Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce

MITRA PENMETSA[1], JAYAKESHAV REDDY BHUMIREDDY[2], RAJIV CHALASANI[3], MUKUND SAI VIKRAM TYAGADURGAM[4], VENKATASWAMY NAIDU GANGINENI[5], SRIRAM PABBINEEDI[6]

[1]University of Illinois at Springfield, USA
[2]University of Houston, USA
[3]Sacred Heart University, USA
[4]University of Illinois at Springfield, USA
[5]University of Madras, Chennai, USA
[6]University of Central Missouri, USA

Abstract: Customer retention is one of the most critical concerns for businesses. Businesses try to decrease customer turnover in order to maximize client lifetime value while lowering the cost of acquiring new customers. By focusing on customer churn prediction and identification, organizations can foresee which customers are most likely to depart. This enables them to implement customized, pertinent actions to lower the rate of client attrition. This paper suggests a machine learning (ML)-based approach that makes use of Random Forest (RF) to predict customer attrition in e-commerce platforms. Recall, accuracy, precision, F1-score, and ROC-AUC were among the metrics used to evaluate the Random Forest model, which demonstrated an impressive accuracy of 95%, a precision of 98%, and a ROC-AUC of 98.51%. Random Forest was shown to be the most successful prediction method based on comparison tests with decision trees (DT) and support vector machines (SVM). With the use of real-time datasets, deep learning techniques, and large-scale deployment for e-commerce enterprises, these results confirm the efficacy of ensemble learning approaches in customer retention activities.

Keywords: E-commerce, big data, Customer churn, Customer retention, Predictive analytics, Machine learning, E-commerce Customer Churn data.

Received: June 17, 2025. Revised: July 23, 2025. Accepted: August 11, 2025. Published: September 4, 2025.

1. Introduction

E-commerce expansion fuelled by digital platforms now drives substantial fundamental shifts to worldwide commerce through the creation of customer interest in online shopping platforms. The digital transformation creates enormous data volumes labeled as big data that give us unprecedented abilities to grasp and steer consumer actions [1]. Since data-driven purchasing methods became standard in business decisions, organizations need quick changes to their buying practices. In the online retail sector, customer churn stands as a primary challenge since people stop using platforms and brands [2]. Digital business profitability suffers through customer churn, so businesses must focus on retaining customers. A business achieves better long-term growth when it retains customers because returning customers tend to be less expensive to maintain than acquiring new customers [3][4]. As such, understanding why customers churn—and how to prevent it—has become a key focus area for businesses aiming to optimize customer experience and retention strategies[5].

Organizations implement analytical models and predictive tools to track customer patterns that help them make anti-churn predictions [6][7]. The analysis of present-era customer relations proves difficult through conventional analytic techniques, particularly in markets showing economic and cultural variations, according to research by [8][9]. The establishment of advanced data-based techniques has led to better retention methods. Companies implement ML systems to analyze big data, which grants them exact modeling capabilities in consumer understanding. ML algorithms process large, complex datasets to generate specific customer retention forecasts [10]. This converts into useful information that supports tailored marketing in the context of e-commerce, customer segmentation, and proactive engagement strategies.

1.1 Motivation and Contribution of Paper

Businesses operating in e-commerce find it more economical to maintain existing customers instead of spending resources on acquiring new ones. The prediction of customer needs with quick problem resolution leads to business sustainability due to revenue decline when customers discontinue business activities. Established through observation of real-time customer data, behaviour prediction systems facilitate optimal churn evaluation. The group came to significant results using SMOTE and RF models in conjunction with contemporary preprocessing techniques, which improved retention rates and decreased operating costs. A practical, scalable model for e-commerce customer churn prediction has been presented through ML technologies. The key contributions include:

- Actual e-commerce data from Kaggle to create a systematic technique for predicting customer attrition.
- Carried out comprehensive data preparation, which included deleting outliers, addressing missing values, and changing features using Z-score normalization and one-hot encoding.
- Improved representation of minority classes through the use of SMOTE to solve class imbalance.
- Used the RF technique to categorize churned consumers due to its efficacy and reliability.
- The model's performance was evaluated using ROC-AUC, F1-score, recall, accuracy, and precision metrics.

1.2 Justification and Novelty of paper

This study supports the use of ML in e-commerce platform customer churn prediction by demonstrating the RF algorithm's good predictive performance, including accuracy, precision, and ROC-AUC scores. The novelty of this work lies in the application of the RF model to customer churn prediction, enhanced by addressing class imbalance through the SMOTE and comprehensive data preprocessing, which are often overlooked in existing studies. Additionally, the study's comparative analysis with traditional models like SVM and DT highlights the superior performance of ensemble learning methods in handling complex customer data, making it a valuable contribution to the use of predictive analytics to changing e-commerce settings in order to retain customers.

