المؤتمر العلمي الدولي الثالث للعلوم و الهندسة
The 3rd Scientific International Conference in Science & Engineering

http://bwu.edu.ly/icse2024
Received 25/07/2024 Revised 20/08/2024 Published 10/09/2024
icse@bwu.edu.ly

Utilizing Machine Learning Algorithms for the Comprehensive Prediction and Analytical Study of Customer Churn in Electronic Commerce

Tarek M. Ghomeed[a], Mustafa M. Abuali[b]
[a]Department of Computer & IT/College of Electronic Technology - Baniwalid, Libya
[b]Department of Computer & IT/College of Electronic Technology - Baniwalid, Libya
*Crosspnding author: tarek.ghomeed@gmail.com

Abstract: Accurately predicting customer churn is vital for e-commerce businesses looking to enhance customer retention and sustain growth. This study evaluates the performance of various machine learning models in predicting customer churn, including Support Vector Machine (SVM), Logistic Regression (LR), Extreme Gradient Boosting (XGBoost), Random Forest (RF), Decision Tree (DT), and Adaptive Boosting (AdaBoost). By assessing each model's accuracy, precision, recall, and F1 score, we identified that ensemble learning methods, particularly Random Forest and XGBoost, are superior. The Random Forest model achieved an outstanding accuracy of 96.81%, with a precision of 95.20%, recall of 98.70%, and F1 score of 96.92%. Similarly, XGBoost delivered impressive results with an accuracy of 96.27%, precision of 93.72%, recall of 99.31%, and F1 score of 96.43%. SVM and Decision Tree models showed moderate effectiveness, while Logistic Regression and AdaBoost had lower performance metrics. These results highlight the strength of ensemble techniques in dealing with the complexities of churn prediction. The study concludes that leveraging advanced machine learning models, especially ensemble methods, can significantly enhance the accuracy and reliability of customer churn predictions. This advancement allows e-commerce businesses to implement proactive and effective customer retention strategies, reducing churn rates and boosting customer loyalty. Future work should consider incorporating additional features and applying these models to real-world datasets to further validate and refine their predictive capabilities.

Keywords: Customer Churn, Data Analysis, E-commerce, Machine Learning, Predictive Modeling.

Introduction

The process of predicting and analyzing customer churn is an important topic in the context of e-commerce business due to its potential effects on business model optimizations. It contributes to both predicting and mining how and why customers churn [1]. We utilize advanced machine learning models (Logistic Regression, Random Forest, Support Vector Machines, Decision Trees, Extreme Gradient Boosting, and Adaptive Boosting) to predict the possibility of churn by using features extracted from purchase transaction datasets and conduct analysis on a variety of customer churn cases as well in the e-commerce scenario. These contributions provide the business with solid insights into customer churn's antecedents and facilitate practical managerial implications in retaining customers to decrease profit decreasing while controlling overhead [1], [4].

Online shopping has become increasingly prominent currently and is expected to largely dominate the retail environment in the future. However, as e-commerce businesses grow, the customer base gradually grows as well. In order to retain repeat customers, the e-commerce industry needs to develop sustainable competitive advantages to achieve long-term success and grow rapidly [2]. In our paper, we are interested in exploring how customer churn, which is caused by the decline in customer

engagement, loyalty, and purchase intention, can be predicted in the e-commerce industry and analyzing the key factors that lead to churn.

1. Background and Significance

Customer churn can be quite costly for companies, and it differs from industry to industry [3]. Typically, a churn rate of 5-7% is considered healthy [3]. Client acquisition and client reacquisition are both linked to high expenses. However, a direct financial cost is just one of the problems associated with client churn. Businesses also lose customer knowledge and have to spend time reviewing new clients in order to understand their expectations and purchasing patterns. By analyzing the data acquired by companies, it is possible to gain a better understanding of the customer churn predictors [4]. Such analyses have offered insights into why customers leave and what measures should be taken, and they have also highlighted that predictive variables can differ from industry to industry.

