45	0.84	0.83	0.76	83.87	0.86	0.84	0.77

For the customer annual spending dataset, we used the linear regression algorithm, and the performance results of the algorithm is shown in Table 6. The value of $R2_score$ for test data is 0.978, and the value of $R2_score$ for train data is 0.985 where both are close to 1, and also get moderate values of MAE, MSE, and RMSC. So, the linear regression algorithm is regarded to be a suitable fit for the customer annual spending dataset.

Table 6. Performance results of customer annual spending dataset

Algorithm	Test data				Train data			
	R2-score	MAE	MSE	RMSC	R2-score	MAE	MSE	RMSC
Linear regression	0.9782	8.45	117.67	10.84	0.9859	7.58	89.26	9.44

In this experiment, the product on-time delivery dataset is used to train four machine learning algorithms, each of which is given specific parameters that are explained in the previous section. Table 7 displays the performance metrics of all models. We get 68.48% test data accuracy using the decision tree algorithm and 68.42% test data accuracy using the SVM algorithm which is the closed value. But the difference between test data and train data accuracy is smaller in the SVM algorithm than in others algorithms. Obtaining good precision, recall, and F1_score values as well. Therefore, of these four algorithms, the SVM algorithm gives the finest output for the on-time product delivery dataset to encourage us to deploy it into the e-commerce site for prediction and analysis.

Table 7. Performance metrics of all models on on-time delivery dataset

Algorithm	Test data				Train data			
	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score
XGBoost	67.09	0.72	0.67	0.67	76.03	0.72	0.67	0.67
DT	68.48	0.77	0.68	0.68	69.37	0.78	0.69	0.69
RF	66.93	0.68	0.67	0.67	100	1.00	1.00	1.00
SVM	68.42	0.79	0.68	0.68	68.67	0.80	0.69	0.68

3.2. Customer data analysis

In this research, our main aim is to analyze customer-predicted data to determine the causes of customer churn, customer annual spending, and late product deliveries on e-commerce sites. We then aim to improve customer service to reduce customer churn, increase customer annual spending, and increase on-time product deliveries on e-commerce sites, all of which are beneficial to both customers and the e-commerce industry.

The customer churn dataset predicted result analysis of test data based on different parameters is shown in Figure 4. The percentage of the customer who will churn and the percentage of the customer who will not churn are shown in the pie chart in Figure 4(a). If analyzing Figure 4(b), we can see that the male customer is more than the female customer but the ratio of churn is almost the same. On the other hand, in Figure 4(c) the number of customers whose preferred categories are laptop and accessory is highest but the ratio of churn whose preferred categories are mobile & accessory is highest. When analyzing Figure 4(d), the highest number of customers who are married but the highest ratio of churn who are single.

