ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

An Integrated Machine Learning Model for E-Commerce Churn Prediction

Nashak Danaan[1]; Abraham Deme[2]; Gideon Bibu[3];
Mustapha Abdulrahman Lawal[4]; Ismail Zahraddeen Yakubu[5]

[1,2,3]Department of Computer Science, Faculty of Natural Sciences, University of Jos, P.M.B 2084 Jos, Nigeria.
[4]Department of Computer Science, Abubakar Tafawa Balewa University Bauchi, Nigeria.
[5]Department of Computing SRM Institute of Science and Technology, Chennai, India

Publication Date: 2025/07/07

Abstract: Customer churn is a major challenge in the e-commerce industry, where customers end their relationship with an online business due to reasons like dissatisfaction with product quality, poor customer service, pricing concerns, fierce competition, or changing preferences. This study introduces an integrated machine learning approach to predict customer churn in e-commerce, combining k-means clustering for customer segmentation and XGBoost for classification within the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The model aims to deliver a comprehensive, stable, and reliable churn-prediction solution by analyzing customer data such as purchase history and demographics. The methodology ensures a thorough and insightful analysis of customer data to improve prediction accuracy. The model achieved an accuracy of 98.68%, precision of 96.19%, recall of 94.39%, and F1 score of 95%, outperforming individual algorithms used in earlier or similar studies. These results demonstrate the effectiveness of the integrated approach in predicting customer churn and offer valuable insights for e-commerce businesses, highlighting the importance of using advanced machine-learning techniques to boost customer retention and profitability. The study adds to the less-explored area of churn prediction in e-commerce and shows the potential of combined machine learning approaches to solve this critical issue.

Keywords: *Customer Churn, E-Commerce, Machine Learning.*

How to Cite: Nashak Danaan; Abraham Deme; Gideon Bibu; Mustapha Abdulrahman Lawal; Ismail Zahraddeen Yakubu; (2025) An Integrated Machine Learning Model for E-Commerce Churn Prediction. *International Journal of Innovative Science and Research Technology*, 10(6), 2762-2778. https://doi.org/10.38124/ijisrt/25jun1816

I. INTRODUCTION

A. Background of the Study

Electronic commerce (e-commerce), as defined by Jain et al. (2021), leverages electronic media and the Internet to buy and sell goods and services, providing consumers with accessibility and comparative advantages across multiple platforms. This transformation in retail has introduced both opportunities and challenges, with one notable challenge being customer churn. Churn refers to customers discontinuing their relationship with an online business because of dissatisfaction, poor service, pricing issues, or changing preferences (Chinnu et al., 2017). Churn can occur voluntarily or involuntarily, necessitating proactive strategies for its identification and management.

Provost and Fawcett (2013) highlighted that predictive techniques are crucial for addressing churn. These techniques range from traditional statistical methods to advanced machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and artificial neural networks (ANN). Owing to its flexibility and ability to optimize predictive accuracy, machine learning has shown superiority over traditional approaches by adapting to evolving datasets and improving the model performance over time.

Integrating multiple machine learning algorithms in sectors such as banking and telecommunications has demonstrated a higher predictive accuracy (Hu et al., 2018; Zhang et al., 2022; Xiahou & Harada, 2022). This integrated approach combines algorithms such as SVM, XGBoost, ANN, KNN, and LR, often enhanced by k-means clustering techniques for data segmentation. Such methodologies aim to develop comprehensive churn prediction models capable of handling complex customer behavior patterns.

This study aims to contribute to e-commerce churn prediction by analyzing customer data attributes such as purchase history and demographics. The most effective algorithms will be explored and combined through the CRISP-DM workflow to develop robust predictive models. By comparing the performance of these models with existing

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

benchmarks, this study seeks to provide actionable insights and recommendations to mitigate churn and enhance customer retention strategies in e-commerce settings.

The remainder of this paper is organized as follows: Section 2 presents related work on customer churn prediction in telecommunication, banking, and e-commerce, and Section 3 discusses the methodology and implementation details of the proposed model. Section 4 covers the study's results and findings, and Section 5 discusses the conclusions and future work.

II. RELATED WORK

A. Churn Prediction in the Telecommunication Sector

Multiple studies have explored diverse machine-learning techniques for predicting customer churn in the telecommunications industry. Lalwani et al. (2021) compared algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), Decision Trees, Random Forest, XGBoost, CatBoost, AdaBoost, and Extra Tree, with AdaBoost and XGBoost achieving the highest accuracy and AUC scores. Singh et al. (2022) applied Gaussian Naive Bayes, Extra Tree, K-nearest neighbors (KNN), Random Forest, Neural Network, Light GBM, XGBoost, and Logistic Regression, highlighting logistic regression's top accuracy of 80.49%. Rani and Kant (2020) emphasized Gradient Boosting and e-Xtreme Gradient Boosting classifiers in supervised and semi-supervised learning contexts, achieving notable accuracy improvements with pseudo-label techniques. Mittal (2022) focused on the efficacy of logistic regression in predicting churn based on customer behavior and service usage patterns. Ahmad et al. (2019) highlighted the capability of XGBoost to manage high-dimensional data and effectively capture nonlinear relationships for churn prediction. Ullah et al. (2019) utilized Neural Networks to achieve approximately 90% accuracy in churn prediction. Caigny (2018) proposed an integrated learning strategy combining neural networks, random forests, and SVMs to enhance prediction accuracy by leveraging the strengths of each algorithm.

B. Churn Prediction in the Banking Sector

Research on customer churn prediction in the banking and financial sectors has leveraged machine learning techniques to improve prediction accuracy. Gafari and Osman (2021) employed XGBoost and Random Forest in a B2B context, achieving an impressive 99.78% accuracy with XGBoost. Tran et al. (2023) focused on customer segmentation, finding that Random Forest achieved 97.4% and 97% accuracy for sample and cluster mean accuracy, respectively, outperforming logistic regression post-segmentation. Durkaya et al. (2023) emphasized the importance of feature selection using Support Vector Machines to identify key variables in churn prediction. Yahaya et al. (2021) demonstrated the effectiveness of Artificial Neural Networks (ANNs) in predicting churn within segmented customer groups. Long et al. (2020) developed a hybrid model combining deep neural networks, decision trees, and logistic regression to enhance predictive performance and interpretability. Gonzalez-Rodriguez et al. (2019) improved churn prediction by integrating recurrent neural networks and gradient boosting, effectively modeling temporal dynamics in customer behavior. Shirazi et al. (2019) highlighted the benefits of using external data sources, such as social media activity and economic indicators, to enhance churn prediction accuracy in the e-commerce sector.

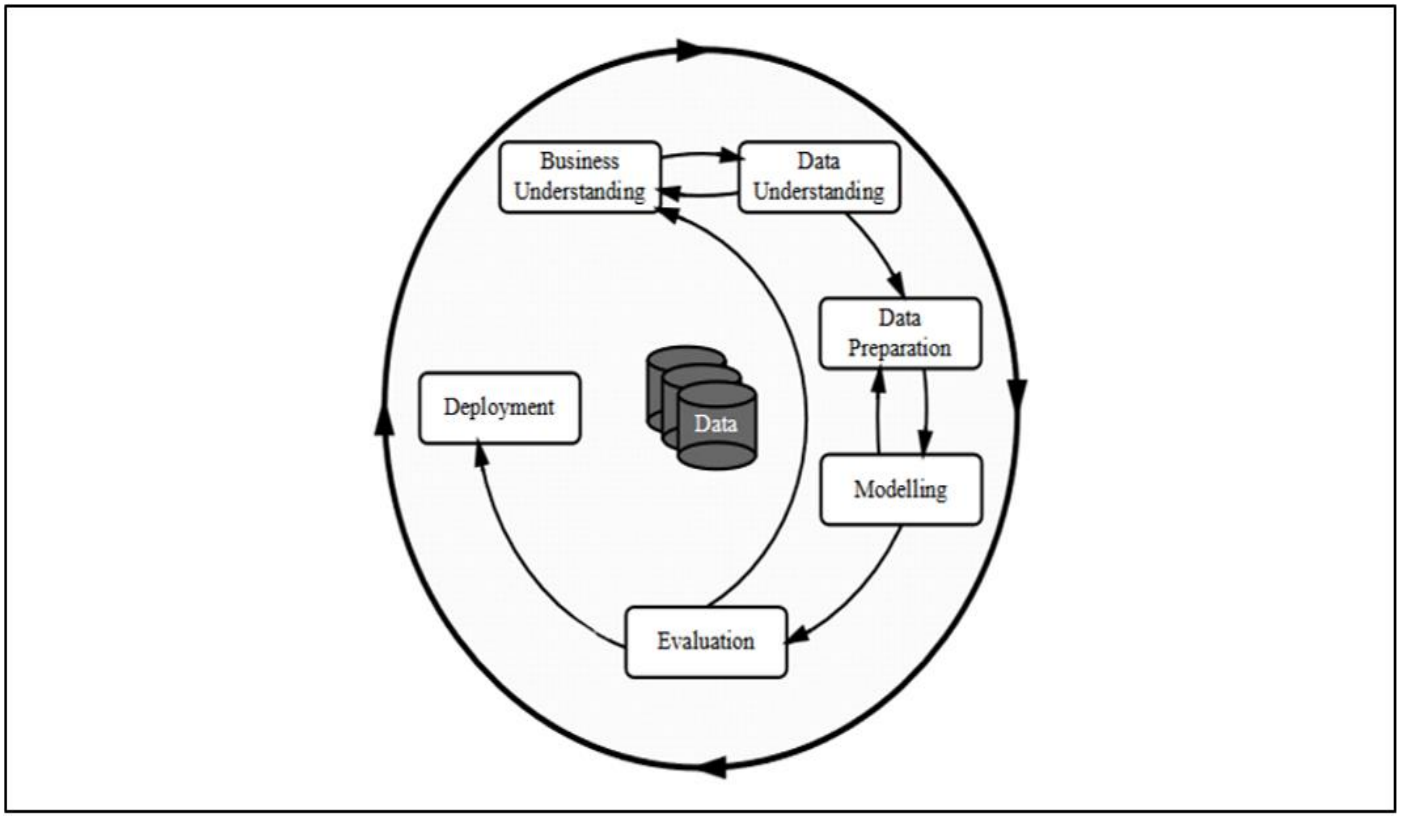

Fig 1 CRISP-DM Cycle. (Chapman et al., 1999)

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

C. Churn Prediction in the E-Commerce Sector

Several studies have contributed to customer churn prediction using diverse methodologies and techniques. After extensive exploratory analysis and data visualization, Alshamsi (2022) employed a Decision Tree, Logistic Regression, and Random Forest, revealing associations between churned customers and male gender and single marital status. Random Forest achieved the highest accuracy and kappa scores of 93.5% and 0.75, respectively. Wu & Meng (2016) introduced an enhanced SMOTE technique to balance datasets and tested several algorithms, including LSSVM and BP neural. Smith et al. (2012) innovatively integrated customer sentiment from social media data into churn prediction, enhancing predictive accuracy by capturing client sentiments effectively through online sentiment analysis.

D. Research Gap and Criticisms of Related Studies

Extensive research in the telecommunications and banking sectors has explored machine learning algorithms such as neural networks, SVMs, decision trees, random forests, logistic regression, and XGBoost for churn prediction, with notable performance evaluation and methodological gaps. For instance, Durkaya et al. (2023) and Mittal (2022) narrowly focused on accuracy metrics, neglecting other crucial evaluation criteria, whereas Alshamsi (2022) incorporated the kappa value and achieved 93.5% accuracy without integration. Lalwani et al. (2021) demonstrated 81.71% accuracy with XGBoost. However, they needed more comparative analysis with other algorithms, and Bindu et al. (2020) achieved a high accuracy of 99.62% but relied on pseudo-analysis without rigorous validation. Despite these advancements, e-commerce churns still need to be explored.

III. RESEARCH METHODOLOGY

➢ *Dataset Description and Exploration*

The dataset chosen for this study was sourced from the Kaggle repository. It comprises customer transaction records from an e-commerce website spanning six months, specifically from June 2021 to November 2021. It consists of 005,630 rows of data with 20 attributes.