1.3 Structure of paper

The paper is structured as follow: Section II provides a background study on Predictive Analytics for Customer Retention. In Section III, the methodology is detailed. In Section IV, the results, analysis, and discussion are compared. Section V presents the study's conclusion and plans for further research.

2. Literature Review

Several significant research studies related to customer retention have been reviewed and analyzed to support the development of the current work. Table I presents the background study on customer behaviour aimed at improving retention, including details on datasets used, models applied, performance metrics, and key contributions.

Ullah et al (2019) provides a churn prediction model that use classification and clustering algorithms to identify churning consumers and explain why telecom industry customers leave. features are picked utilizing a correlation attribute ranking filter and information gain. With 88.63% of correctly categorized cases, the RF method performed well in the proposed model's first categorization of churn customer data, utilizing classification techniques. One of the CRM's primary responsibilities is developing efficient retention rules to stop churners [11].

Ahmad, Jafar and Aljoumaa (2019) create a model for predicting customer attrition this assists telecom providers in determining which customers are most likely to quit. By applying ML approaches to a huge data platform, the model presented in this study provides a novel approach to feature engineering and selection. The model's performance is tested using the AUC standard measure, which yields an AUC value of 93.3%. Another important innovation is the usage of the customer's social network to extract SNA attributes. Compared to the AUC benchmark, the model's performance improved from 84 to 93.3% when SNA was implemented. Using a large dataset created by translating enormous volumes of raw data given by the Syriatel telephone operator, the model was created and assessed using the Spark environment [5].

Ebrah and Elnasir (2019) To predict churn, three ML algorithms—two benchmark datasets have been analyzed using NB, SVM, and DT. The IBM Watson dataset has 7033 observations and 21 attributes, whereas the cell2cell dataset contains 71,047 observations and 57 attributes. According to the AUC, in a measure used to assess model performance, the models scored 0.82, 0.87, and 0.77 for the IBM dataset and 0.98, 0.99, and 0.98 for the cell2cell dataset. Additionally, the accuracy of the suggested models was higher than that of earlier research employing the same datasets [12].

Saghir et al. (2019) assess current individual and ensemble classifiers based on neural networks and suggest an ensemble classifier that improves performance metrics and raises the accuracy of churn prediction by combining bagging and neural networks. To compare and assess the suggested approach, this study uses two benchmark datasets those purchased from GitHub. The suggested model has an average accuracy of 81% [13].

Sabbeh (2018) attempts to analyse and compare the efficiency of numerous ML approaches utilised for the churn prediction problem. The methods that were chosen include discriminant analysis, DT (CART), instance-based learning KNN, SVM, LR, ensemble-based learning (RF, Ada Boosting trees, and Stochastic Gradient Boosting), NB, and MLP. 3333 records from a telecommunications dataset were used to test the models. According to the results, ADA boost and RF both perform better than any other method, with around the same 96% accuracy. With 94% accuracy, SVM and MLP can also be suggested. 90% was attained by DT, 88% by NB, and 86.7% by LR and linear discriminant analysis (LDA) [14].

Li et al (2016) make a novel supervised one-sided sampling strategy is offered for preprocessing the unbalanced dataset. Following the K-means technique's clustering of the dataset into meaningful clusters, each cluster uses one-sided sampling to remove noise and unnecessary negative samples. The RF approach is used to choose key variables and reduce dimensions. In this study, the classifier used to forecast client attrition in two or three months is the C5.0 DT. The tests involve around 2.7 million 4G telecommunications consumer data points. With a recall ratio of 52.43%, it achieves a precision ratio of 80.42%. The suggested methodology offers accurate prediction results that may be applied in practice to win back possible lost clients [15].

TABLE I. SUMMARY OF RELATED WORKS ON CUSTOMER CHURN PREDICTION USING MACHINE LEARNING TECHNIQUES

Author	Proposed Work	Dataset	Key Findings	Challenges/Gaps
Ullah et al. (2019)	Churn prediction via clustering and classification; used InfoGain and correlation filters for feature selection	Telecom sector data	RF achieved 88.63% classification accuracy; helped identify churn factors	Lack of big data handling; limited algorithmic diversity
Ahmad, Jafar, and Aljoumaa (2019)	Telecom churn prediction with big data and ML; introduced SNA features	SyriaTel big telecom data	AUC improved from 84% to 93.3% with SNA features; scalable Spark-based system	High dependency on large infrastructure (Spark); data complexity
Ebrah and Elnasir (2019)	Decision Trees, SVM, and Naïve Bayes	IBM Watson (7033 records), Cell2Cell (71,047 records)	High AUCs: up to 0.99 for Cell2Cell dataset; models outperformed previous studies	Limited interpretability; no deep learning models used; domain focused on telecom
Saghir et al. (2019)	Ensemble Neural Network with Bagging	Two GitHub benchmark datasets	Ensemble model achieved 81% average accuracy	Lack of real-world deployment context; datasets unspecified; potential data preprocessing issues