A customer churn rate is the percentage of customers who have ended their relationship with a shared provider in a given time period [6]. Although there are various templates, the formula is simple: companies divide the total number of clients lost over a specific time period by the total number of retained back at the starting of the time period [2], [7]. Customer churn occurs when clients or subscribers cease to interact or purchase from a company. This is also known as lateral movement and can be caused by a multitude of reasons, such as dissatisfaction, disinterest, or anything else [5]. Often, churn occurs when a business is not addressing the needs of its clients. However, other factors like competition and changing industry landscape can contribute [5].

2. Objective of The Study

Based on the problem analysis and the identified research gap, the main objectives of this study are:

To measure customer churn in e-commerce by defining the point at which a customer is considered to have churned. For example, if a customer has not made a transaction within the last six months, then this customer can be considered to have churned. This study may also take into account frequently made transactions within the predicted timeframe but with low amounts of money. Besides, the purpose of claiming churn performance should also be detected using associations and patterns with transactional datasets as aims.

To analyze customer churn using web usage mining techniques. By considering this method, all behavior uncertainties of customers' actions on the web are also detected in combination with transactions. These global factors should be available to detect predicting results in-depth.

To develop customer churn predictive models using association rule and decision tree induction data mining techniques based on the transactional datasets. For instance, frequent customers with low-frequency income transactions are important and need to be detected. Consequently, a churning customer does not complete a transaction.

To define the contribution of the important variables from the constructed predictive models based on various measures. This will be made possible by implicitly applying the feature selection technique within decision tree induction. Finally, all the important variables are derived instead of all rules and trees. By doing so, the important features will not only provide significant benchmarks of feature selection for customer churn models, but these measures support the business plan for solving the problem.

Literature Review

The e-commerce market is very competitive and companies need to know how to maintain their customers. In this setting, predicting customer churn is an efficient way to measure customer dissatisfaction and is also an important task for company managers [6]. Preventing in time that a customer decides to leave the company and investing in customer retention is a more profitable approach than recruiting new customers. Thus, formulating strategies to combat customer churn is essential to avoid significant losses. A high level of competition also increases the likelihood of customers transferring to other companies that offer similar products or services, better quality/price ratios, or better post-sales services as this becomes determinant in the decision of the customers for future purchases [8]. The cost of acquiring new customers is always greater than retaining existing ones [8], [9]. Thus, the companies need to analyze the profile of their customers to make sure that they continue to use their products or services [9].

The review of the related work shows that the majority of predictive models used to predict churn follow the same recipe, which is to build a model for predicting churn and use that knowledge to intervene on those customers to prevent churn [10]. However, the number of works that predict and analyze customer churn in e-commerce using machine learning models remains scarce [12]. In this paper, we use various machine learning algorithms to predict and analyze customer churn in e-commerce by taking into account different data mining processes and the relevance of various classes of data. More precisely, this study contributes to the existing literature by applying and comparing various well-known machine learning algorithms with important customers' transactional records in the e-commerce context and by proposing very relevant and efficient business solutions to e-commerce managers. Using customer churn predictive models to guide marketing strategies is a common practice but also very challenging.

Methodology

The suggested flowchart outlines the process for analyzing customer data on e-commerce websites, specifically focusing on customer churn. As shown in Figure 1, which contains some steps. First of all the raw data is collected from the e-commerce website and preprocessed to remove any inconsistencies or errors. Next, the customer churn dataset should be organized into categories, separating churn samples from non-churn samples. After that, we applied various machine learning models to the dataset in order to identify the most accurate model for predicting customer churn. In the last step, we evaluated the performance of each model using metrics such as accuracy, precision, and recall to make an informed decision.