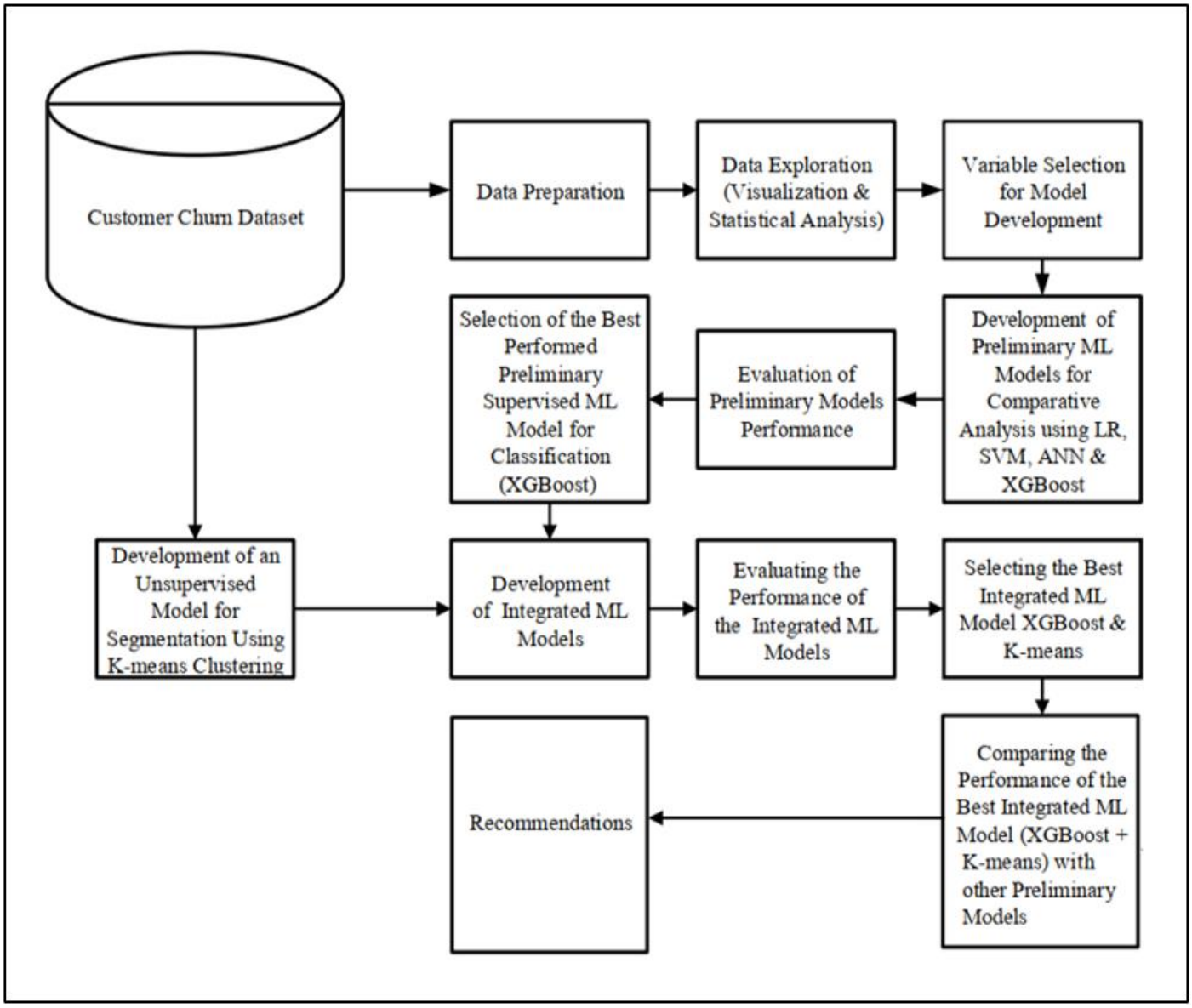

Fig 2 Machine Learning Pipeline.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

➢ *CRISP-DM Cycle.*

Fig. 1 depicts the CRISP-DM cycle, a well-established framework utilized in this research for data mining and machine learning processes. This iterative framework comprises six stages that guide structured approaches to handling data-driven projects.

Fig. 2 shows the study's machine learning pipeline, and a simple workflow diagram of the activities required to achieve its objectives.

➢ *Business Understanding*

The business understanding phase in this study involves grasping consumer behavior dynamics and identifying potential churn triggers using a dataset featuring 20 attributes, such as preferred login device, warehouse-home distance, payment method, marital status, and gender. Historical customer data were analyzed to uncover patterns contributing to churn, facilitating the creation of predictive models that estimate churn probability. The optimal model, an integrated approach, predicts customer behavior based on demographic data and attributes. This enables strategic recommendations to prioritize high-risk customers and implement retention strategies such as personalized offers, improved customer support, or product enhancements.

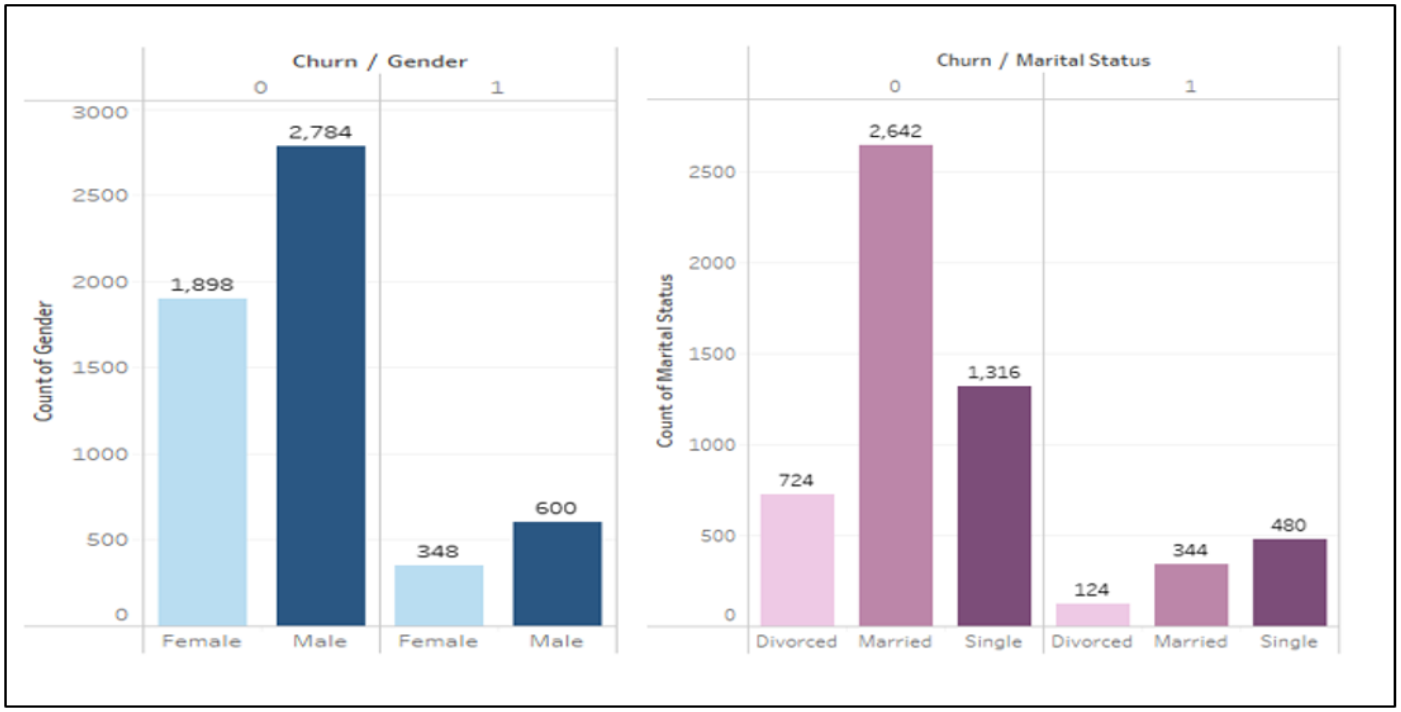

Fig 3 Churn Based on Gender and Marital Status

➢ *Data Understanding*

The data-understanding phase is essential for analyzing a dataset and laying the groundwork for data mining. It involves exploring data sources, formats, and structures, identifying challenges such as missing values and outliers, and employing exploratory data analysis techniques such as visualizations and summary statistics to uncover patterns and form hypotheses.

➢ *Data Preparation*

Eliminating discrepancies in the dataset before proceeding to the modeling phase is crucial because data preparation significantly influences the performance and fit of the models used in data mining, ensuring accurate predictions. As part of the data cleaning process, null entries were dropped from the dataset to improve predictive model outcomes and performance.

➢ *Data Cleaning*

The dataset used in this study included redundant attributes such as labels, phones, and mobile phones, which essentially convey the same information. Therefore, the 'mobile phone' was replaced with the phone to enhance precision.

➢ *Variable Transformation*

This study's primary data transformation approach involves converting categorical variables into numeric formats using a label encoder. This transformation was applied using the Scikit-learn library label encoder class, which assigns unique integers to each categorical label. Variables such as preferred login device, payment method, gender, marital status, and preferred order category were encoded into numeric values to facilitate processing and analysis.

➢ *Model Selection*

Model Selection: Selecting the most suitable model involves evaluating performance metrics, such as accuracy and cross-validation scores, to balance complexity and generalization. Algorithms such as KNN, XGBoost, ANN, Logistic Regression, K-means, and SVM were chosen based on their strengths in classification, regression, and clustering tasks.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

IV. RESULTS AND DISCUSSIONS

➢ *Churn Exploration Based on Gender and Marital Status*

Based on the visualization presented in Fig. 4, it is evident that a significant number of customers who churned were males, with a count of 600. Similarly, most of the churned customers were single, totaling 480 individuals. This suggests that single males constitute a notable segment of customers who churned, highlighting a potential demographic trend influencing customer attrition within the dataset.

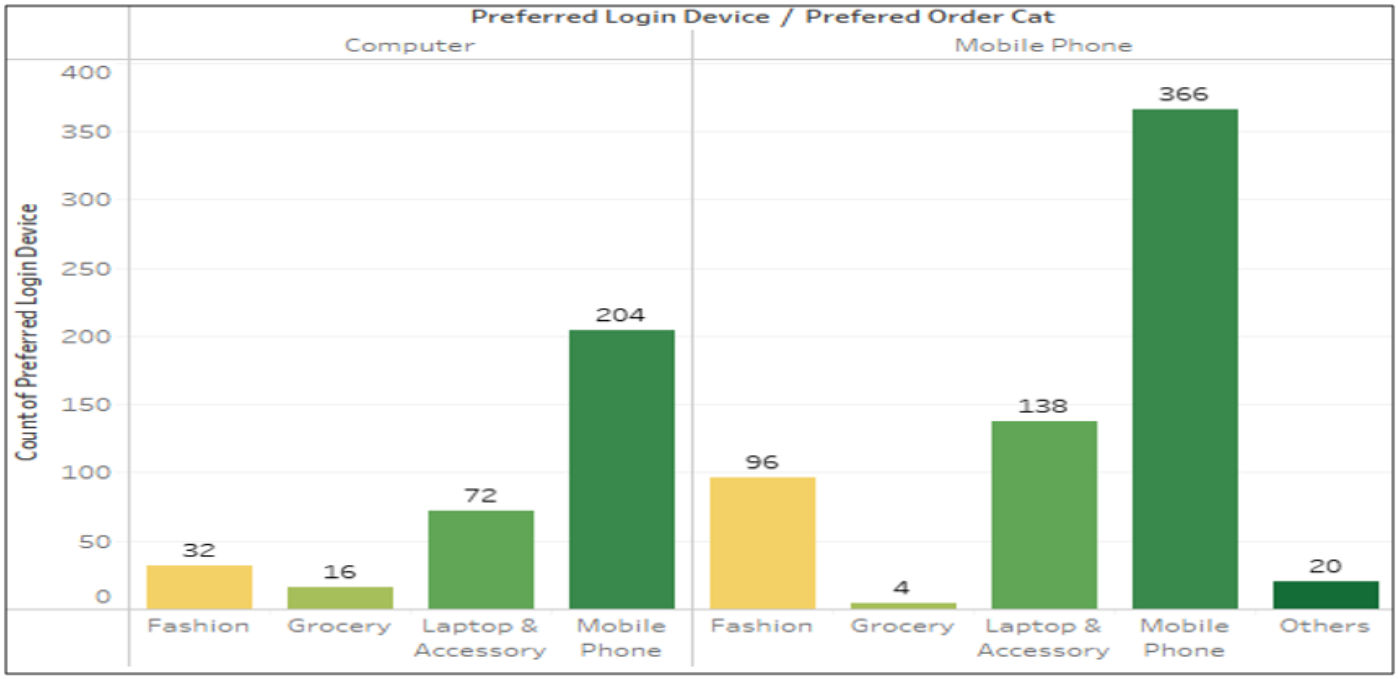

Fig 4 Churn Based on Purchase Preferences.

