International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Predicting Customer Churn in E-Commerce Using Statistical and Machine Learning Methods

Vidya Rajasekaran[1] , Latha Tamilselvan[2]
[1,2] Department of Information Technology
[1,2] B.S. Abdur Rahman Crescent Institute of Science and Technology
[1,2] Chennai,India
[2] latha.tamil@crescent.education

Abstract— This research work aims to develop prediction models and analytical insights to overcome customer churn issues through data-driven approaches. The attrition rate of consumers in e-commerce is a significant issue requiring effective retention strategies. A novel methodology is proposed comprising data preprocessing, using statistical analysis techniques developing the model and carrying out tailored retention strategies. The model is used to identify crucial churn influencers and propose practical recommendations for enhancing consumer retention. The significance of this work lies in its potential to allow e-commerce ventures with insights with the intention of price savings strategies, enhanced revenue measures, and better consumer fulfillment. This research will influence the e-commerce business by facilitating evidence-based methods for reducing customer turnover and increasing long-term customer value. The resulting accuracy of the proposed model using Logistic Regression results in 87 percentage of accuracy which is a good metric to assess overall model performance. The Kaplan-Meier curve is used to check the survival probability of consumers and identify consumers more likely to churn over time.

Keywords-Consumer churn; Statistical analysis; Kaplan-Meier curve; Consumer retention; Logistic regression; Prediction.

I. INTRODUCTION

Retaining consumers is vital for an online business to continue to be successful in today's vibrant and intensely competitive e-commerce environment. The e-commerce business has grown dramatically over the last decade, and its ongoing expansion has introduced new obstacles. Customers today have access to a plethora of alternatives often just a few clicks away, making it easier than ever for them to switch their loyalty to a new platform or brand. Consequently, to be competitive, e-commerce companies need to be quick to react, flexible, and acutely aware of client preferences. E-commerce businesses continue to struggle with the phenomena of customer churn when customers impede engaging with a brand or platform [1]. It is difficult to fully grasp the intricate mechanics of client attrition. Pricing, product quality, website usability, customer service, and even external variables like economic conditions and industry trends can all influence a consumer's choice to churn. E-commerce businesses must thus embrace cutting-edge analytical and predictive techniques to acquire insights. Understanding the variables that contribute to customer churn and implementing practical ways to reduce it are crucial business imperatives for e-commerce companies as the digital market persists to enlarge and diversify by developing applications [2]. In addition to affecting an e-commerce company's profits, customer churn also influences the revenue of the overall business. Businesses started increasingly relying on statistical and machine learning techniques to predict and decrease customer attrition rates to solve this problem. These methods provide e-commerce businesses the capability to proactively spot consumers in danger of leaving the e-commerce business and apply specific and relevant client retention efforts.

To foresee consumer attrition in the context of e-commerce, this proposed research investigates the relevance of statistical and machine-learning approaches. Businesses may get insights that enable them to take prompt and data-driven steps to retain important consumers by examining the gradations of the customer's behavior, their purchase patterns, and engagement following indicators. A general overview of the problems with client turnover faced by e-commerce firms and the importance of resolving these issues are focused initially. The second section delves into the statistical and machine learning techniques that may be used to forecast customer turnover in e-commerce, providing a methodical overview of the several options open to e-commerce companies. The tactics and approaches that may be used in this model help e-commerce businesses create practical, customer-based strategies to lower churn, enhance customer pleasure, and eventually fuel sustainable business development in an increasingly competitive business [3].

E-commerce enterprises have supreme access to information about their consumers' interactions, tastes, and habits in the age of data-driven decision-making. This profusion of data has the potential to be a potent tool in the fight against client turnover if properly utilized. Organizations may get useful insights from this data and forecast when a client may be about to churn, enabling timely and specialized actions. This is done by using statistical and machine learning methodologies. The objective of this work is to provide e-commerce specialists, data analysts, and business executives with the skills and resources they need to accurately forecast and deal with consumer attrition. E-commerce companies can predict client performance and acclimatize necessary marketing by leveraging the power of statistical and machine learning approaches.