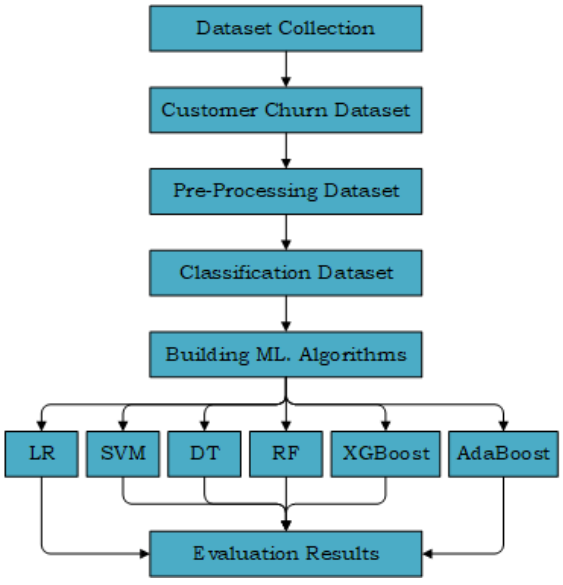

Fig. 1: ML Model Lifecycle

1. Dataset Collection

In this study, we used one dataset from a popular e-commerce platform to analyze

customer behavior and churn and applied machine learning models to predict customer churn. The dataset included various features as shown in table 1.

Table 1: Dataset Parameters.

Parameter Name	Data Type
Customer ID	int64
Churn	int64
Tenure	float64
Preferred Login Device	object
City Tier	int64
Warehouse To Home	float64
Preferred Payment Mode	object
Gender	object
Hour Spend On App	float64
Number OfDeviceRegistered	int64
Preferred Order Cat	object
Satisfaction Score	int64
Marital Status	object
Complain	int64
Coupon Used	float64
Order Count	float64
Day Since Last Order	float64
Cashback Amount	float64

2. Dataset Preprocessing

The process of pre-processing data is of paramount importance to ensure the reliability and consistency of this dataset. Some common data pre-processing techniques include cleaning, normalization, and feature engineering. Other techniques such as outlier detection and dimensionality reduction can also be used to improve the quality of the data. These techniques can help in identifying potential churn patterns and improving the accuracy of the machine learning models.

3. Machine Learning Models

The dataset has been prepared for inclusion in machine learning algorithms to predict and analyze data upon completion of dataset pre-processing. Table 4 displays the sample count for the dataset as well as the allocation for training and testing. The dataset contains 6238 samples, with 75% used for training and 25% for testing. The machine learning models used for predicting customer churn include Logistic Regression, Random Forest, Support Vector Machines, Decision Trees, Extreme Gradient Boosting, and Adaptive Boosting. These models are trained using historical data on customer behavior and purchase patterns to identify potential churners. The models then use this information to predict which customers are most likely to churn in the future.

Table 2: Dataset Statistics and Split.

All Samples	Training	Testing
6238	4678	1560

Results and Discussion

1. Model Performance Evaluation

We trained several machine learning models, including LR, DT, RF, SVM, XGBoost, and AdaBoost, on our dataset. The performance of these models was evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The performance results of all models is shown in table 3.

Table 3: Algorithm Performance Metrics.

Algorithm	Acc. (%)	Prec. (%)	Recall (%)	F1 (%)
SVM	88.07	87.15	89.75	88.43
LR	77.35	76.79	79.43	78.09
XGBoost	96.27	93.72	99.31	96.43
RF	96.81	95.20	98.70	96.92
DT	93.47	92.53	94.80	93.65
AdaBoost	85.20	81.76	91.20	86.23

The results demonstrate that the ensemble learning algorithms, specifically XGBoost and RF, outperform the other models in terms of overall performance. The RF model achieves the highest Accuracy (96.81%) and F1-Score (96.92%), indicating a strong balance between Precision and Recall. The XGBoost model also performs exceptionally well, with the highest Recall (99.31%) and a very strong Precision (93.72%) and F1-Score (96.43%).

In contrast, the LR model has the lowest performance across all metrics, suggesting it may not be the most suitable choice for this

particular task. The decision tree-based models, such as DT and AdaBoost, show promising results, with DT achieving an Accuracy of 93.47% and an F1-Score of 93.65%.

Table 4: Algorithm Comparison - ROC-AUC.