➢ *Exploring Churn Based on Purchase Preferences.*

The dataset contains attributes that describe customer preferences during purchases. Some instances included the preferred order cat, preferred login device, and preferred payment method. Fig. 5 shows that most customers' preferred login devices are computers, and their preferred chat device is a mobile phone. A similar pattern can be observed for customers who chose mobile phones as their preferred login devices.

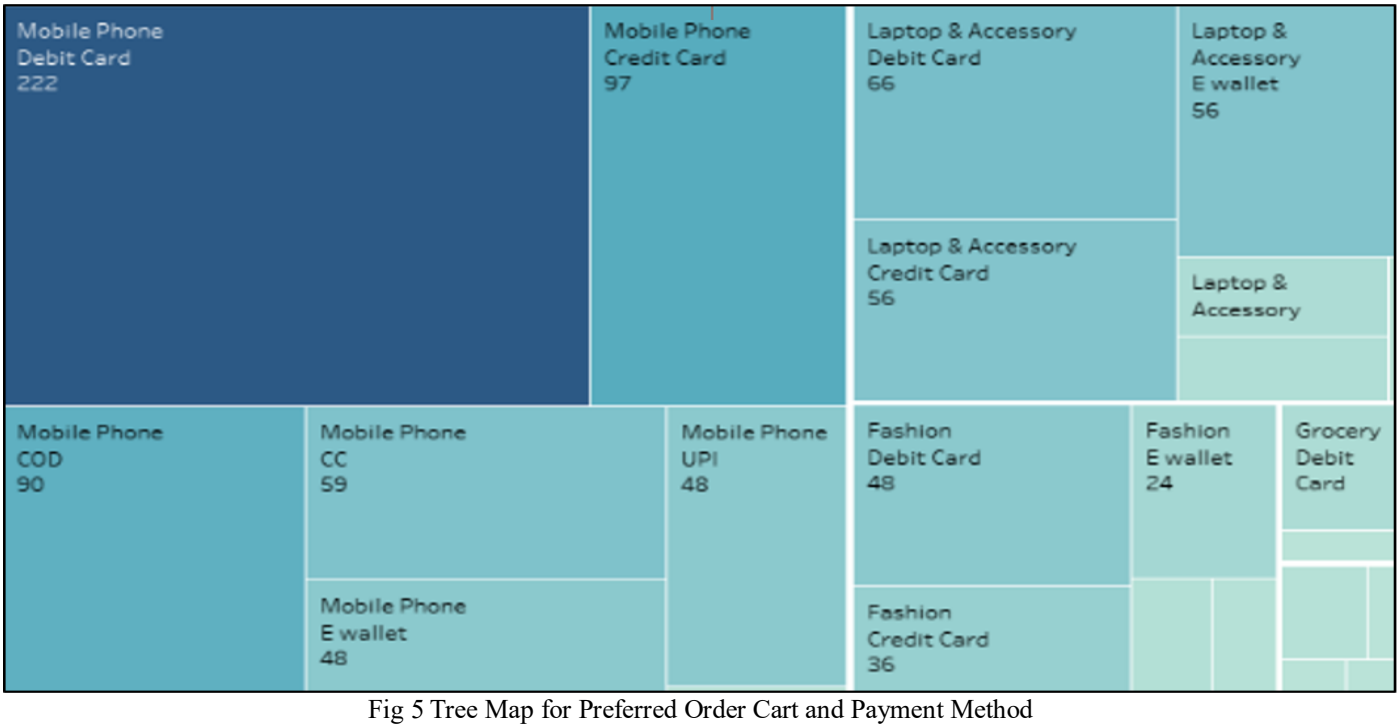

Fig 5 Tree Map for Preferred Order Cart and Payment Method

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

➢ *Exploring churn based on preferred order cart and payment methods.*

The preferred or order cart, which includes fashion, mobile phones, grocery, laptop accessories, and others about the preferred payment method, is visualized to reveal patterns that suggest churning based on the two criteria. This is illustrated in the previous tree map. As shown in Fig. 5, most customers churn based on the preferences presented, using debit cards as their preferred payment method and mobile phones as their preferred category. This is an essential insight that predictive models can be used to predict churn for new unclassified entries.

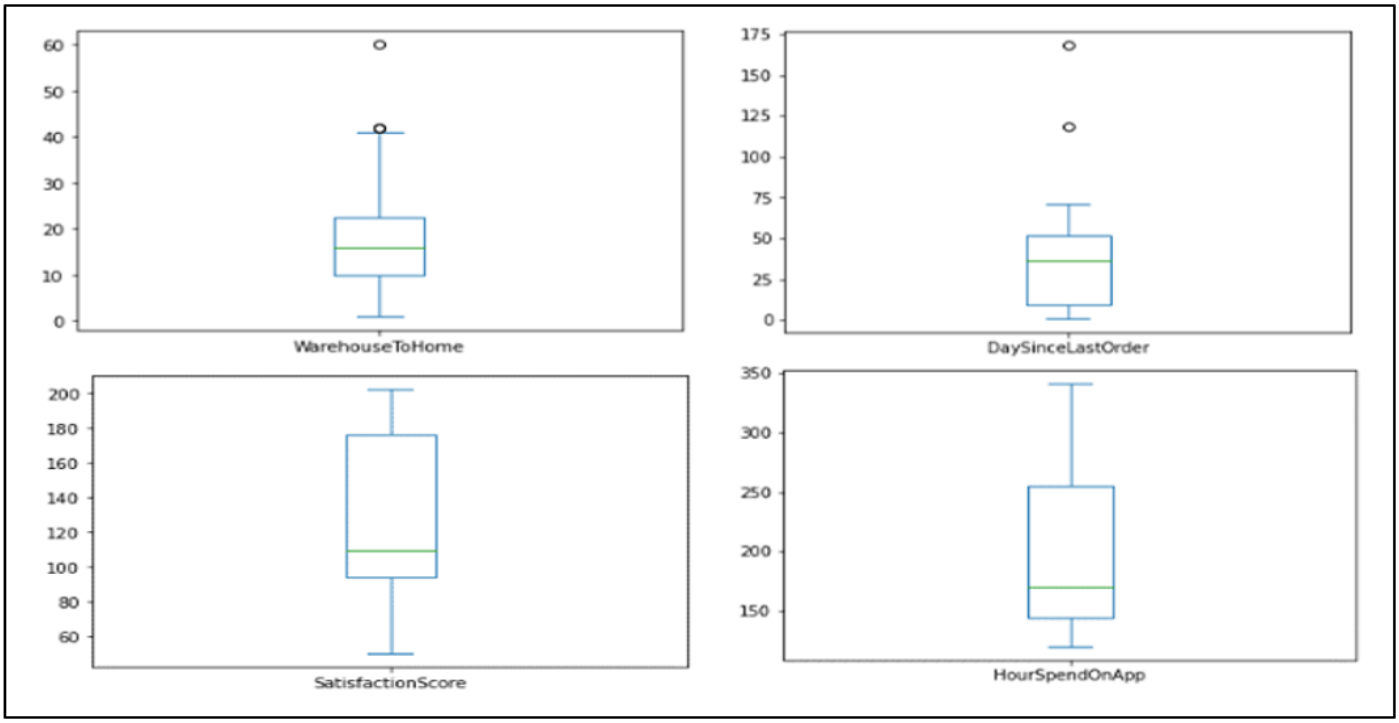

Fig 6 Box Plot Outlier Analysis.

➢ *Outlier Analysis*

Sometimes, the datasets contain entries that do not align with other values within the dataset. These entries appear to be errors and may negatively affect the overall representation of the data. Box plots were used to check for attributes that contained outliers to understand the dataset better. Fig. 7 shows the outcome of the outlier analysis of the selected continuous variables and ordinal attributes. The boxplots used for the outlier analysis show outliers in the distance from the warehouse to home and the number of days since the last order, in contrast to the hours spent on the app and satisfaction scores. It is important to note that outliers play a vital role in unsupervised learning, although they may negatively impact supervised learning.

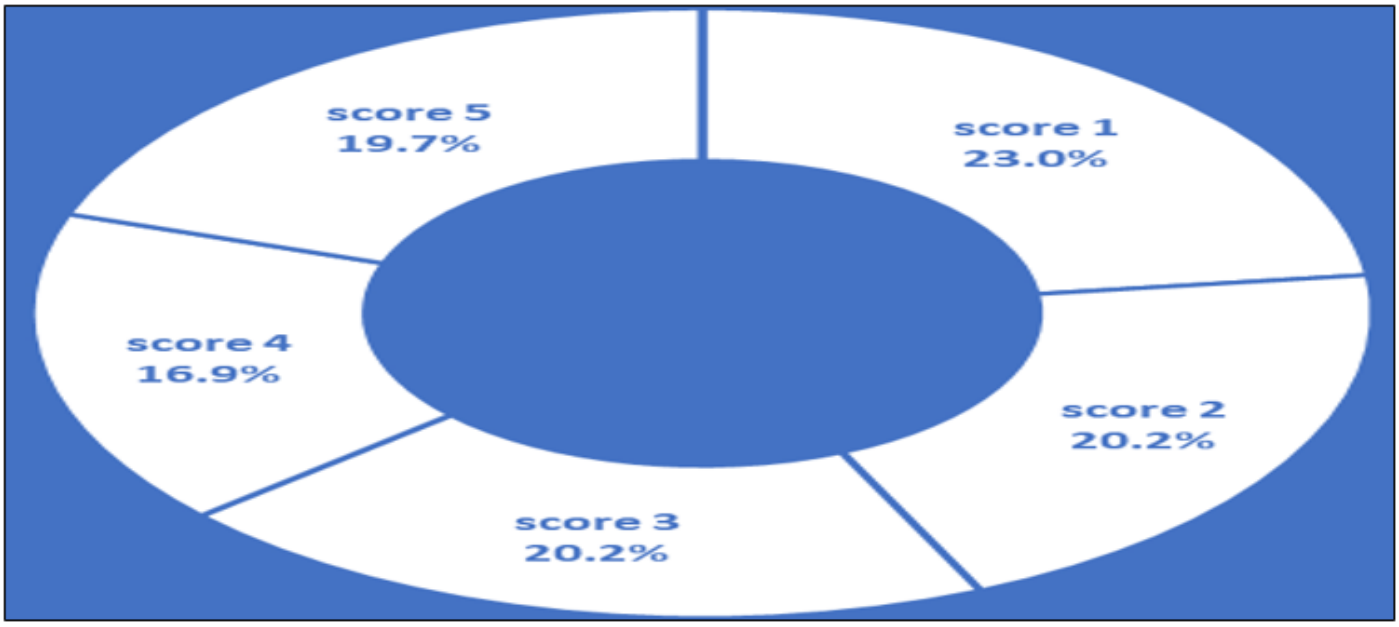

Fig 7 Pie Chart Showing Churn Based on Satisfaction Score.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

➢ *Satisfaction Score Versus Number of Complaints.*

The satisfaction score of customers with a service is hypothetically considered to correlate with the number of complaints; the relationship should be negative. This meas the satisfaction score should be relatively low for customers who complain, as seen in the following visual. As shown in Fig. 7, most customers who churned had a satisfaction score of 1 (23.0%), whereas the least satisfied customers had a score of 4 (16.9%). This is an expected outcome because it is expected that the least satisfied customers are most likely to churn.

	WarehouseToHome	HourSpendOnApp	NumberOfDeviceRegistered	SatisfactionScore	OrderAmountHikeFromlastYear
WarehouseToHome	1.000000	0.052731	0.024582	0.000434	0.031975
HourSpendOnApp	0.052731	1.000000	0.293021	0.039879	0.096827
NumberOfDeviceRegistered	0.024582	0.293021	1.000000	-0.017788	0.083342
SatisfactionScore	0.000434	0.039879	-0.017788	1.000000	-0.008143
OrderAmountHikeFromlastYear	0.031975	0.096827	0.083342	-0.008143	1.000000
CashbackAmount	-0.012433	0.131281	0.113504	0.012061	0.014215

Fig 8 Correlation Analysis for Variable Selection.

➢ *Correlation Analysis for Variable Selection*

Correlation analysis is vital for selecting variables for predictive models in machine learning because it identifies significant correlations with the target variable, helps reduce multicollinearity, and enhances model interpretability and generalization. It uncovers hidden patterns and relationships in the data, improving the understanding of its structure. Selecting features with meaningful correlations with the target can improve the accuracy of the predictive model. However, correlation does not imply causation; domain expertise is needed to avoid misinterpretation. Fig. 8 shows predominantly low positive correlations among variables, indicating minimal multicollinearity concerns in the dataset.