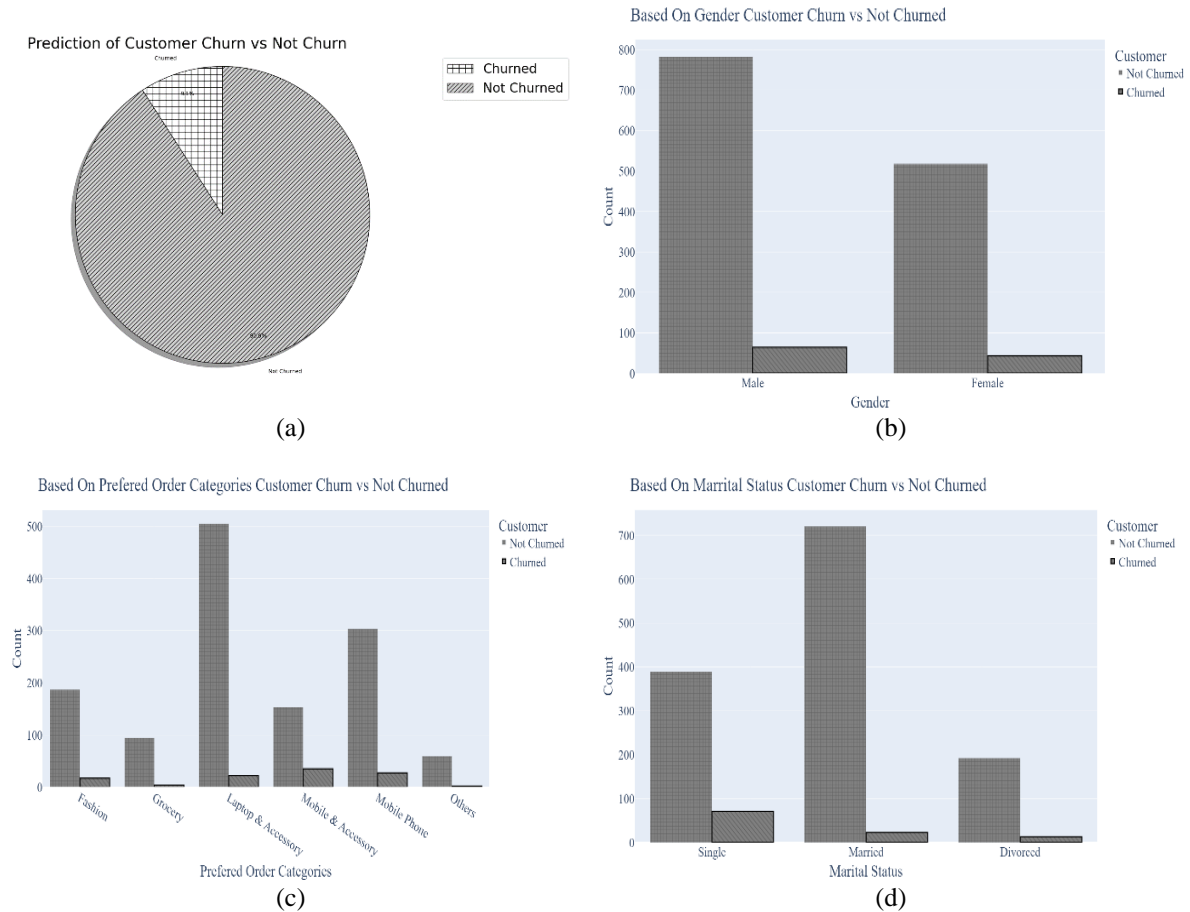


Figure 4. Customer churn data statistical analysis based on different parameters: (a) prediction of churn and not churn; (b) gender customer vs not churn; (c) preferred order categories customer vs not churn; and (d) marital status customer vs not churn

Figure 5 displays the customer annual spending predicted outcome analysis of test data based on various factors. If we analyze Figures 5(a)-(c), we can see that the average session length, time on the app, and length of membership are increasing the customer's annual spending is increasing. The customer's annual spending is high when the average session length is 34 minutes to 36 minutes. Also, when the time spent on the app is 13 minutes to 14 minutes, and the length of membership is 5 to 6, the customer's annual spending is significant. On the other hand, the time on the website is increasing the customer's annual spending is not increasing as shown in Figure 5(d). There is no effect on customer annual spending on time on the website, it is always almost the same.

Figure 6 shows the predicted outcome analysis of the test data that is split from the product on-time delivery dataset based on different parameters. The pie chart in Figure 6(a) shows that, the proportion of the product that will be delivered on time and the proportion that will not. If we analyze Figure 6(b), the number of products whose importance is low is high but the ratio of product late delivery is high whose product importance is high. On the other hand, the number of products whose customer rating is 1 is high but the ratio of products with late delivery is high whose customer rating is 2 as shown in Figure 6(c).