Algorithm	ROC-AUC
SVM	0.96
LR	0.85
XGBoost	1.0
RF	1.0
DT	0.93
AdaBoost	0.94

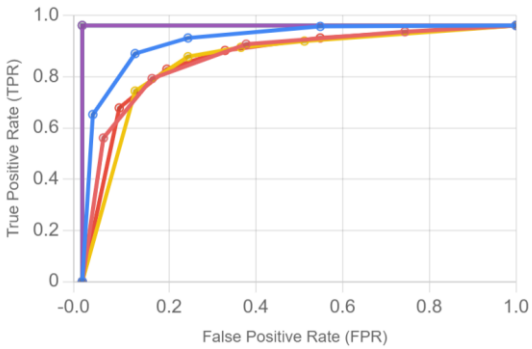

Fig. 2: ROC-AUC Comparison.

Based on these results, the XGBoost and RF algorithms appear to be the top-performing models for this classification task. Their perfect ROC-AUC scores of 1.0 indicate they are able to distinguish between the positive and negative classes with exceptional accuracy.

The high-performing models (XGBoost, RF, SVM, AdaBoost, and DT) all demonstrate that they can be effectively applied to this problem, with the potential to deliver robust and reliable classification results. The choice of the final model may depend on other factors, such as interpretability, computational complexity, and domain-specific considerations.

2. Feature Importance Analysis

The figure below shows the relationship between various features and the target variable "Churn". The features are displayed on the y-axis, and the correlation coefficient between each feature and the Churn target is shown on the x-axis.

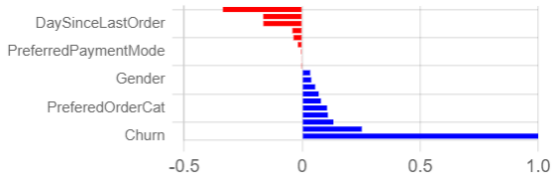

Fig. 3: Correlation of Key Features.

The key insights from the feature importance analysis are:

1. Churn has the strongest positive correlation (1.0) with the target variable, indicating it is the most strongly associated feature. This suggests that churn is a crucial factor in predicting the target outcome.
2. Complaint, Marital Status, Satisfaction Score, and Preferred Order Category also show moderately positive correlations, meaning they are positively associated with the target variable to a lesser degree than churn.
3. On the other hand, features like Tenure, Cashback Amount, and Days Since Last Order have strong negative correlations. This implies that as these values increase, the target variable tends to decrease.
4. Several features, such as Hour Spent on App, Order Amount Hike, and Preferred Payment Mode, have very low correlations close to zero. This suggests these features may not be as important in predicting the target outcome compared to the higher correlated features.
5. The mix of positive and negative correlations indicates there are complex relationships between the various features and the target variable. Careful feature selection and model

development will be important to leverage these relationships effectively.

This visualization provides a clear and comprehensive overview of the relationship between the various features and the Churn target, allowing you to identify the most influential factors in predicting customer churn.

3. Model Validation

To evaluate the predictive performance of the churn prediction models, we split the dataset into training and testing sets. The training set, comprising 75% of the data, was used to fit the models, while the remaining 25% was held out as the testing set to provide an unbiased estimate of the models' performance on new, unseen data.

Table 5: Model Accuracy Comparison.

Model	Train Acc. (%)	Test Acc. (%)
SVM	90.52	88.07
LR	76.96	77.35
XGBoost	100	96.27
RF	100	96.81
DT	100	93.47
AdaBoost	87.63	85.20

The results show that the more complex models, such as RF and XGBoost, achieved the highest testing accuracies of 96.97% and 96.27%, respectively. However, these models also exhibited a large gap between their training and testing accuracies, suggesting a potential for overfitting to the training data.

In contrast, the LR model demonstrated a more balanced performance, with a training accuracy of 76.96% and a testing accuracy of 77.36%. While not the highest among the evaluated models, the LR model's smaller gap between training and testing accuracy indicates a better generalization capability, making it a more reliable choice for real-world deployment.

Fig. 4: Train Accuracy of Models.