Sabbeh (2018)	Comparative Analysis: 10 ML techniques (e.g., RF, AdaBoost, SVM, MLP)	Telecom dataset (3,333 records)	RF and AdaBoost ~96% accuracy; SVM & MLP ~94%	Small dataset; limited scalability; no big data platform used
Li et al. (2016)	One-sided sampling, k-means clustering, and C5.0 for churn prediction	2.7 million 4G telecom customer records	Achieved 80.42% precision and 52.43% recall	Imbalanced data remains a challenge; moderate recall

3. Research Methodology

The study approach for Customer Retention in E-commerce is a systematic, data-driven strategy that works well for practical e-commerce applications. Initially, the commerce customer churn dataset was collected from Kaggle. Then, pre-process the data for missing values handled and outliers removed. Then, categorical features were transformed using one-hot encoding, and numerical attributes were standardized using Z-score normalization to maintain consistency across variables. The SMOTE was then utilized to increase minority class representation in order to reduce the class gap between churned and kept clientele. The processed data was divided into training and testing sets with an 80:20 ratio. Following that, RF was used to categorize customer turnover. To ensure the robustness and generalizability of the prediction outputs, the models were assessed using performance measures such as accuracy, precision, recall, F1, and ROC-AUC. Figure 1 provides an illustration of the methodology's general phases.

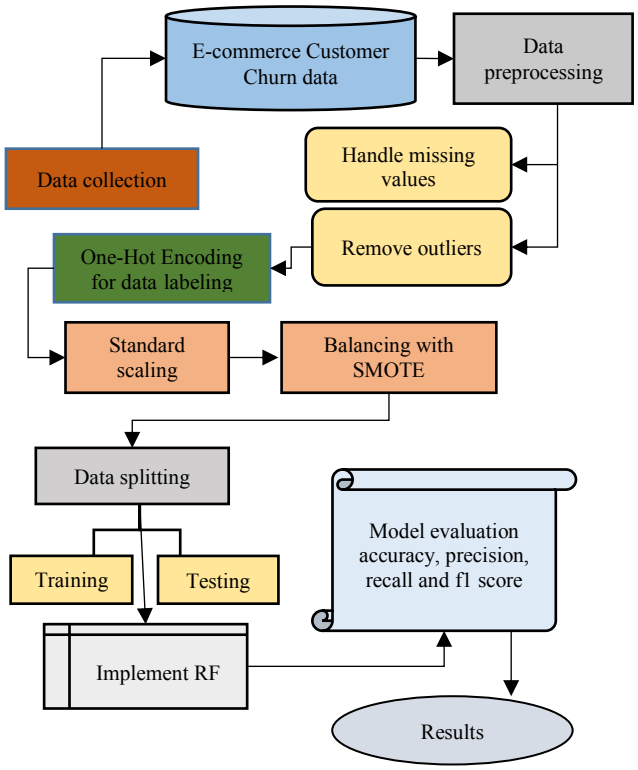

Fig. 1. Proposed flowchart for Customer Retention

Each step of a proposed flowchart for Retention Optimization through Predictive Analytics is provided below:

1.1 Data Collection

The e-commerce customer churn dataset used in this study contains detailed behavioral and transactional information of customers from an online retail platform. It includes key variables such as customer demographics, order history, purchase frequency, last purchase date, number of complaints, and payment methods, which are critical in analyzing churn behaviour. Each record represents a customer and is labeled to indicate whether they have churned or remained active. The visual inspect of data are provide in below:

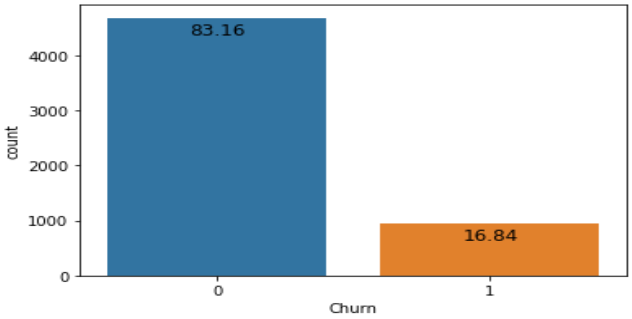

Fig. 2. Churn Distribution

Figure 2 illustrates the distribution of churn in the dataset, where '0' (Not Churn) is substantially taller than the bar for '1' (Churn), indicating a much larger number of customers who did not churn. Specifically, approximately 83.16% of the customers in while only around 16.84% of consumers churned, the dataset did not. This notable class imbalance is an important characteristic of the dataset.