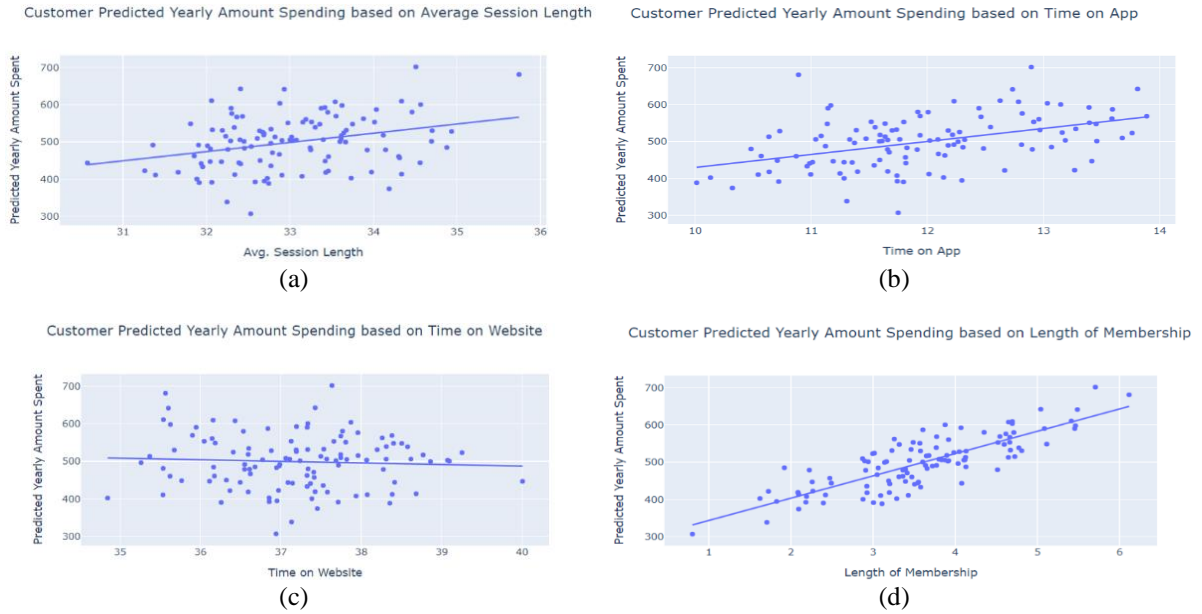


Figure 5. Customer annual spending data statistical analysis based on different parameters: (a) average session length, (b) time on app, (c) time on website, and (d) length of membership

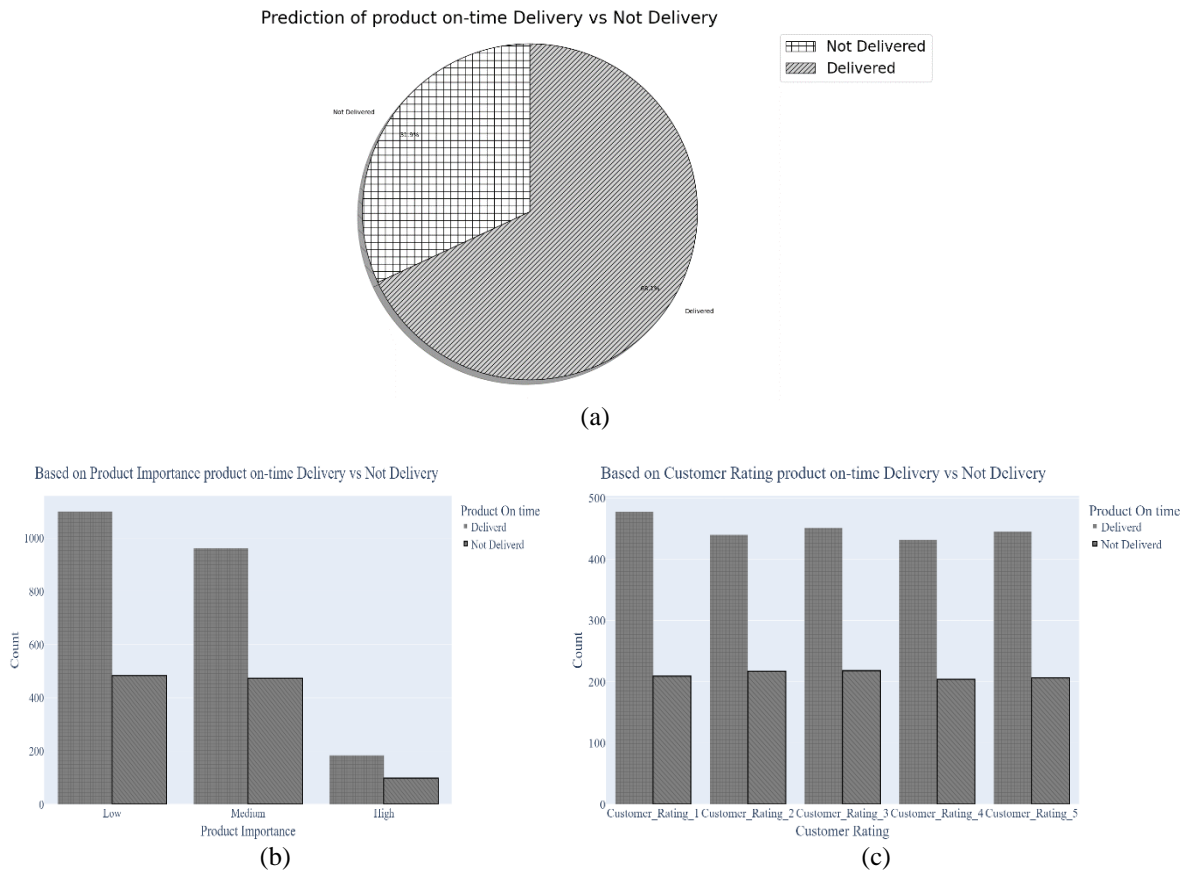


Figure 6. Customer ordered product on-time delivery test data statistical analysis based on different parameters: (a) prediction of ordered product on-time delivery vs not delivery, (b) based on product importance, and (c) based on customer rating

4. CONCLUSION

The use of machine learning algorithms to predict customer annual spending, customer churn and on-time product delivery in the e-commerce industry is valuable for businesses looking to improve their customer retention and increase their revenue. The results of this work demonstrate that regression and classification algorithms can be effectively used to analyze customer data. The study also found that factors such as purchase frequency, product categories, and average session length are important predictors of customer churn, annual spending, and product on-time delivery in e-commerce. The results of this paper can contribute to a better understanding of customer behavior in the e-commerce industry and provide insights into ways to improve customer retention and increase revenue.