II. LITERATURE SURVEY

The rise of the industrial agglomeration e-commerce business has been strongly influenced by the quick development of technology and the economy, increasing customer shopping experiences and removing logistical obstacles [4]. This development potential is supported by contemporary information technology and networks, with multimedia and industrial cluster marketing increasing relevance. Virtual e-commerce platforms in industrial clusters haven't, however, received much research. This study examines virtual e-commerce models in diverse industrial agglomerations, discussing big data, industrial agglomeration, and their connection to e-commerce. It emphasizes the significant growth in domestic internet users over the past ten years and the significance of e-commerce for current and upcoming company endeavors. The study concludes that various e-commerce models in industrial agglomerations must be adjusted to various contexts since they are context-specific and cannot be generalized.

This work [5] focuses on how crucial client retention is to corporate success, particularly in the cutthroat environment of today. Losing clients, also known as customer churn, is a major problem for newly established businesses. The idea of employing a Stacking Classifier, an ensemble learning approach, to assess and forecast customer turnover in e-commerce data is introduced in the article as a way to address this. This classifier includes knowledge from four basic learners: KNN, SVM, RF Classifiers, and Decision Trees.

The research [6] focuses on dealing with customer churn, which is an important issue for large companies, especially in the telecommunications industry. Using machine learning methods on a big data platform, the authors created a churn prediction model to address this problem. They increased prediction accuracy by utilizing Social Network Analysis (SNA) elements as well.

Customer retention is a crucial issue since the telecommunications business is both oversaturated and very competitive [7]. Data mining techniques and data science technology provide useful tools for anticipating client turnover and enhancing customer loyalty. The study intends to analyze alternative methods and build data science models to categorize consumers based on their propensity to quit a telecom firm. This study shows how these models may forecast and lower customer turnover by identifying the major causes of churn and improving services. Enhancing customer loyalty, lowering attrition, and improving company results may all be achieved by implementing customer churn prediction models.

Machine learning techniques for anticipating client attrition were compared and analyzed by the authors in [8]. The research assesses the effectiveness of several machine learning algorithms in the context of customer attrition prediction.

The authors [9] explore the world of e-commerce, paying close attention to client turnover prediction. The research results presented in the article can give e-commerce enterprises insightful information that will help them better understand consumer behavior and foresee turnover. The creation of more successful customer engagement programs, retention tactics, and marketing campaigns that are specifically tailored to each customer can result from this.

The authors in [10] combine machine learning and geographical analytic methods, which proposes a thorough approach for researching and forecasting consumer attrition in the e-commerce industry. Spatial analysis is commonly used to evaluate geographical or location-based data and patterns, and it may be used to identify spatial trends or dependencies linked to customer turnover in this context. Contrarily, machine learning techniques are useful for processing and analyzing enormous volumes of data, making it possible to spot patterns, trends, and prediction models.

E-commerce enterprises must be able to forecast client turnover to improve customer retention and marketing tactics [11]. The model presented in this research uses support vector machine (SVM) prediction and k-means customer segmentation to detect customer loss in B2C e-commerce. Customers are divided into three groups using the approach, which also identifies important customer segments. The work shows that k-means clustering enhances prediction accuracy when comparing SVM and logistic regression for churn prediction. The accuracy of SVM is higher than that of logistic regression. The management of customer interactions can benefit from these insights for B2C e-commerce businesses.

This study explores several data analysis strategies, techniques, and tools that may be used to examine consumer behavior data and find trends or patterns that could point to impending churn [12]. These techniques could consist of data mining, machine learning techniques, and statistical analysis. The study's goal is to investigate how these methods might be used in the particular context of cargo and logistics, where keeping clients is crucial to preserving the profitability and long-term viability of the company.

E-commerce businesses gather a lot of client information, such as search history, buying trends, reviews, and comments. This data may be used using machine learning and data mining tools to assess customer behavior and detect possible attrition issues [13]. A popular supervised learning technique in machine learning, the support vector machine focuses on both regression and classification issues in predictive analysis. The approach for forecasting E-Commerce customer attrition presented in this research uses support vector machines in conjunction with a hybrid recommendation strategy. Empirical data show that employing the integrated forecasting model significantly improves several parameters, including coverage ratio, hit ratio, lift degree, and accuracy rate.

Considering all the relevant research mechanisms a novel mechanism is proposed to predict consumer churn in e-commerce business platforms.