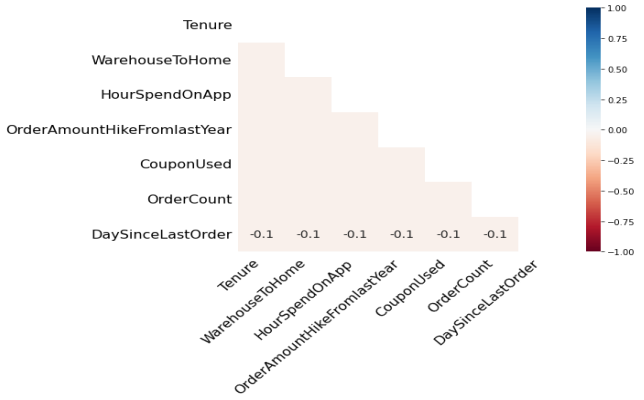

Fig. 3. Correlation Heatmap

Figure 3's correlation heatmap shows that the associations are generally weak among the numerical features, with 'Day Since Last Order' showing a consistent weak negative correlation of -0.1 with others. Overall, no strong correlations are observed.

1.2 Data Preprocessing

A crucial step in guaranteeing the dataset's quality and consistency is preprocessing the benchmarked customer data. Data preprocessing involves several steps, each addressing particular difficulties with data structure and quality, and relevance. The following steps of pre-processing are as follows:

- **Handle missing value:** This step involves processing and imputation of missing data. A few of the chosen algorithms let the mean, median, or zero to be used in place of missing values.

- **Remove Outliers:** There are a lot of outliers in the dataset, which might distort the findings and alter the sample mean and variance. Using the fourth quartile approach, outliers in the characteristics were eliminated.

1.3 One-Hot Encoding for Categorical Features

The label values seen in categorical data are regarded as nominal values. The ML model's efficiency can be increased by converting the categorical input into numerical information. Presenting the data as a feature and encoding it in binary code is one method of hot encoding. One of the most popular approaches compares each numerical variable level to a predetermined beginning point.

1.4 Standard Scaling

Standardizing data is a typical preprocessing procedure that helps the model converge more quickly and perform better by ensuring that features are of the same size. Z-score normalization is a popular technique for standardizing data and is defined as Equation (1):

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

where z is the standardized value, x is the original value, d is the feature standard deviation, and d is the feature mean. Each feature is subjected to this procedure separately, changing the data in order for its standard deviation to be 1 and its mean to be 0.

1.5 Class Imbalance Handling

The statistics showed a notable class imbalance, with around 17% of consumers churning and 83% of customers not churning. The SMOTE was used on the training dataset in order to solve this problem. In order to create a balanced class distribution and reduce the possibility of bias in the prediction model, this strategy intentionally oversamples the minority class (churned consumers).

1.6 Data Splitting

The training and testing subsets are the two divisions of the dataset. 20% of the data is used for model assessment, while 80% is used for model training.

1.7 Random Forest (RF) Model

RF is an approach to supervised learning that employs a group of tree-structured classifiers, or DT, applied to a number of sub-samples in a given dataset. The DT is made up of multiple nodes connected by branches that stretch from the root node, which is usually located at the top, to the leaf nodes. At the decision nodes leading to a branch, characteristics are considered [16][17]. These branches may terminate in a leaf node or lead to another decision node [18]. The algorithm is an example of supervised learning, which necessitates a training dataset that contains the target variable's values. The CART technique was initially the particular decision tree to be employed; Two branches are produced by each decision node, making the tree binary. It grows by choosing the best measurement vectors that lower the maximum impurity using a "exhaustive search of all available variables and all possible splitting values." Dividing the root node into binary segments is the initial stage in building a DT. The Equation (2) of potential splits s at node t serves as the foundation for the splitting process:

$$\Delta \mathrm{i}(\mathrm{s},\mathrm{t}) = \mathrm{i}(\mathrm{t}) - p_L \mathrm{i}(t_L) - P_R \mathrm{i}(t_r) \quad (2)$$

where $\Delta \mathrm{i}(\mathrm{s},\mathrm{t})$ is a metric for reducing impurities from split s, $i(t)$ symbolizes the impurities before to separation, and $i\ (t_L)$ and $i(t_R)$ demonstrate After halving node t with split s, the left child's impurity node T_L and the right child's node T_R. There are a number of estimates for measuring these impurities, but the Gini impurity—which quantifies the likelihood that a random distribution of labels in the subset results in an inaccurate label for a randomly chosen element from the set is the condition for splitting by default.