ACKNOWLEDGMENTS

We would like to express our gratitude to DIU NLP and ML Research Lab at Daffodil International University's for their assistance in preparing the datasets and laboratory work. No specific grant from a public, private, or nonprofit funding organization was given for this research.




REFERENCES

- [1] K. Matuszelański and K. Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," *Journal of Theoretical Applied Electronic Commerce Research*, vol. 17, no. 1, pp. 165–198, 2022, doi: 10.3390/jtaer17010009.
- [2] X. Xiahou and Y. Harada, "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM," *Journal of Theoretical Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, 2022, doi: 10.3390/jtaer17020024.
- [3] M. Pondel *et al.*, "Deep learning for customer churn prediction in e-commerce decision support," *Business Information Systems*, vol. 1, pp. 3–12, 2021, doi: 10.52825/bis.v1i.42.
- [4] X. Xiahou and Y. Harada, "Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-Commerce Industry," *American Journal of Industrial and Business Management*, vol. 12, no. 03, pp. 277–293, 2022, doi: 10.4236/ajibm.2022.123015.
- [5] H. K. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with Cost-Enabled Cost-Sensitive Classifier for Customer Churn Prediction," *Computational Intelligence and Neuroscience*, vol. 2022, p. 9028580, 2022, doi: 10.1155/2022/9028580.
- [6] A. Yaseen, "Next-Wave of E-commerce: Mobile Customers Churn Prediction using Machine Learning," *Lahore Garrison University Journal of Computer Science and Information Technology*, vol. 5, no. 2, pp. 62–72, 2021, doi: 10.54692/igurjcsit.2021.0502209.
- [7] S. Kim and H. Lee, "Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees," *Procedia Computer Science*, vol. 199, pp. 1332–1339, 2021, doi: 10.1016/j.procs.2022.01.169.
- [8] S. Agrawal, A. Das, A. Gaikwad, and S. Dhage, "Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis using Deep Learning," *International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Shah Alam, Malaysia, 2018, pp. 1-6, doi: 10.1109/ICSCEE.2018.8538420.
- [9] L. Gan, "XGBoost-Based E-Commerce Customer Loss Prediction," *Computational Intelligence and Neuroscience*, vol. 2022, p. 1858300, 2022, doi: 10.1155/2022/1858300.
- [10] H. Li, Z. Guan, and Y. Cui, "Customer Churn Prediction Based on BG / NBD Model," *16th Wuhan International Conference on e-Business WHICEB 2017*, 2017, pp. 386–393.
- [11] E. J. Kędziora and G. K. Maksim, "Performance analysis of machine learning libraries," *Journal of Computer Sciences Institute*, vol. 20, no. 2, pp. 230–236, 2021, doi: 10.35784/jcsi.2693.
- [12] K. Chen, Y. H. Hu, and Y. C. Hsieh, "Predicting customer churn from valuable B2B customers in the logistics industry: a case study," *Information Systems and e-Business Management*, vol. 13, no. 3, pp. 475–494, 2015, doi: 10.1007/s10257-014-0264-1.
- [13] P. Sharma, A. Chakraborty, and J. Sanyal, "Machine Learning based Prediction of Customer Spending Score," *Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2019, pp. 1-4, doi: 10.1109/GCAT47503.2019.8978374.
- [14] C. J. Liu, T. S. Huang, P. T. Ho, J. C. Huang, and C. T. Hsieh, "Machine learning-based e-commerce platform repurchase customer prediction model," *PLoS One*, vol. 15, no. 12 December, pp. 1–15, 2020, doi: 10.1371/journal.pone.0243105.
- [15] M. A. Karim, P. Samaranayake, A. J. R. Smith, and S. K. Halgamuge, "An on-time delivery improvement model for manufacturing organisations," *International Journal of Production Research*, vol. 48, no. 8, pp. 2373–2394, 2010, doi: 10.1080/00207540802642245.
- [16] J. Li, T. Wang, Z. Chen, and C. Luo, "Machine Learning Algorithm Generated Sales Prediction for Inventory Optimization in Cross-border E-Commerce," *International Journal of Frontiers in Engineering Technology*, vol. 1, no. 1, pp. 62–74, 2019, doi: 10.25236/IJFET.2019.010107.
- [17] Z. Mohammed and S. Kadhem, "A Study about E-Commerce Based on Customer Behaviors," *Engineering and Technology Journal*, vol. 39, no. 7, pp. 1060–1068, 2021, doi: 10.30684/etj.v39i7.1631.
- [18] X. Yu, S. Guo, J. Guo, and X. Huang, "An extended support vector machine forecasting framework for customer churn in e-commerce," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1425–1430, 2011, doi: 10.1016/j.eswa.2010.07.049.
- [19] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 4, pp. 2367–2383, 2018, doi: 10.11591/ijece.v8i4.pp2367-2383.
- [20] S. Raeisi and H. Sajedi, "E-Commerce Customer Churn Prediction by Gradient Boosted Trees," *10th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 55–59, 2020, doi: 10.1109/ICCKE50421.2020.9303661.
- [21] H. L. Wu, W. W. Zhang, and Y. Y. Zhang, "An empirical study of customer churn in e-commerce based on data mining," *International Conference on Management and Service Science*, no. 70971124, pp. 0–3, 2010, doi: 10.1109/ICMSS.2010.5576627.
- [22] C. C. Wang, C. H. Chien, and A. J. C. Trappey, "On the application of ARIMA and LSTM to predict order demand based on