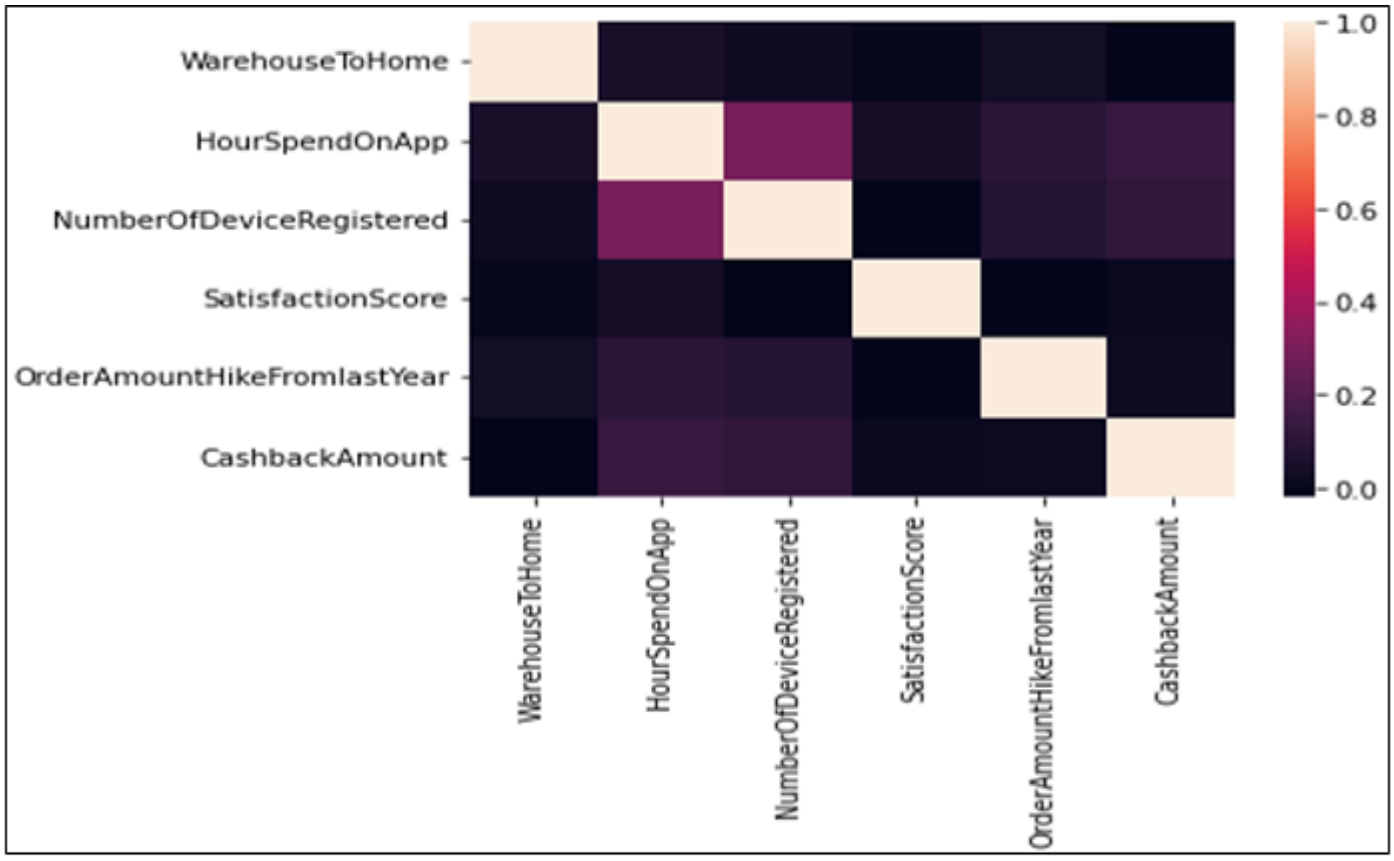

Fig 9 Correlation Analysis Heat Map.

➢ *Correlation Analysis*

Fig. 9 shows a heatmap of the results of the correlation analysis. It presents the levels chromatically in terms of absolute values of the correlation coefficient.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

➢ *Performance of the Algorithms Used in this Study.*

- *Logistic Regression*

Fig. 10 depicts the performance evaluation results. Most performance matrices could be better, emphasizing the need for improvement. Fig. 11 depicts the Confusion Matrix for the Logistic Regression Model. This matrix indicates inferior performance in predicting positive instances, which is unacceptable for all standards, with only 49 true positive predictions. Therefore, an improved model was required. Fig. 12 depicts the ROC and AUC curves, indicating that the logistic regression model does not reliably discriminate between the classes and trade-offs between true positive and true negative, as shown below, where both lines almost converge. Fig. 13 depicts the Precision-Recall Curve with the precision-recall trade-offs, which could be better.

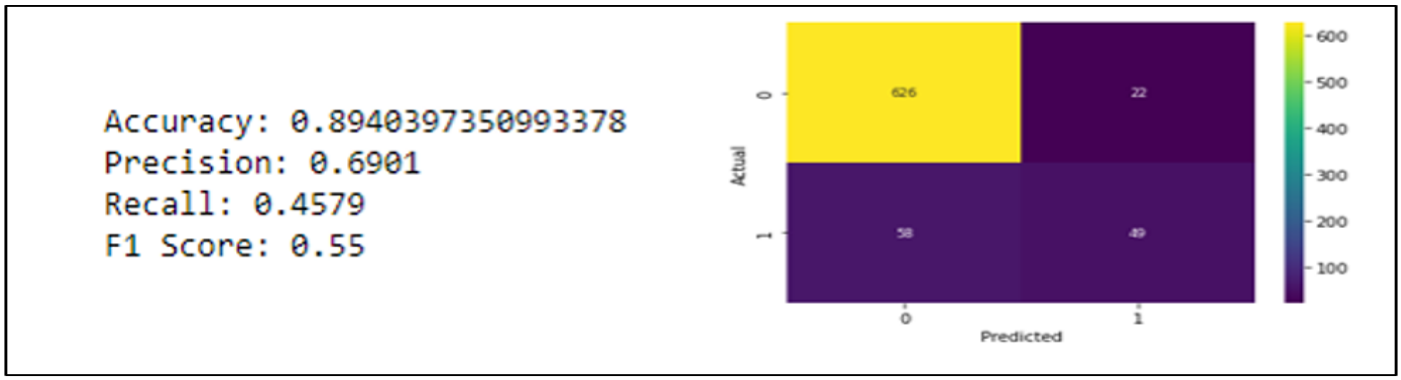

Fig 10 LR Evaluation Matrix. Fig. 11: LR Confusion Matrix

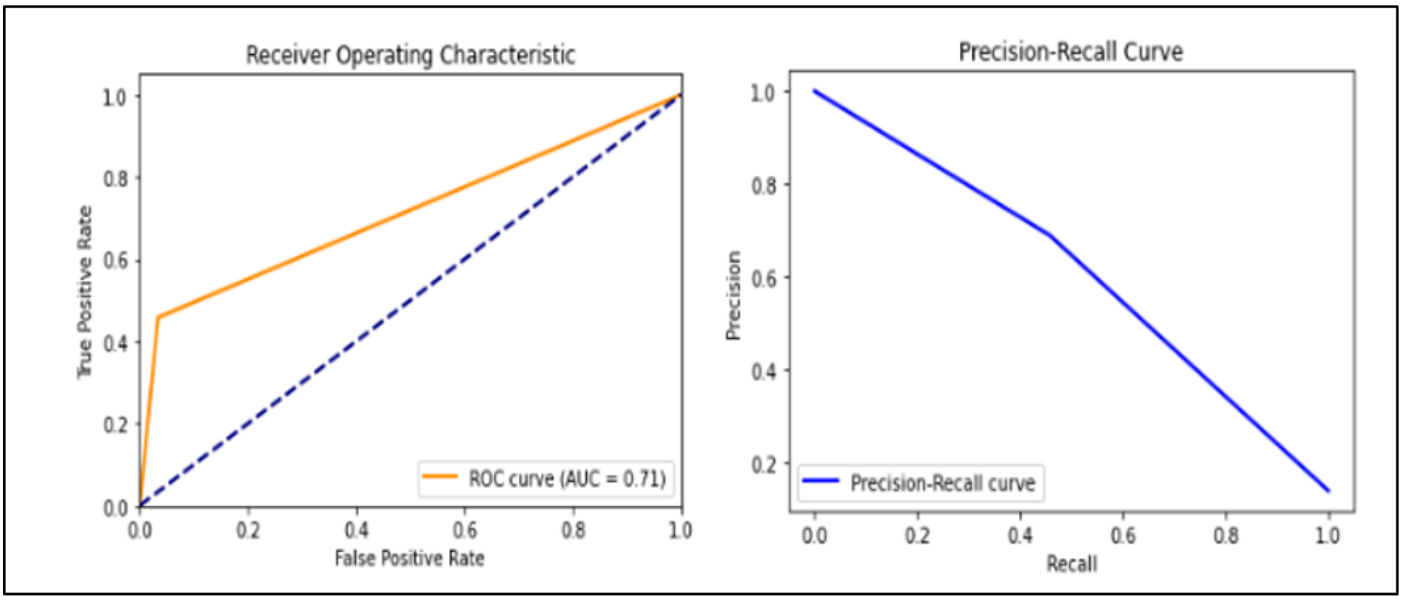

Fig 12 LR ROC and AUC Curve. Fig 13 LR Precision-Recall Curve.

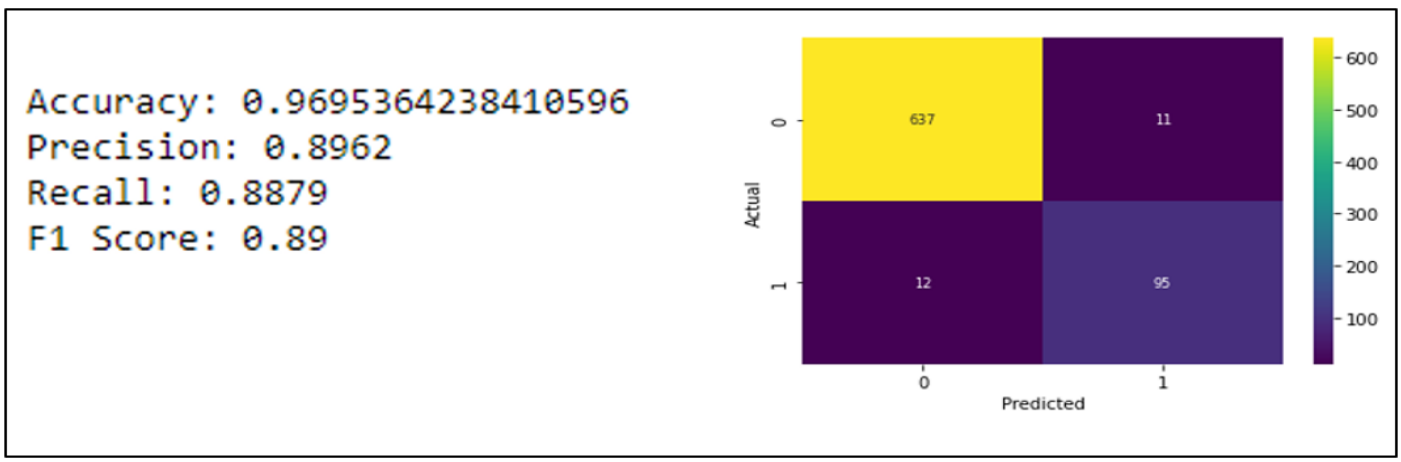

Fig 14 XG Boost Evaluation Matrix. Fig 15 XG Boost Confusion Matrix.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