III. PROPOSED METHODOLOGY

The initial process to start with is data preprocessing. The efficiency of the prediction techniques relies on the quality of the data because raw data causes inaccuracies and inconsistencies and impacts the accuracy of the results. The dataset is preprocessed by handling missing values through removal, identifying and removing duplicates and outliers, finally, the data consistency is ensured. The target variable churn is set by finding the consumers who haven't made successful purchases in the last six months. Feature engineering is carried out to create new features and transform the existing features to enhance the performance of the model. The features involved in predicting the churn value are identified. The features include frequency, purchase history, monetary value, and consumer demographic information like age, gender, location, and salary. Additional features like time-stamped

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

data, successful transactions, average order value, and return information are also considered.

The dataset is split in a ratio of 70:15:15. 70% of the data is used in training the model, 15% is used in validation, and the remaining 15% is used for testing the model. The dataset is balanced using techniques like oversampling and under sampling. Scaling is carried out to transform the data for good model convergence.

Exploratory Data Analysis is conducted to gain insights from the data and to identify and understand the significant patterns, the structures, and their relationships. The basic statistics are displayed for the dataset having numeric features using the Python panda library to describe the function and the results are stated in Fig.1.

	cust_id	sex	Age	city_code	total_tran
count	5502.000000	5502.000000	5502.000000	5502.000000	5502.000000
mean	271035.526536	1.486550	32.808615	5.482734	3.791349
std	2453.093450	0.499865	6.613921	2.858343	1.835860
min	266783.000000	1.000000	21.000000	1.000000	1.000000
25%	268911.000000	1.000000	27.000000	3.000000	2.000000
50%	271026.500000	1.000000	33.000000	5.000000	4.000000
75%	273172.750000	2.000000	39.000000	8.000000	5.000000
max	275265.000000	2.000000	44.000000	10.000000	11.000000

	return_tran	success_tran	Qty_purchase	total_amt	frequency
count	5502.000000	5502.000000	5502.000000	5502.000000	5502.000000
mean	0.395493	3.395856	10.184478	8823.023930	2.395856
std	0.657679	1.767623	5.932385	5874.936059	1.767623
min	0.000000	-2.000000	-8.000000	-9352.720000	-3.000000
25%	0.000000	2.000000	6.000000	4320.550000	1.000000
50%	0.000000	3.000000	9.000000	7898.540000	2.000000
75%	1.000000	4.000000	14.000000	12423.238750	3.000000
max	4.000000	11.000000	37.000000	41510.430000	10.000000

	recency	monetary_value	return	churn
count	5502.000000	5502.000000	5502.000000	5502.000000
mean	589.624682	2157.944500	0.313886	0.294984
std	333.147681	1352.431316	0.464113	0.360205
min	0.000000	0.000000	0.000000	0.000000
25%	348.000000	1231.083333	0.000000	0.000000
50%	633.500000	2091.523809	0.000000	0.000000
75%	846.000000	2952.937500	1.000000	0.500000
1 max	1420.000000	7500.000000	1.000000	1.000000

Figure 1. Statistical Representation of the Operational Dataset.

Univariate analysis for the data distributions is represented in Fig.2 . The histogram plot represents the total purchase amount and the distribution of purchase amounts.

A counterplot as shown in Fig. 3 is used to visualize the distribution of categorical variables of the customer segments. It displays the number of customers in each segment.

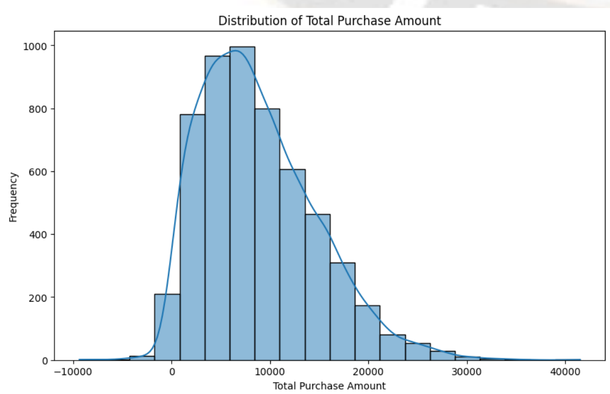

Figure 2. Histogram Representation of Total purchase amount and frequency of purchase.