1.8 Evaluation Metrics

This is the last level of the prediction model. Here, analyse the prediction outcomes using a number of evaluation measures, such as classification accuracy, confusion matrix, and F1-score. Every measure is dependent on statistical variables that come from a class confusion matrix. The following instances of the confusion matrix are:

- **TP (True Positive):** TP reflects the total number of clients that were properly identified as churn.
- **FP (False Positive):** FP represents the total number of clients who are wrongly classified as churn.
- **TN (True Negative):** TN represents the total number of accurately recognized non-churn clients.
- **FN (False Negative):** It represents the total number of consumers who were wrongly labelled as non-churn.

Accuracy: The model's performance is evaluated in terms of accuracy. Accuracy is the ability to discriminate between actual and unrealistic possibilities. It is provided as Equation (3)-

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+Fp+TN+FN}} \quad (3)$$

Precision: Precision refers to the number of positive class forecasts that really become true positive class predictions. It is determined by taking the number of accurately predicted positive observations and dividing it by the total expected positive observations. It is written as Equation (4)-

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \quad (4)$$

Recall: Recall is defined as the number of accurate positive predictions divided by the overall number of correct positive samples. It is expressed mathematically as Equation (5)-

$$Recall = \frac{\mathrm{TP}}{TP+FN} \quad (5)$$

F1 score: The F1 score is computed by taking the harmonic mean of the model's accuracy and recall, and evaluates a model's performance on a dataset. Mathematically, it is given as Equation (6)-

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

ROC-AUC: The area under the ROC curve, which compares the TPR (Recall) to the FPR, is known as ROC-AUC. Equation (7) represents the ROC.

$$AUC = \int_0^1 TPR(x)dx \quad (7)$$

According to Equation(7), a higher AUC denotes a better-performing model that can successfully differentiate across classes.

4. Results And Discussion

The efficiency of the suggested ML-based strategy was measured and validated in this section through a variety of experiments conducted on the datasets referenced. For this experiment, a laptop with a Core i3 CPU operating at 2.0 GHz and 4 GB of RAM is utilized. The performance of the suggested RF model on consumer data from e-commerce is shown in Table II. It achieved a high accuracy of 95%, indicating its strong overall predictive capability. The model's precision of 98% highlights its excellent ability to correctly identify customers who are likely to remain loyal, while a recall of 83% reflects its competence in detecting actual churners. The chosen model attains adequate performance levels in real-life scenarios with an F1-score balance of 86%. Most notably, the model attained a ROC-AUC score of 98.51%, underscoring its exceptional discrimination power between churn and non-churn classes and validating its effectiveness for predictive analytics in retention optimization.

TABLE II. EXPERIMENT RESULTS OF PROPOSED MODEL FOR CUSTOMER RETENTION

Performance Matrix	Random Forest (RF)
Accuracy	95
Precision	98
Recall	83
F1-score	86
ROC-AUC	98.51

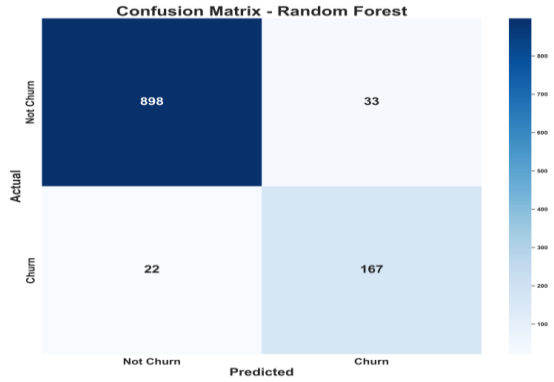

Fig. 4. Confusion Matrix for Random Forest

The RF model's confusion matrix in Figure 4, which has 898 TN and 167 TP, demonstrates excellent churn prediction ability. 33 non-churners were incorrectly categorized as churners FP, and 22 churners as non-churners FN, suggesting that while overall accuracy was good, precision and recall might be improved.

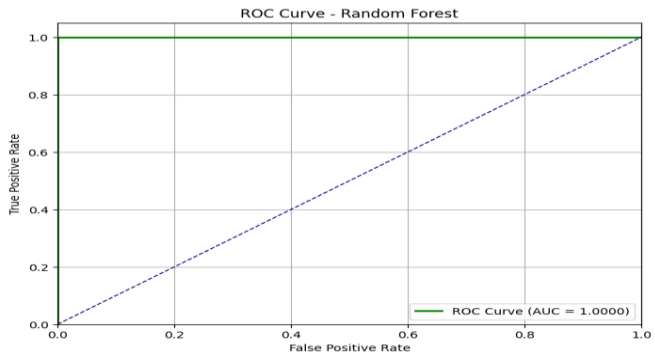

Fig. 5. ROC Curve for the Random Forest

Figure 5 shows the ROC curve for the RF model, with an AUC value of 1.0000, demonstrates perfect categorisation. The curve rapidly goes towards the top-left corner, which exhibits ideal performance for the TP rate and very low FP rates. This demonstrates how well the model separates churn from non-churn clients.