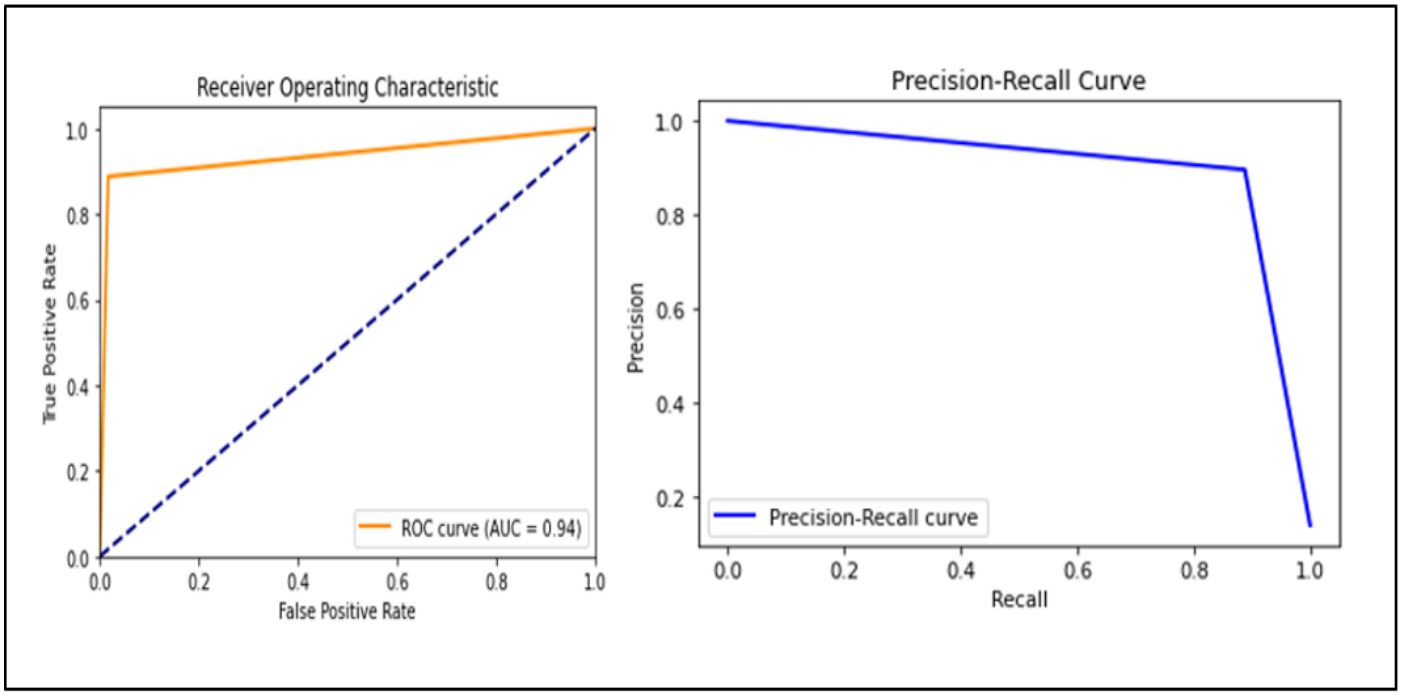

Fig 16 XG Boost ROC and AUC Curve. Fig 17 XG Boost Precision-Recall Curve

- *Extreme Gradient Boosting (XG Boost)*

Fig. 14 shows the XG Boost evaluation matrix. Compared to the LR model, all the criteria have improved, but there is still room for improvement. Fig. 15 shows the confusion matrix indicating a general improvement over the logistic regression model. Customer churn prediction increased from 49 to 95. Similarly, an increase can be observed in the true negative predictions. Fig. 16 shows the ROC and AUC curves with a clear improvement in class discrimination, especially compared to the LR model. Fig. 17 shows the precision-recall curve with a massive improvement over the LR.

- *Support Vector Machine (SVM)*

Fig. 18 shows the evaluation matrix for the SVM model. Although the model was significantly accurate, it had low precision, recall, and F1-score values. Fig. 19 shows the confusion matrix. Fig. 20 shows the ROC and AUC curves with poor class discrimination because of lower churn prediction. Fig. 21 shows the precision-recall curves indicating lower values. This model has the least true positive prediction compared to XG Boost, yet it has better true negative predictions than the logistic regression model. Although the rate of true positive prediction was significant compared to logistic regression, overall, it did not better predict churn and was considered unsuitable.

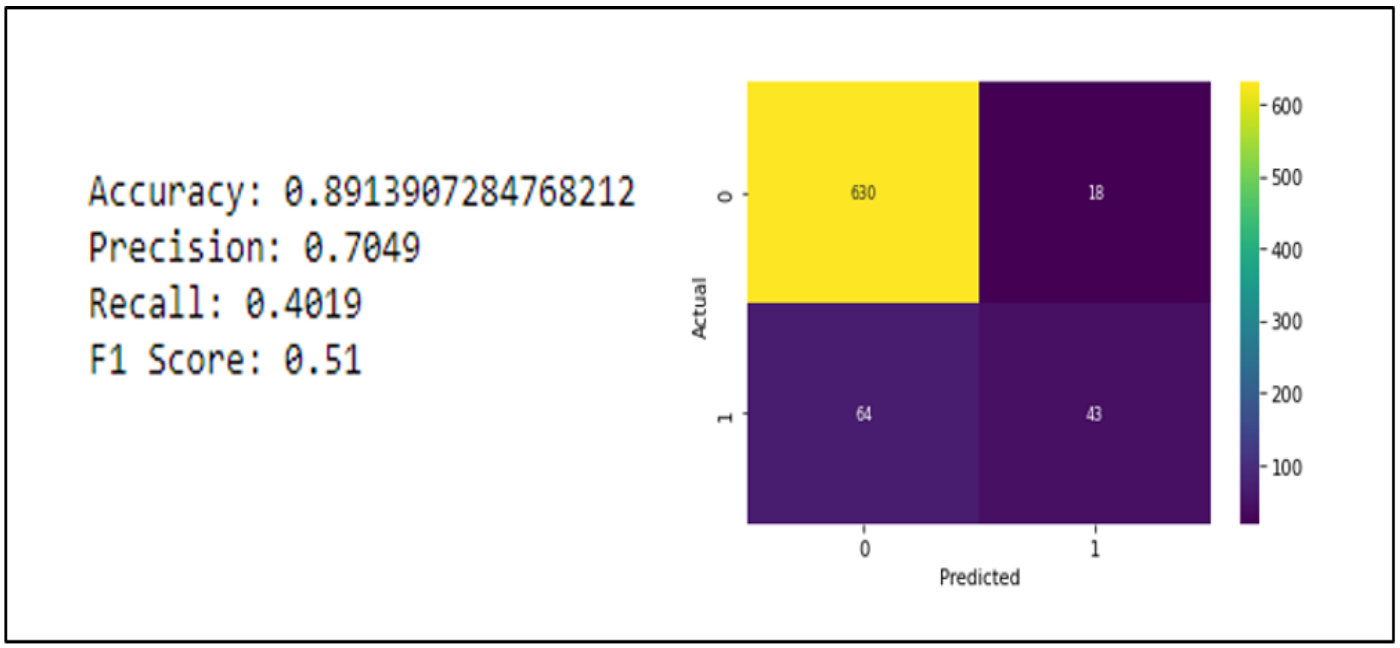

Fig 18 SVM Evaluation Matrix. Fig 19 SVM Confusion Matrix.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

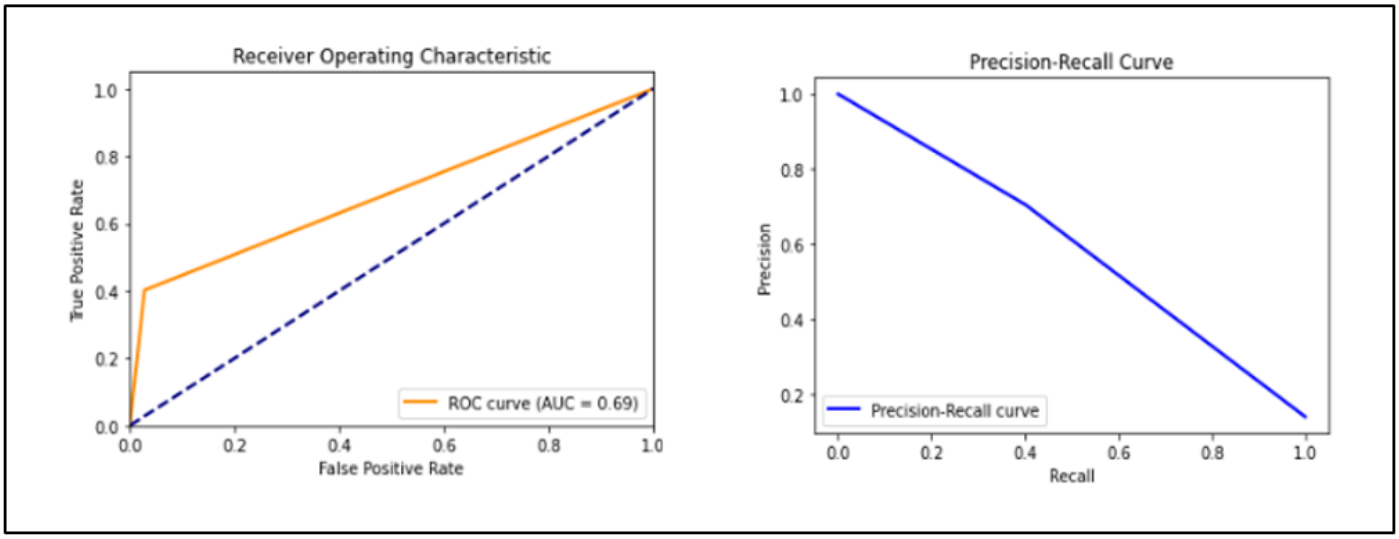

Fig 20 SVM ROC and AUC Curve

Fig 21 SVM Precision-Recall Curve

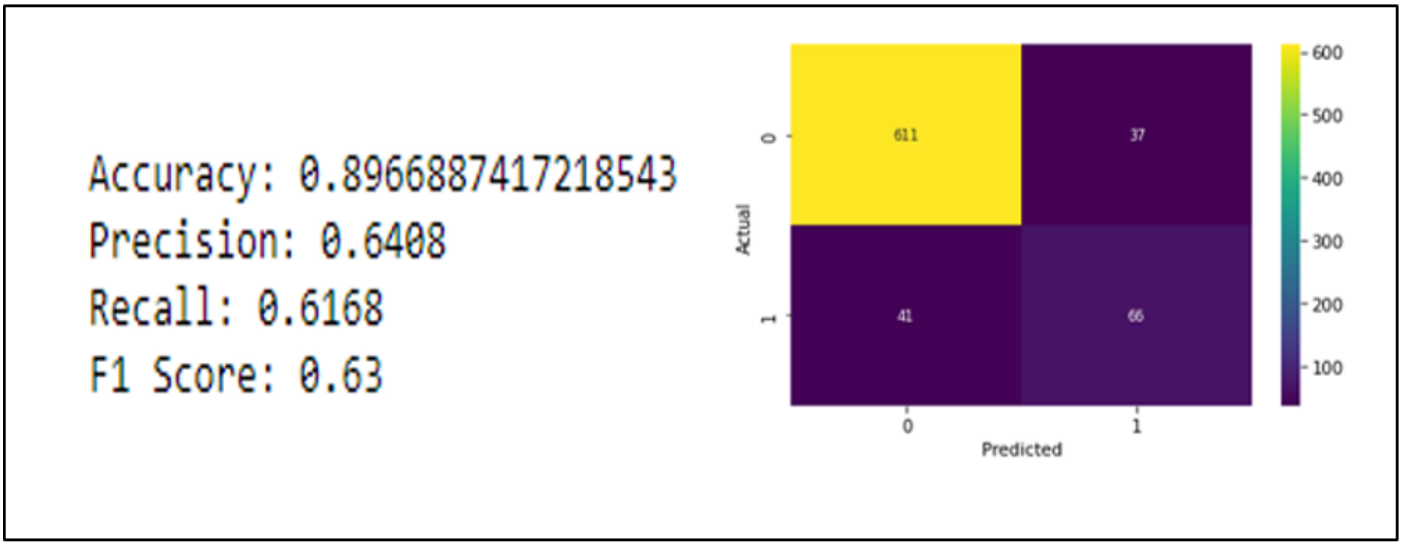

Fig 22 ANN Evaluation Matrix

Fig 23 ANN Confusion Matrix.