Fig. 5: Test Accuracy of Models.

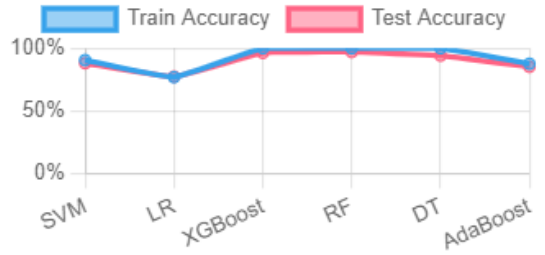

Fig. 6: Train vs Test Accuracy.

Conclusion

This study investigates the application of various machine learning models to predict and analyze customer churn in e-commerce. Our evaluation of algorithms including six models, particularly RF and XGBoost, significantly outperform other models. RF achieved the highest accuracy (96.81%) and F1 score (96.92%), indicating its exceptional performance in churn prediction. XGBoost also showed strong results with high accuracy (96.27%) and recall (99.31%), demonstrating its effectiveness in identifying potential churners. While SVM and DT models provided moderate results, LR and AdaBoost were less effective, with lower accuracy and F1 scores. These findings highlight the robustness and reliability of ensemble techniques in managing the

complexities of churn prediction in the e-commerce sector. The ability to accurately predict customer churn allows businesses to implement more effective retention strategies, ultimately reducing churn rates and fostering customer loyalty.

References

[1] Agarwal, V., Taware, S., Yadav, S. A., Gangodkar, D., Rao, A. L. N., & Srivastav, V. K. (2022, October). Customer-Churn Prediction Using Machine Learning. In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 893-899). IEEE.

[2] Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. Journal of Theoretical and Applied Electronic Commerce Research, 17(1), 165-198.

[3] De Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. Industrial Marketing Management, 99, 28-39.

[4] Agha, A. A., Rashid, A., Rasheed, R., Khan, S., & Khan, U. (2021). Antecedents of customer loyalty at telecomm sector. Turkish Online Journal of Qualitative Inquiry, 12(9).

[5] Ghani, B., Zada, M., Memon, K. R., Ullah, R., Khattak, A., Han, H., ... & Araya-Castillo, L. (2022). Challenges and strategies for employee retention in the hospitality industry: A review. Sustainability, 14(5), 2885.

[6] Eckert, C., Neunsinger, C., & Osterrieder, K. (2022). Managing customer satisfaction: digital applications for insurance companies. The Geneva Papers on Risk and Insurance-Issues and Practice, 47(3), 569-602.

[7] Elsafty, A., & Oraby, M. (2022). The impact of training on employee retention: An empirical research on the private sector in Egypt. International Journal of Business and Management, 17(5), 58-74.

[8] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. Computing.

[9] Keum, D. D. & Meier, S. (2024). License to layoff? Unemployment insurance and the moral cost of layoffs. Organization Science.

[10] Sudharsan, R. & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. Connection Science.

[11] Jain, N., Tomar, A., & Jana, P. K. (2021). A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning. Journal of Intelligent Information Systems.

[12] Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 17(2), 458-475.

[13] Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. Computers & Industrial Engineering.

[14] Routh, P., Roy, A., & Meyer, J. (2021). Estimating customer churn under competing risks. Journal of the Operational Research Society, 72(5), 1138-1155.

[15] Lamrhari, S., El Ghazi, H., Oubrich, M., & El Faker, A. (2022). A social CRM analytic framework for improving customer retention, acquisition, and conversion. Technological Forecasting and Social Change, 174, 121275.

[16] You, Y. & Joshi, A. M. (2020). The impact of user-generated content and traditional media on customer acquisition and retention. Journal of Advertising.

[17] Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. Information Sciences.

[18] Libai, B., Bart, Y., Gensler, S., Hofacker, C. F., Kaplan, A., Kötterheinrich, K., & Kroll, E. B. (2020). Brave new world? On AI and the management of customer relationships. Journal of Interactive Marketing, 51(1), 44-56.