- short lead time and on-time delivery requirements,” *Processes*, vol. 9, no. 7, 2021, doi: 10.3390/pr9071157.
- [23] K. Singh, P. M. Booma, and U. Eaganathan, “E-Commerce System for Sale Prediction Using Machine Learning Technique,” *Journal of Physics: Conference Series*, vol. 1712, no. 1, 2020, doi: 10.1088/1742-6596/1712/1/012042.
- [24] B. Shen, “E-commerce Customer Segmentation via Unsupervised Machine Learning,” *ACM International Conference on Computing and Data Science*, no. 1994, 2021, doi: 10.1145/3448734.3450775.
- [25] X. Wu and S. Meng, “E-commerce customer churn prediction based on improved SMOTE and AdaBoost,” *13th International Conference on Service Systems and Service Management (ICSSSM)*, pp. 0–4, 2016, doi: 10.1109/ICSSSM.2016.7538581.
- [26] B. P. Chamberlain, A. Cardoso, C. H. B. Liu, R. Pagliari, and M. P. Deisenroth, “Customer lifetime value prediction using embeddings,” *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2017, pp. 1753–1762, doi: 10.1145/3097983.3098123.

BIOGRAPHIES OF AUTHORS






Md Abdullah Al Rahib    received a Bachelor degree in Computer Science and Engineering from Daffodil International University, Dhaka, Bangladesh in 2022. His research interests are machine learning, deep learning, data mining, data science, natural language processing, and artificial intelligence. He can be contacted at email: abdullah15-12247@diu.edu.bd.






Nirjhor Saha    received a Bachelor degree in Computer Science and Engineering from Daffodil International University, Dhaka, Bangladesh in 2022. His research interests include machine learning, data science, and artificial intelligence. He can be contacted at email: nirjhor15-12207@diu.edu.bd.



Raju Mia    is a junior Software Engineer at Nimusoft Technologies Ltd. Hereceived a Bachelor degree in Computer Science and Engineering from Daffodil International University, Dhaka, Bangladesh in 2022. His research interests include machine learning, data science, and artificial intelligence. He can be contacted at email: raju15-11995@diu.edu.bd.



Abdus Sattar    is currently doing a Doctor of Philosophy (Ph.D.) in Computer Science and Engineering at Bangladesh University of Professionals (BUP). Previously, he received a Bachelor of Science in Computer Science and Engineering (CSE) from Ahsanullah University of Science and Technology (AUST) and a Master’s Program of Interactive Systems Engineering (ISE) from KTH-Royal Institute of Technology, Sweden. During the master’s thesis research and development project, he was employed as a research assistant on a research project. This Research Project work was carried out in collaboration with Södertörn University, Stockholm University, and Karolinska Institutet Innovations, Sweden. Currently, he is employed as an Assistant Professor in the Department of Computer Science and Engineering (CSE) at Daffodil International University (DIU). Previously, he was employed as an Assistant Professor in the Department of Computer Science and Engineering (CSE) at Britannia University, Comilla. He can be contacted at email: abdus.cse@diu.edu.bd.