4.1 Comparison with Discussion

The section provides a comprehensive review of alternative procedures against present-day approaches. Table III demonstrates how predictive models work to keep customers by comparing their approaches. ML and DL received research attention for accuracy testing purposes. Through analysis of the SVM model registered 77% accuracy in identifying customers who would depart the organization. Non-linear data structures run through DT, which generated model predictions with 87.9% accuracy. Using many models to provide 95% correct predictions, RF outperforms previous ensemble learning techniques and helps avoid overfitting. This comparison highlights the superior effectiveness of ensemble-based models like RF in predictive customer retention tasks.

TABLE III. COMPARISON BETWEEN VARIOUS MODELS PERFORMANCE FOR CUSTOMER RETENTION

Models	Accuracy
Support Vector Machine (SVM) [19]	77
Decision Tree (DT) [14]	87.9
Random Forest (RF)	95

The RF model brings multiple benefits to predictive customer retention while achieving 95% accuracy in predictions. RF ensemble learning system utilizes numerous DT to minimize overfitting, yet it improves the accuracy of predictions for new data points. Additionally, the RF model is capable of handling large datasets with high dimensionality and is proficient at handling both categorical and numerical inputs.

5. Conclusion and Future Study

The accurate prediction of customer turnover by companies that operate e-commerce becomes necessary for designing successful retention approaches and marketing strategies. For modern businesses, predicting customer attrition has become vital to achieve organizational success. A new, reliable ML method based on RF predicts customer churn behaviour in e-commerce platforms. The model outperformed its competitors by reaching 95% accuracy and maintaining 98% precision, along with 83% recall, and achieving a 98.51% ROC-AUC score. Data preprocessing methods, along with the SMOTE, helped the model become more efficient in recognizing churn and non-churn customers. The analysis of SVM and DT models proved that ensemble learning outperforms both in predicting customer retention. The performance of the model was strong, but critics pointed out its dependency on one dataset, together with its insufficient ability to recognize customers. Future research will explore the application of DL models, real-time datasets, and optimization for large-scale deployment, enhancing In changing e-commerce situations, the model's capacity to yield practical insights for client retention strategies.

References

[1]. J. P. Dias and H. S. Ferreira, “Automating the Extraction of Static Content and Dynamic

Behaviour from e-Commerce Websites," in *Procedia Computer Science*, 2017. doi: 10.1016/j.procs.2017.05.355.

[2]. M. Voss, "Impact of customer integration on project portfolio management and its success—Developing a conceptual framework," *Int. J. Proj. Manag.*, vol. 30, no. 5, pp. 567–581, Jul. 2012, doi: 10.1016/j.ijproman.2012.01.017.

[3]. K. Coussement and K. W. De Bock, "Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning," *J. Bus. Res.*, 2013, doi: 10.1016/j.jbusres.2012.12.008.

[4]. A. M. Tartaglione, Y. Cavacece, G. Russo, and G. Granata, "A systematic mapping study on customer loyalty and brand management," *Adm. Sci.*, 2019, doi: 10.3390/admsci9010008.

[5]. A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0191-6.

[6]. A. S. Al-Adwan and M. A. Al-Horani, "Boosting customer e-loyalty: An extended scale of online service quality," *Inf.*, 2019, doi: 10.3390/info10120380.

[7]. N. H. M. Ariffin, "The Development of a Strategic CRM-i Framework: Case Study in Public Institutions of Higher Learning," *Procedia - Soc. Behav. Sci.*, 2013, doi: 10.1016/j.sbspro.2013.04.005.

[8]. V. Kolluri, "A Comprehensive Analysis on Explainable and Ethical Machine: Demystifying Advances in Artificial Intelligence," *Int. Res. J.*, vol. 2, no. 7, 2015.

[9]. J. Sujata, D. Aniket, and M. Mahasingh, "Artificial intelligence tools for enhancing customer experience," *Int. J. Recent Technol. Eng.*, 2019, doi: 10.35940/ijrte.B1130.0782S319.

[10]. K. L. J. Shannahan, R. J. Shannahan, and A. Alexandrov, "Strategic Orientation and Customer Relationship Management: A Contingency Framework of CRM Success," *J. Comp. Int. Manag.*, 2010.

[11]. 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*, 2019, doi: 10.1109/ACCESS.2019.2914999.

[12]. K. Ebrah and S. Elnasir, "Churn prediction using machine learning and recommendations plans for telecoms," *J. Comput. Commun.*, vol. 7, no. 11, pp. 33–53, 2019.