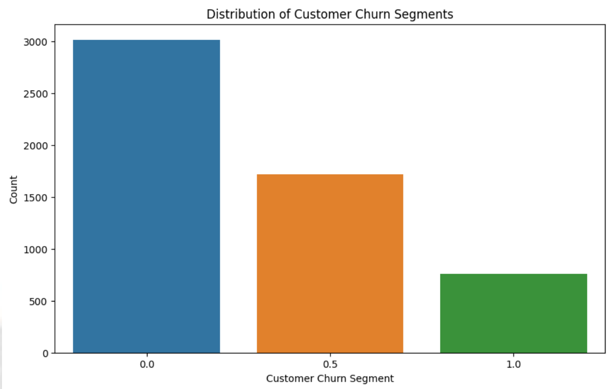

Figure 3. Countplot representing the distribution of customer segments.

The Bi-variate analysis is a statistical analysis technique used to examine the relationship between two variables and their relationships. A heat map is generated to calculate and visualize the correlation matrix and its influences on the churn. Fig. 4 represents the correlation matrix and Fig. 5 denotes the churn analysis. The churn analysis provides insights into the proportion of customers churned.

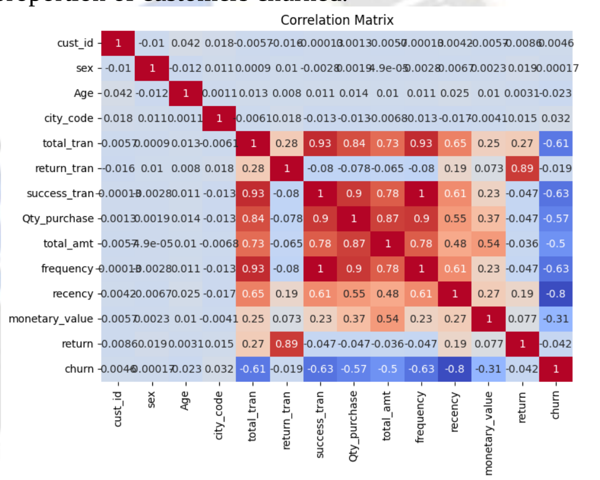

Figure 4. Relationship analysis using Correlation Matrix.

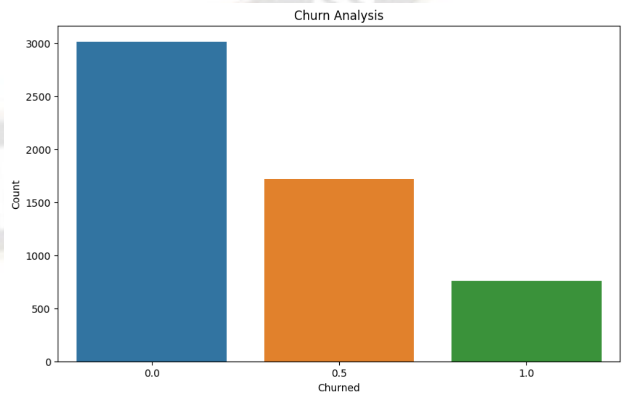

Figure 5. Countplot representing distribution of customer segments.

Fig.6 denotes demographic analysis representing the age and Fig.7 represents the behavioral analysis through frequency distribution based on the customer churn analysis.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

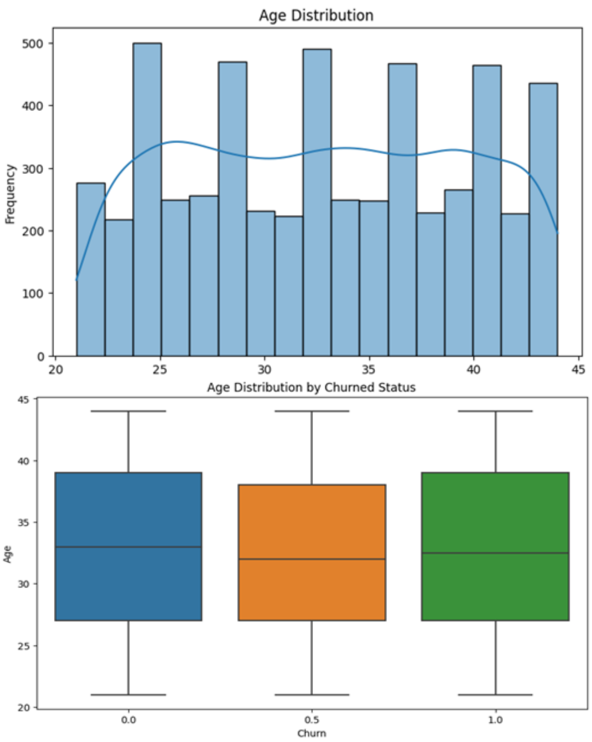

Figure 6. Age Distribution graph based on frequency and customer churn.