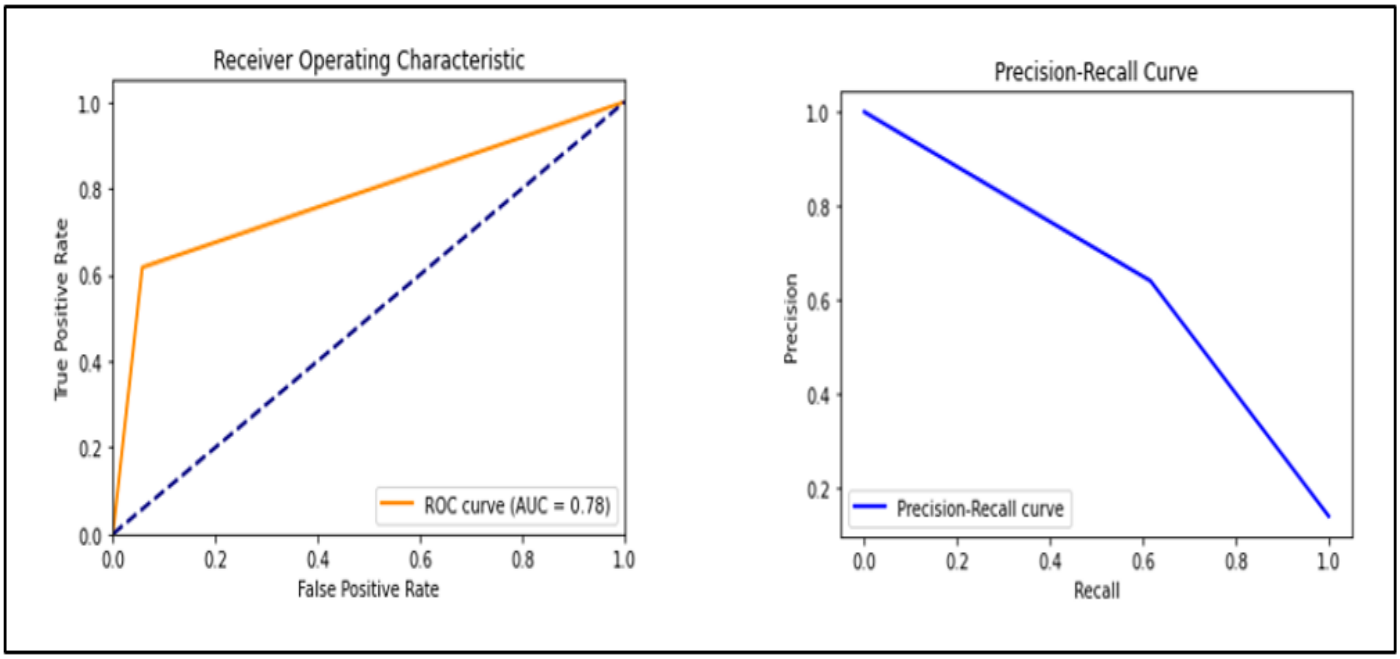

Fig 24 ANN ROC and AUC Curve

Fig 25 ANN Precision-Recall Curve

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

- *Artificial Neural Network (ANN)*

The evaluation matrix of the ANN in Fig. 22 shows a relatively inferior performance in terms of accuracy, precision, recall, and F1 score. Fig. 23 shows 611 true negative predictions and 66 true positive predictions for the ANN model. This performance is superior to the LR model, which made significant false negative and false positive predictions. Thus, the ANN model alone is unsuitable for making accurate predictions regarding this dataset, implying room for improvement. Fig. 24 shows the ANN's ROC and AUC curves, which indicate diminished performance relative to the other supervised learning models. Fig. 25 shows that the Precision-Recall curve for ANN was lower, both in precision and recall, which explains the diminished shape of the curve. An increased performance would present a curve with a more acute angle approaching 90 degrees.

- *Integrated Models*

The best-performing supervised learning model was combined with an unsupervised model to develop the integrated model. The unsupervised learning model used was k-means clustering. To perform K-means clustering, a suitable number of clusters (k) was first sorted, after which the cluster was added as an additional attribute, which was then applied to the XGBoost (best-performing supervised learning model). The processes are discussed thus.

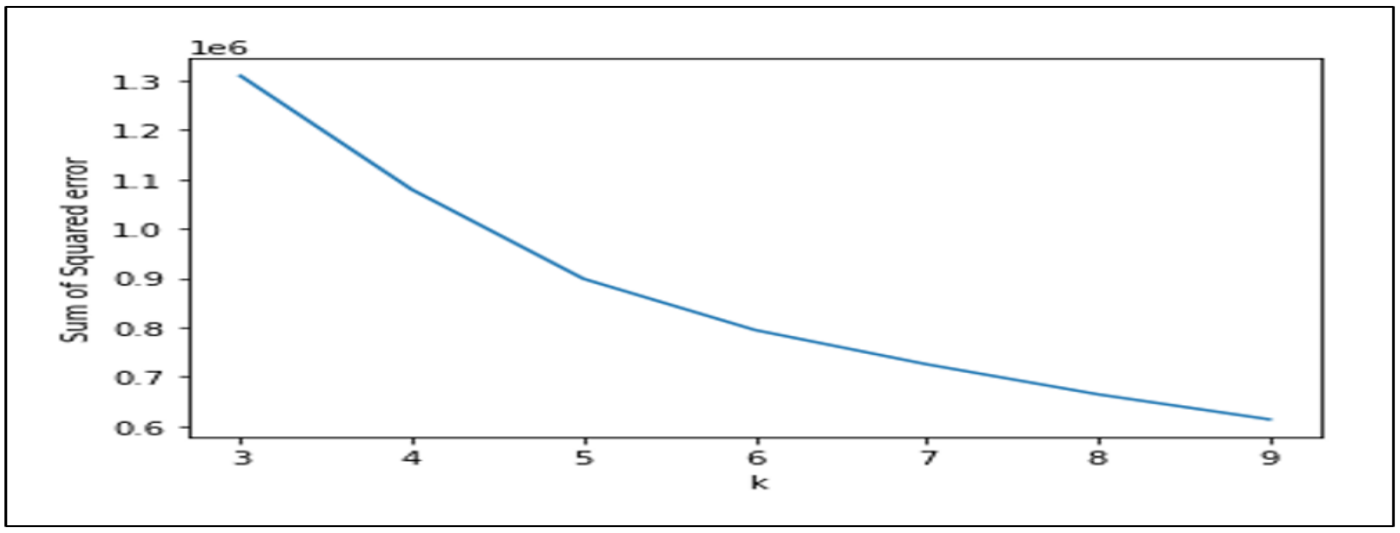

Fig 26 Elbow Method for Finding K Value

✓ *K-means clustering*

K-means clustering is a popular unsupervised machine learning algorithm used to partition a dataset into k clusters, where each data point belongs to the cluster with the nearest mean. K-means aims to minimize the within-cluster sum of squares (WCSS), which measures the variance within each cluster.

✓ *Finding a suitable value of K (elbow method).*

The elbow method helps select an appropriate value for the number of clusters (k) by observing the rate of variance reduction as the number of clusters increases. From Fig. 26, the elbow is at k = 6. Thus, the number of clusters was assigned 6, and the K-means clustering algorithm was performed.

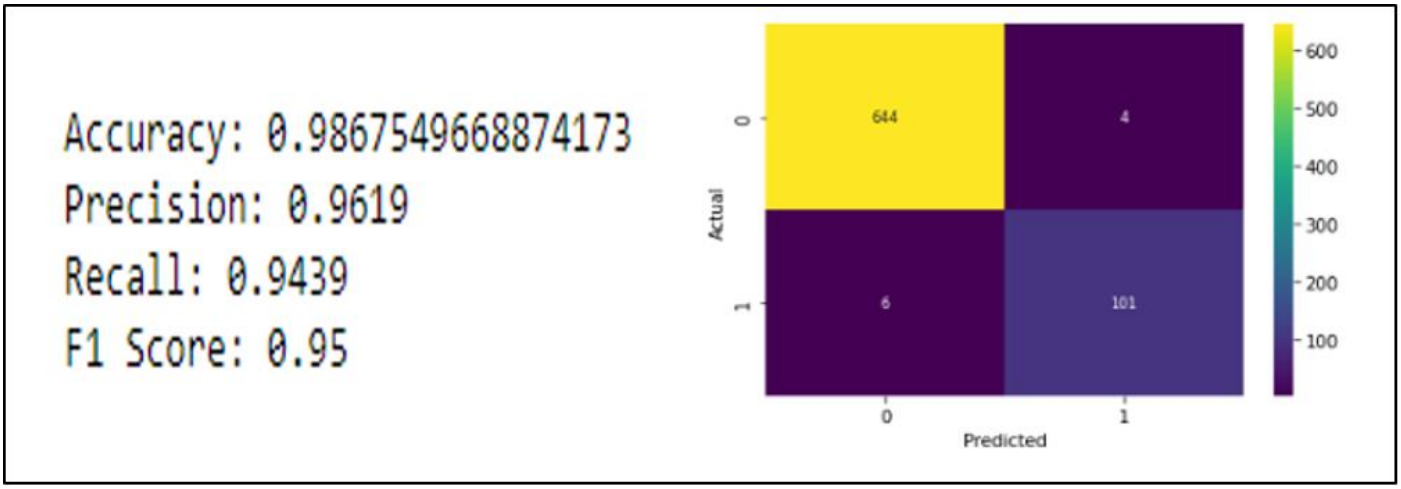

Fig 27 Integrated Model 1 Evaluation Matrix.

Fig 28 Integrated Model 1 Confusion Matrix.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

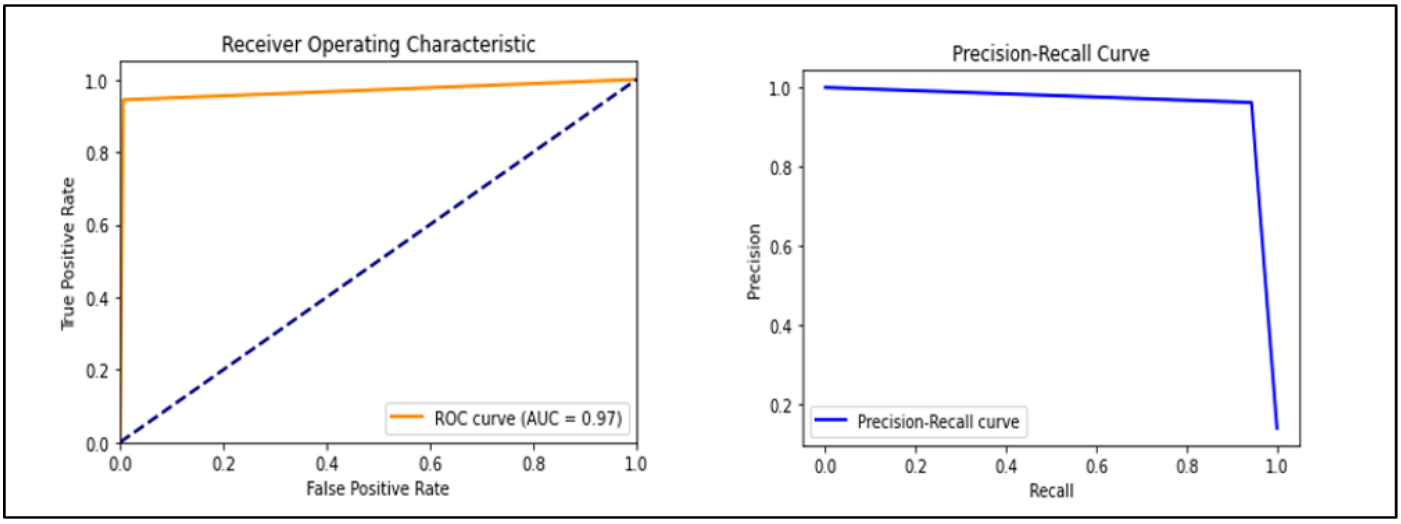

Fig 29 Integrated Model 1 ROC & AUC Curve. Fig 30 Integrated Model 1 Precision-Recall Curve.

- *Integrated Model 1*

Fig. 27 depicts the evaluation matrix for the integrated model 1. All four criteria performed superiorly to the previously discussed models. Fig. 28 depicts the confusion matrix with high true positive and negative predictions, indicating a balance. The improved churn prediction from 95 in XGBoost to 101 in the integrated model was commendable. This implies that, although the XGBoost algorithm is used in both cases, the impact of the K-means clustering is eminent. Fig. 29 is the ROC and AUC curves, which depict clear class discrimination, and a significant improvement compared to the previously discussed models. Fig. 30 shows the precision-recall curve, indicating a significant improvement from the isolated XGBoost model. This shows that the integrated model performs best when considered by all the criteria mentioned.