[13]. M. Saghir, Z. Bibi, S. Bashir, and F. H. Khan, "Churn Prediction using Neural Network based Individual and Ensemble Models," in *Proceedings of 2019 16th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2019*, 2019. doi: 10.1109/IBCAST.2019.8667113.

[14]. S. F. Sabbeh, "Machine-learning techniques for customer retention: A comparative study," *Int. J. Adv. Comput. Sci. Appl.*, 2018, doi: 10.14569/IJACSA.2018.090238.

[15]. H. Li, D. Yang, L. Yang, Y. Lu, and X. Lin, "Supervised massive data analysis for telecommunication customer churn prediction," in *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, 2016. doi: 10.1109/BDCloud-SocialCom-SustainCom.2016.35.

[16]. R. Tarafdar and Y. Han, "Finding Majority for Integer Elements," *J. Comput. Sci. Coll.*, vol. 33, no. 5, pp. 187–191, 2018.

[17]. G. Biau, "Analysis of a Random Forests Model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012, [Online]. Available: https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

[18]. H. M. Gomes *et al.*, "Adaptive random forests for evolving data stream classification," *Mach. Learn.*, 2017, doi: 10.1007/s10994-017-5642-8.

[19]. K. Lu, X. Zhao, and B. Wang, "A Study on Mobile Customer Churn Based on Learning from Soft Label Proportions," in *Procedia Computer Science*, 2019. doi: 10.1016/j.procs.2019.12.005.

[20]. Bodepudi, V., & Chinta, P. C. R. (2024). Enhancing Financial Predictions Based on Bitcoin Prices using Big Data and Deep Learning Approach. *Available at SSRN 5112132*.

[21]. Chinta, P. C. R. (2023). The Art of Business Analysis in Information Management Projects: Best Practices and Insights. *DOI*, *10*.

[22]. Chinta, P. C. R., & Katnapally, N. (2021). Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures. *Neural Network-Based Risk Assessment for Cybersecurity in Big Data-Oriented ERP Infrastructures*.

[23]. Katnapally, N., Chinta, P. C. R., Routhu, K. K., Velaga, V., Bodepudi, V., & Karaka, L. M. (2021). Leveraging Big Data Analytics and Machine Learning Techniques for Sentiment Analysis of Amazon Product Reviews in Business Insights. *American Journal of Computing and Engineering*, *4*(2), 35-51.

[24]. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., Bodepudi, V., & Maka, S. R. (2025). Building an Intelligent Phishing Email Detection System Using Machine Learning and Feature Engineering. *European Journal of Applied Science, Engineering and Technology*, *3*(2), 41-54.

[25]. Moore, C. (2024). Enhancing Network Security With Artificial Intelligence Based Traffic Anomaly Detection In Big Data Systems. Available at SSRN 5103209.

[26]. Chinta, P. C. R., Moore, C. S., Karaka, L. M., Sakuru, M., & Bodepudi, V. (2025). Predictive Analytics for Disease Diagnosis: A Study on

Healthcare Data with Machine Learning Algorithms and Big Data. *J Cancer Sci*, *10*(1), 1.

[27]. Chinta, P. C. R., Jha, K. M., Velaga, V., Moore, C., Routhu, K., & SADARAM, G. (2024). Harnessing Big Data and AI-Driven ERP Systems to Enhance Cybersecurity Resilience in Real-Time Threat Environments. *Available at SSRN 5151788*.

[28]. Chinta, P. C. R. (2023). Leveraging Machine Learning Techniques for Predictive Analysis in Merger and Acquisition (M&A). *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

[29]. Chinta, P. C. R. (2022). Enhancing Supply Chain Efficiency and Performance Through ERP Optimisation Strategies. *Journal of Artificial Intelligence & Cloud Computing*, *1*(4), 10-47363.

[30]. Chinta, P. C. R., & Karaka, L. M. AGENTIC AI AND REINFORCEMENT LEARNING: TOWARDS MORE AUTONOMOUS AND ADAPTIVE AI SYSTEMS.

[31]. Sadaram, G., Karaka, L. M., Maka, S. R., Sakuru, M., Boppana, S. B., & Katnapally, N. (2024). AI-Powered Cyber Threat Detection: Leveraging Machine Learning for Real-Time Anomaly Identification and Threat Mitigation. MSW Management Journal, 34(2), 788-803.

[32]. Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[33]. Sadaram, G., Sakuru, M., Karaka, L. M., Reddy, M. S., Bodepudi, V., Boppana, S. B., & Maka, S. R. (2022). Internet of Things (IoT) Cybersecurity Enhancement through Artificial Intelligence: A Study on Intrusion Detection Systems. *Universal Library of Engineering Technology*, (2022).

[34]. Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. *J Cardiobiol*, *9*(1), 1.