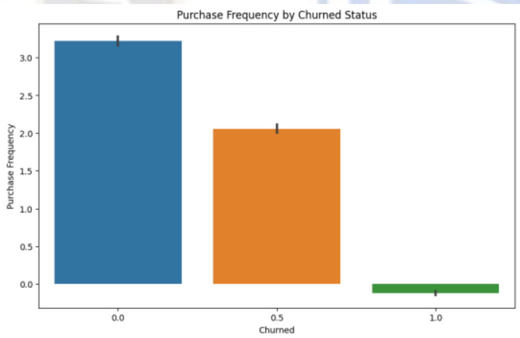

Figure 7. Consumer Purchase Frequency Graph based on Churn analysis.

The time series analysis is depicted in Fig. 8. The timestamp is represented in date time format and is set as an index. It helps in calculating the last purchase and purchase frequency of the consumers.

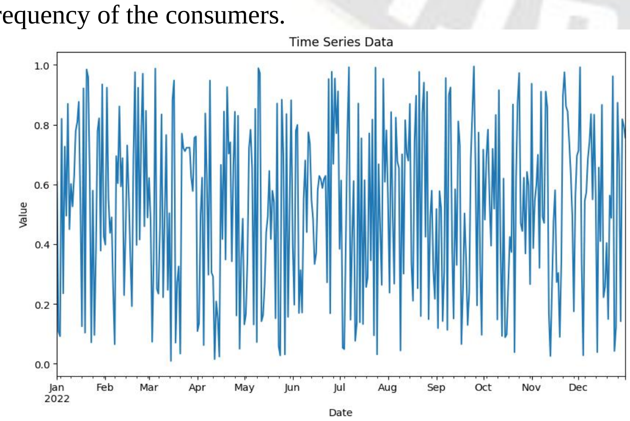

Figure 8. Time series analysis representation graph.

A Statistical test is carried out to compare the characteristics and behaviors of churned and non-churned consumers. The significant differences can be calculated using either the t-test, chi-square test, or ANOVA test. ANOVA (Analysis of Variance) is applied to analyze the variation between different groups and categories to explore the statistically significant differences among them. The obtained values are F-Statistic: 7.2470099773916745 and P-Value: 0.007123421087408745. The resulting interpretation states that the analysis has found statistically significant evidence that the mean total purchase amounts differ among the consumers churned and not. The obtained p-value is less than 0.05 and it indicates that observed differences are statistically significant. This insight is used for decision-making and developing customer retention strategies.

A model is developed using Logistic Regression to predict customer churn. The model estimates the probability of the customer churn based on the features. Logistic Regression algorithm is used in several prediction applications [14] [15] [16] [17]. Logistic Regression is applied to binary classification problems and here the binary classification task of categorizing consumers as churned or not based on the features . The model is used to estimate the probability of customer churn. The Sigmoid function is applied to transform the linear combination of features between a probability value of 0 and 1. The equation of the Logistic function is stated below in (1).

$$P(X=1) = \frac{1}{1+e^{-(a_0+a_1y_1+a_2y_2+\cdots+a_ny_n)}} \quad (1)$$

Here, P(X=1) is the probability of customer churn, y_1, y_2, y_3, and y_n are the features, and a_0, a_1, a_2 and a_n, denote the logistic model coefficient values.

After the proposed model is trained and evaluated it is applied to make predictions on the unseen data. The probability of customer churn is calculated for the given consumer based on their feature values. Finally, a decision threshold is set to segregate the consumers as either churned "1" or non-churned "0". We set a threshold score of 0.5 stating that if the predicted probability is greater than 0.5 the consumers are classified as churned and less than 0.5 as non-churned. If the value is exactly 0.5, they are kept in further investigation analysis for further upcoming months.