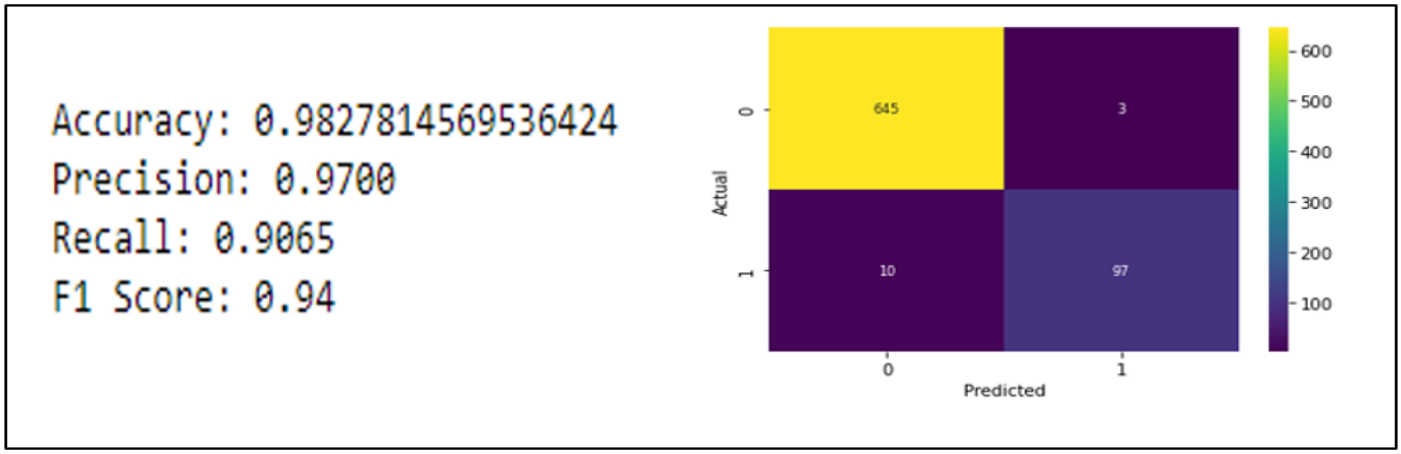

Fig 31 Integrated Model 2 Evaluation Matrix. Fig 32 Integrated Model 2 Confusion Matrix.

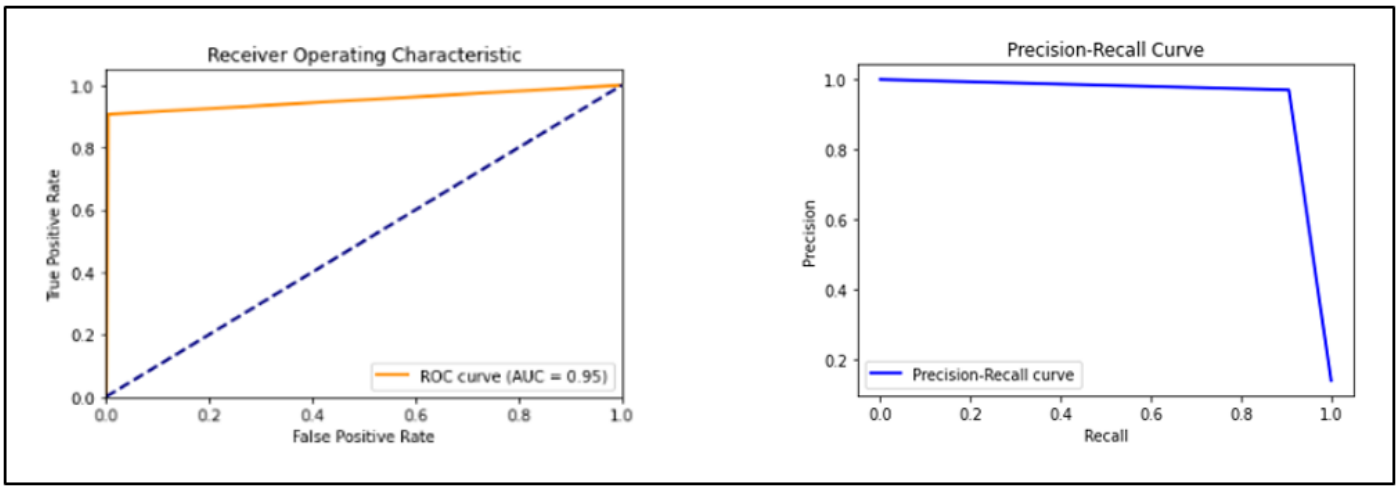

Fig 33 Integrated Model 2 ROC & AUC Curve. Fig 34 Integrated Model 2 Precision-Recall Curve

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

- *Integrated Model 2*

A second integrated model was developed using a combination of additional models to ascertain further the opportunity to improve the system's performance. The second integrated model contains K-mean clustering, XGBoost, Logistic regression, and ANN. The result is presented herein. 31 shows the performance evaluation matrix indicating an increased precision score compared to other models. This is the only improvement made over the integrated model 1. Fig. 32 shows the confusion matrix. Fig. 33 shows the ROC and AUC curves, and Fig. 34 shows the precision-recall curve. The curve shows that the integrated model performed better than the individual models in isolation.

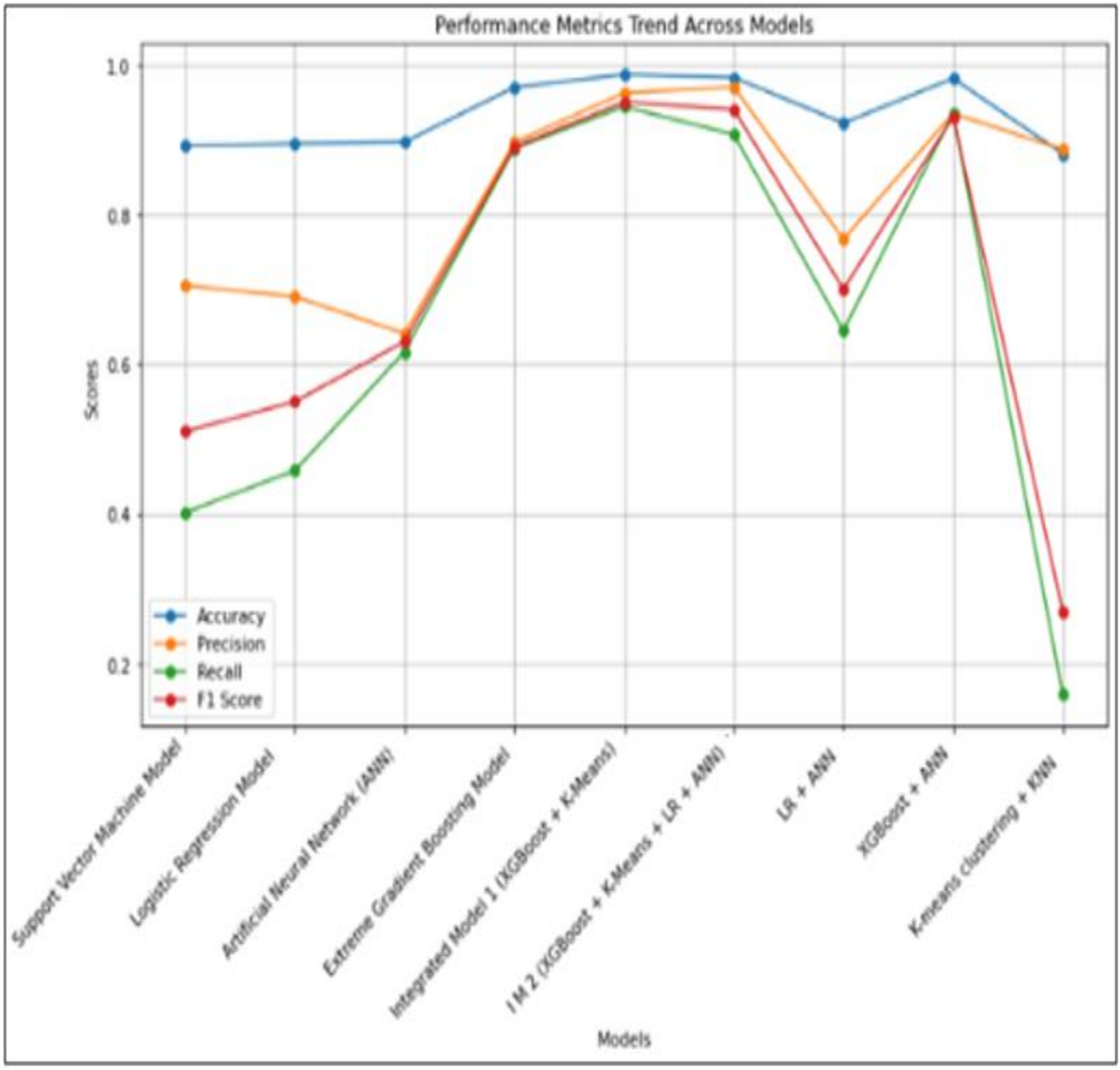

Fig 35 Line Plot Models Comparison

- *Models Comparisons*

The line chart in Fig. 35 illustrates how the performance metrics vary across the nine machine-learning models. Extreme Gradient Boosting and integrated models consistently maintained high scores across metrics, indicating their robust performance. Conversely, the Support Vector Machine and Logistic Regression models showed lower and scores. This visualization enables the identification of trends in model performance, highlighting the effectiveness of specific algorithms and guiding decision-making in selecting the most suitable models for specific tasks.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

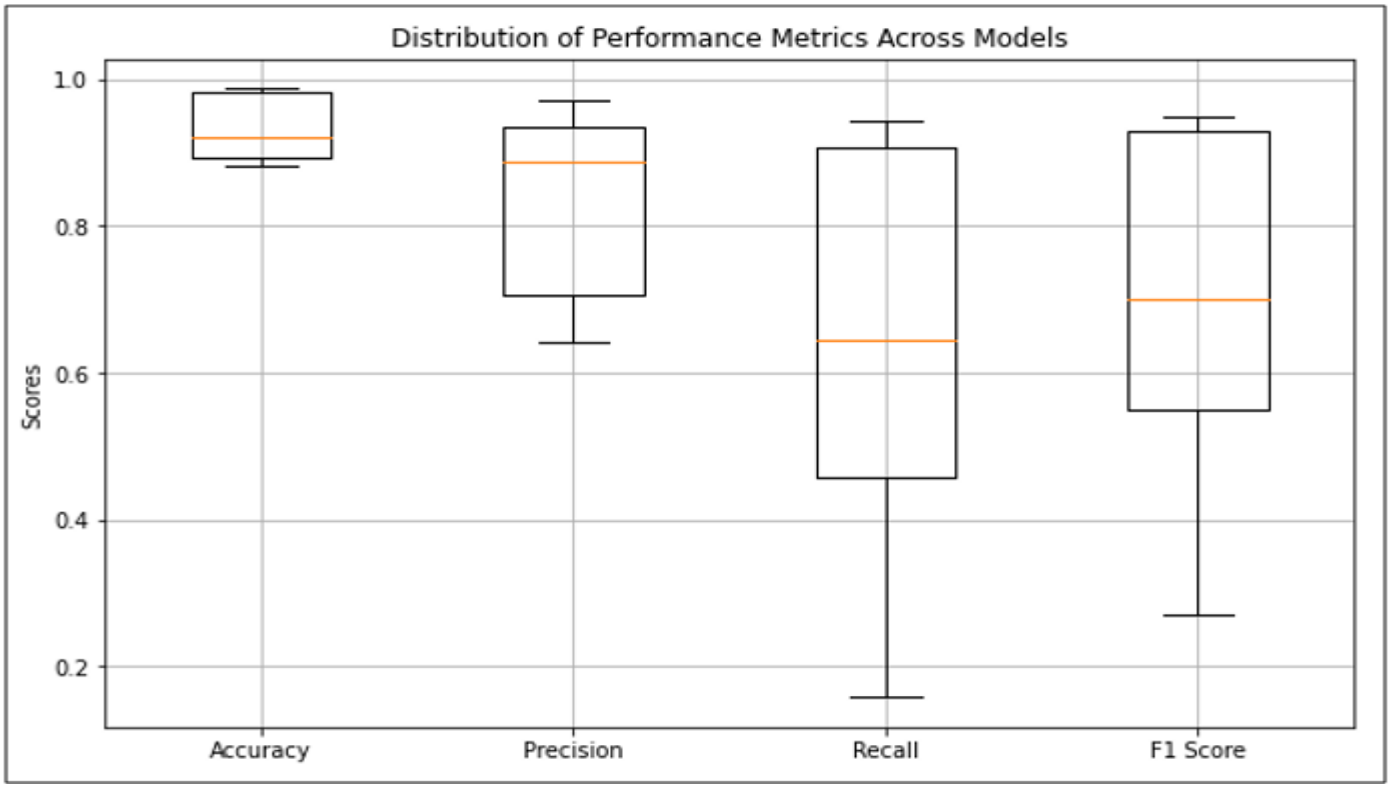

Fig 36 Boxplot Models Comparison.

The box plot in Fig. 36 presents the distribution of performance metrics, accuracy, precision, recall, and F1-score across nine of the nine models used in the study. Extreme Gradient Boosting and integrated models exhibited narrower interquartile ranges and higher medians, indicating a more consistent and superior performance compared to the other models. Conversely, K-means clustering + KNN displayed more comprehensive ranges and lower medians, suggesting more significant variability and inferior performance. This visualization provides a clear overview of performance metrics' spread and central tendency, aiding in model comparison and selection.