[35]. Maka, S. R. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Available at SSRN 5116707*.

[36]. Karaka, L. M. (2021). Optimising Product Enhancements Strategic Approaches to Managing Complexity. *Available at SSRN 5147875*.

[37]. KishanKumar Routhu, A. D. P. Risk Management in Enterprise Merger and Acquisition (M&A): A Review of Approaches and Best Practices.

[38]. Routhu, KishanKumar & Katnapally, Niharika & Sakuru, Manikanth. (2023). Machine Learning for Cyber Defense: A Comparative Analysis of Supervised and Unsupervised Learning Approaches. Journal for ReAttach Therapy and Developmental Diversities. 6. 10.53555/jrtdd.v6i10s(2).3481.

[39]. Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2022). Enhancing Early Diagnosis: Machine Learning Applications in Diabetes Prediction. *Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-205. DOI: doi. org/10.47363/JAICC/2022 (1)*, *191*, 2-7.

[40]. Chinta, Purna Chandra Rao & Moore, Chethan Sriharsha. (2023). Cloud-Based AI and Big Data Analytics for Real-Time Business Decision-Making. 36. 96-123. 10.47363/JAICC/2023.

[41]. Kuraku, D. S., & Kalla, D. (2023). Phishing Website URL's Detection Using NLP and Machine Learning Techniques. *Journal on Artificial Intelligence-Tech Science*.

[42]. Krishna Madhav, J., Varun, B., Niharika, K., Srinivasa Rao, M., & Laxmana Murthy, K. (2023). Optimising Sales Forecasts in ERP Systems Using Machine Learning and Predictive Analytics. *J Contemp Edu Theo Artific Intel: JCETAI-104*.

[43]. Jha, K. M., Velaga, V., Routhu, K., Sadaram, G., Boppana, S. B., & Katnapally, N. (2025). Transforming Supply Chain Performance Based on Electronic Data Interchange (EDI) Integration: A Detailed Analysis. *European Journal of Applied Science, Engineering and Technology*, *3*(2), 25-40.

[44]. Kuraku, D. S., & Kalla, D. (2023). Phishing Website URL's Detection Using NLP and Machine Learning Techniques. *Journal on Artificial Intelligence-Tech Science*.

[45]. Jha, K. M., Velaga, V., Routhu, K. K., Sadaram, G., & Boppana, S. B. (2025). Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. *J Cardiobiol*, *9*(1), 1.

[46]. KishanKumar Routhu, A. D. P. Risk Management in Enterprise Merger and Acquisition (M&A): A Review of Approaches and Best Practices.

[47]. Kalla, D., Mohammed, A. S., Boddapati, V. N., Jiwani, N., & Kiruthiga, T. (2024, November). Investigating the Impact of Heuristic Algorithms on Cyberthreat Detection. In *2024 2nd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)* (Vol. 1, pp. 450-455). IEEE.

[48]. Bodepudi, V. (2023). Understanding the Fundamentals of Digital Transformation in Financial Services: Drivers and Strategic Insights. *Journal of Artificial Intelligence and Big Data*, *3*(1), 10-31586.

[49]. Kalla, D., Smith, N., & Samaah, F. (2025). Deep Learning-Based Sentiment Analysis: Enhancing IMDb Review Classification with LSTM Models. *Available at SSRN 5103558*.

[50]. Jha, K. M., Bodepudi, V., Boppana, S. B., Katnapally, N., Maka, S. R., & Sakuru, M. Deep Learning-Enabled Big Data Analytics for Cybersecurity Threat Detection in ERP Ecosystems.

[51]. Kalla, D., Samaah, F., Kuraku, S., & Smith, N. (2023). Phishing detection implementation using databricks and artificial Intelligence. *International Journal of Computer Applications*, *185*(11), 1-11.

[52]. Boppana, S. B., Moore, C. S., Bodepudi, V., Jha, K. M., Maka, S. R., & Sadaram, G. AI And ML Applications In Big Data Analytics: Transforming ERP Security Models For Modern Enterprises.

[53]. Sreeramulu, M. D., Mohammed, A. S., Kalla, D., Boddapati, N., & Natarajan, Y. (2024, September). AI-driven Dynamic Workload Balancing for Real-time Applications on Cloud Infrastructure. In *2024 7th International Conference on Contemporary Computing and Informatics (IC3I)* (Vol. 7, pp. 1660-1665). IEEE.

[54]. Kalla, D., & Samaah, F. (2023). Exploring Artificial Intelligence And Data-Driven Techniques For Anomaly Detection In Cloud Security. *Available at SSRN 5045491*.

[55]. Kalla, D., Smith, N., & Samaah, F. (2023). Satellite Image Processing Using Azure Databricks and Residual Neural Network.