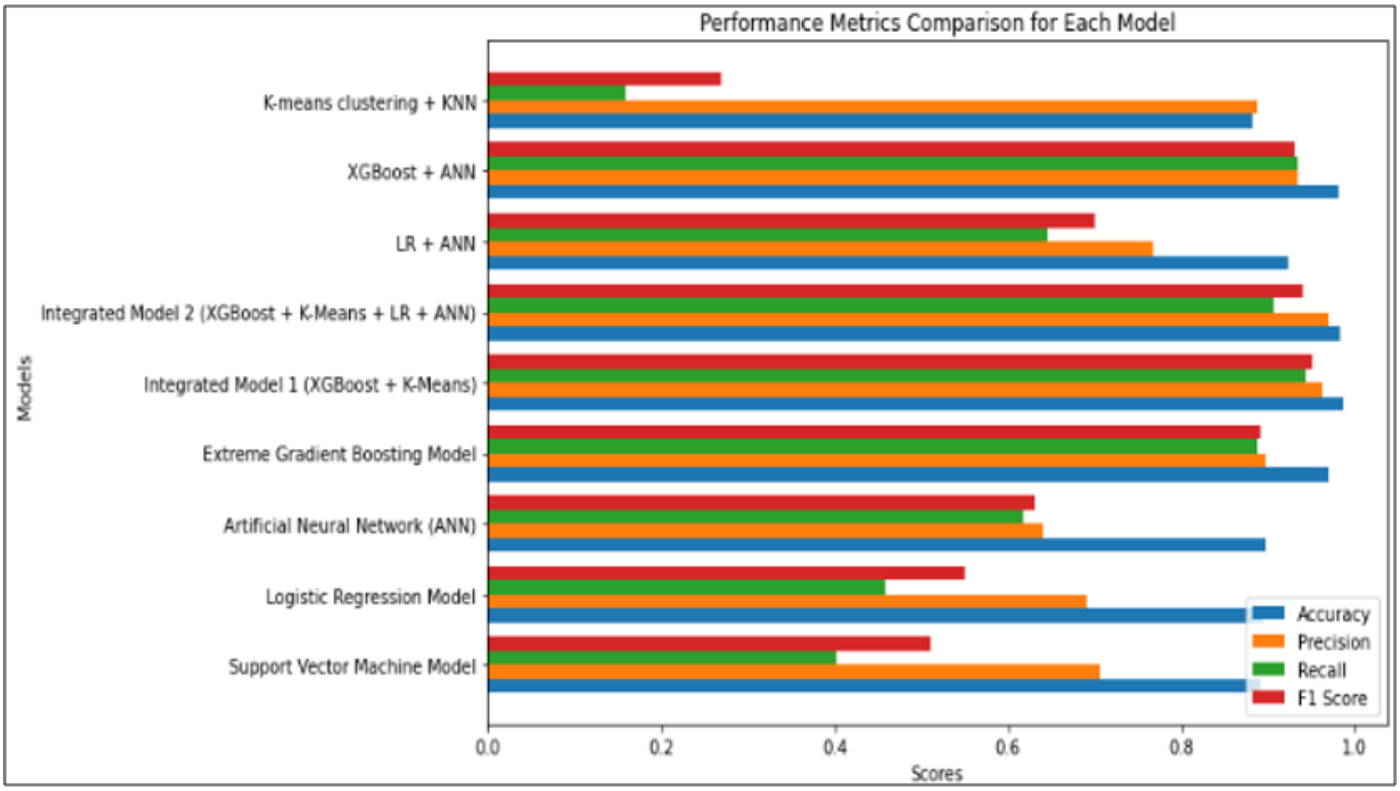

Fig 37 Bar Chart Models Comparison.

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

The horizontal bar chart in Fig. 37 depicts the performance metrics of the models applied in this study. XGBoost performed better for individual algorithms, whereas Integrated Model 1 was superior to XGBoost and other integrated and individual models. The integrated model demonstrates the benefits of combining multiple algorithms. This visualization aids in interpreting and comparing the model's effectiveness for appropriate model selection and decision-making.

Table 1 Model Comparison

MODEL	Accuracy	Precision	Recall	F1-score
Support Vector Machine Model	0.8914	0.7049	0.4019	0.51
Logistic Regression Model	0.8940	0.6901	0.4579	0.55
Artificial Neural Network (ANN)	0.8966	0.6408	0.6168	0.63
Extreme Gradient Boosting Model	0.9695	0.8962	0.8879	0.89
Integrated Model 1 (XGBoost + K-Means)	**0.9868**	0.9619	**0.9439**	**0.95**
Integrated Model 2 (XGBoost + K-Means + LR + ANN)	0.9827	**0.9700**	0.9065	0.94
LR + ANN	0.9219	0.7667	0.6449	0.70
XGBoost + ANN	0.9815	0.9346	0.9346	0.93
K-means clustering + KNN	0.8808	0.8867	0.1587	0.27

Table 1 shows the two integrated models in bold text. The first combines XGBoost and K-means clustering, providing the highest accuracy, recall, and F1 scores. However, the second method, which combined XGBoost, K-means clustering, logistic regression, and ANN, produced a higher precision of 97%. The High accuracy (98.68%) of integrated model 1 implies correct predictions for true negatives and positives. A precision of 97.00% indicates that integrated model 2 has high-quality optimistic predictions. Another essential metric to consider is the recall of 94.39%, which significantly benefits from the implication of missing positive instances; that is, a customer that churns must be predicted. The F1 score of 95% provides a balance between precision and recall; thus, integrated model 1 proved superior in all matrices except precision.

From the results, the best-performing model is Integrated Model 1, a combination of XGBoost and K-means clustering. The two models used in the integrated model were selected based on the accuracy of the supervised learning model. The best-unsupervised learning model was XGBoost, with 96.95% accuracy; however, the performance was significantly improved after applying K-means clustering. Thus, it is evident that the integrated model performs significantly better than individual models. However, there are inevitable trade-offs, which can be seen when comparing integrated models 1 and 2: higher precision was attained in the latter to the detriment of a reduced accuracy score. Several combinations of supervised and unsupervised learning models were used to gain insights into the performance of the integrated models, including LR + ANN, XGBoost + ANN, and K-means clustering + KNN. From the results reported in Table 1, except for the precision of integrated model 2, none of the combinations achieved a higher accuracy, recall, and F1 score in the evaluation matrix than model 1.

➢ *Fig. Assessment of Objectives*

This study successfully applied the CRISP-DM framework for data mining and conducted comprehensive exploratory data analysis (EDA), discovered patterns, and visualized features in the dataset.

A comparative analysis was performed using XGBoost, K-means clustering, logistic regression, SVM, KNN, and ANN in isolation for training, testing, and performance evaluation to guide the selection of a suitable classification algorithm.

An integrated model combining XGBoost for classification and k-means for segmentation was developed. Therefore, the project's goal is to implement an integrated machine-learning model.

The performance of the integrated models was successfully evaluated against individual machine learning algorithms, demonstrating a clear improvement in the prediction accuracy, recall, precision, and F1 score.

The actionable recommendations based on the prediction results align with the project's objective of guiding e-commerce retailers to implement targeted retention strategies, enhance customer satisfaction, and optimize business profitability.

V. CONCLUSION

This study emphasizes the importance of integrated learning methodologies in data science models, specifically combining XGBoost for supervised learning with K-means clustering for unsupervised learning. This has proved highly effective in predicting customer churn in the e-commerce sector. This integrated approach leveraged XGBoost's ability to capture complex data relationships, along with the additional

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

insights provided by the clustering algorithms. This synergistic approach enabled a nuanced understanding of customer churn patterns, empowered businesses to address issues, and proactively optimized retention strategies. To maximize the impact of these findings, organizations are advised to prioritize proactive customer engagement, ensure data quality, explore diverse data sources, continuously evaluate models, and invest in employee training. By implementing these strategies, businesses can enhance their predictive capabilities and translate insights into actionable retention strategies, fostering sustained growth in a dynamic e-commerce environment.

FUTURE WORK

Integrating additional relevant features and exploring new data sources, such as social media interactions and customer feedback, is crucial for enhancing predictive modeling in customer churn analysis. These inputs offer a more comprehensive view of customer behavior and enrich models with contextual information for accurate predictions. Furthermore, reinforcement learning can enhance model adaptability to dynamic customer behavior, allowing continuous refinement based on real-time interactions to strengthen predictive capabilities, effectively mitigate customer churn, and foster sustainable growth in a competitive e-commerce environment.

REFERENCES

[1]. Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in the big data platform. *Journal of Big Data, 6*(1), 1-24.

[2]. Alshamsi, S. A. (2022). Customer churn prediction in the e-commerce sector *Master's thesis, Rochester Institute of Technology, Dubai.* https://core.ac.uk/download/534376391.pdf

[3]. Caigny, D., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research, 269*(2), 760-772.

[4]. Chinnu, P. J., & Paul, P. M. (2017). Customer churn prediction: A survey. *International Journal of Advanced Research in Computer Science, 8*(5), 2178-2181.

[5]. Durkaya, K. B., & Ozcan, T. (2023). Predicting customer churn using grey wolf optimization-based support vector machine with principal component analysis. *Journal of Forecasting*.

[6]. Gonzalez-Rodriguez, N., Osaba, E., Camacho, D., & Yang, X. S. (2019). A hybrid customer churn prediction model uses deep learning and gradient boosting. *Expert Systems with Applications, 137*, 236-247.

[7]. Hu, J., Zhuang, Y., Yang, J., Lei, L., Huang, M., Zhu, R., & Dong, S. (2018). RNN: A recurrent neural network-based approach for customer churn prediction in the telecommunication sector. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 4081-4085). IEEE.

[8]. Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science, 167*, 101-112.

[9]. Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*, 685-695. https://doi.org/10.1007/s12525-021-00475-2

[10]. Khanna, R., & Awad, M. (2015). *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer. https://doi.org/10.1007/978-1-4302-5990-9

[11]. Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2021). Customer churn prediction system: A machine learning approach. VIT Bhopal University, Bhopal, India. https://doi.org/10.1007/s00607-021-00908-y

[12]. Long, J., Chen, Z., He, W., Wu, T., & Ren, J. (2020). An integrated framework of deep learning and knowledge graph for predicting stock price trend: An application in the Chinese stock exchange market. *Applied Soft Computing, 91*, 106205.

[13]. Mittal, M. K. (2022). *Customer churn analysis in telecom using machine learning techniques* (Doctoral dissertation, Dublin, National College of Ireland).

[14]. Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

[15]. Rani, B., & Kant, S. (2020). Semi-supervised learning approach to improve machine learning algorithms for churn analysis in telecommunication. *International Journal of Computer Information Systems and Industrial Management Applications, 12*, 265-275. https://www.mirlabs.net/ijcisim/index.html

[16]. Shirazi, F., & Mohammadi, M. (2019). A big data analytics model for customer churn prediction in the retiree segment. *An International Journal of Information Management, 48*, 238-253.

[17]. Smith, A. N., Fischer, E., & Yongjian, C. (2012). How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *Journal of Interactive Marketing, 32*, 26-36.

[18]. Singh, Y., Pandit, Y., Joshi, N., & Avhad, V. (2022). Prediction of customer churn using machine learning. *International Research Journal of Modernization in Engineering Technology and Science, 4*(4), 2582-5208. www.irjmets.com

[19]. Tran, H., Le, N., & Nguyen, V.-H. (2023). Customer churn prediction in the banking sector using machine learning-based classification models. *Interdisciplinary Journal of Information, Knowledge, and Management, 18*, 87-105. https://doi.org/10.28945/5086

[20]. Umair, S. (2014). A comparative study of data mining process models. *International Journal of Innovation and Scientific Research, 12*(1), 217-222. http://www.ijisr.issr-journals.org/

ISSN No: 2456-2165 https://doi.org/10.38124/ijisrt/25jun1816

[21]. Wu, X., & Meng, S. (2016). E-commerce customer churn prediction based on improved SMOTE and AdaBoost. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)* (pp. 1-5). IEEE.

[22]. 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.

[23]. Yahaya, R., Abisoye, O. A., & Bashir, S. A. (2021). An enhanced bank customers churn prediction model using a hybrid genetic algorithm, k-means filter, and artificial neural network. In *2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA)* (pp. 52-58). IEEE.

[24]. Zhang, G., Zeng, J., Zhao, Z., Jin, D., & Li, Y. (2022). A counterfactual modeling framework for churn prediction. Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China. https://doi.org/10.1145/3488560.3498468