The strategies to enhance customer retention can be practiced based on the derived predictions. A conclusion can be made for consumers with a high probability of churning by targeting marketing strategies like discounts and promotions.

Further, the consumer survival probabilities and analysis can be implemented through Kaplan-Meier analysis techniques. The technique is used in survival analysis estimation and visualization of consumers over a certain period. The resulting curves are used in analyzing the time at which the consumer churn occurs. It is used to calculate the probability of a consumer continuing with an e-commerce platform over a particular time and find the probability of a customer staying over different periods of months. Time–to–event data is the information on the customer churn and the duration the customer remains active on an e-commerce

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

platform before churn. All the consumers will not experience the churn factor in the stated time and some will be still active and not churned at the end of time. This Kaplan-Meier curve is used to handle this type of issue and make predictions for consumers who have not churned.

The formula for the Kaplan-Meier curve estimator for a specific time is represented in the below equation (2). In the below equation, $\widehat{K}(t)$ is the estimated survival probability over a time t. t_x denotes the customer churn time over some time, C_x is the customer churn observed over a certain time and R_x is non-churned consumers over time x.

$$\widehat{K}(t) = \prod_{t_x \le t} \left(1 - \frac{C_x}{R_x}\right) \tag{2}$$

The time series analysis is further extended to find the seasonality changes in consumer behavior and their relationship with consumer churn. Hence a final decision support system is generated to predict he churn and provide relevant insights to enhance consumer retention in e-commerce business. The results of statistical analysis and machine learning algorithms can be compared to derive a holistic view of consumer churn. The Kaplan-Meier curve is examined to check how it differs in terms of survival probability and identify which consumers are likely to churn over time. The results of logistic regression are used to determine the predictor's statistical significance for predicting customer churn. The resulting coefficient values indicate the direction and the strength of the model. The features creating a significant impact on the churn are identified and used in developing targeted retention strategies. Hence a complete and interpretable analysis of consumer churn can be developed for e-commerce businesses to understand the survival mechanism and retention dynamics of consumers.

IV. Results and Discussion

The performance metric of logistic regression is evaluated using various metrics like accuracy, precision, recall, F1-score, and support score [18] [19] [20]. The results are depicted in Fig. 9. The below-calculated metrics are used to determine how well the proposed model can distinguish between churned and retained customers.

```
Accuracy: 0.8782924613987284
Confusion Matrix:
 [[533  56]
 [ 78 434]]
Classification Report:
              precision    recall  f1-score   support

       False       0.87      0.90      0.89       589
        True       0.89      0.85      0.87       512

    accuracy                           0.88      1101
   macro avg       0.88      0.88      0.88      1101
weighted avg       0.88      0.88      0.88      1101
```

Figure 9. Performance analysis results of Logistic Regression model.

Accuracy is a commonly used performance evaluation metric for evaluating classification models used in measuring the proportion of the instances correctly classified from the overall instances. The calculated accuracy measure is 0.8783, which states that the model correctly predicts 87.83 % of the instances in the dataset using the formula in (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

A confusion matrix table is generated to evaluate the performance of the model. There are four values required to calculate the confusion matrix score. The TP (True Positives) are churned consumers that are correctly predicted by the proposed model. The TN (True Negatives) are non-churned consumers that are also correctly predicted as negatives by the proposed model. The FP (False Positives) is negatives which are predicted as positives by the model and FN (False Negatives) is instances that are positive in reality but are predicted as negatives. The result states that there are 533 true positives correctly predicted as churned customers, 434 true negatives correctly predicted as not churned, 56 as false positives that are incorrectly predicted as churned consumers, and 78 as false negatives are churned consumers incorrectly predicted as not churned as shown in Fig.9.

The precision, recall, and F1-score are calculated for both churned (True) and non-churned consumers (False) using formulas in (4) (5), and (6).

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{6}$$

For true the precision score is 89% which indicates that the instances predicted as true are certainly true and recall 85% indicates true instances predicted correctly. The F1-score is 87% which is a balance between precision and recall. The non-churned consumers stated as false resulted in a precision score of 87% indicating instances predicted as false are false, 90% of the recall score stating actual false correctly predicted, and 89% of the F1-score stating balance between the two metrics. The support metric indicates the number of instances in each class. The proposed model performs well with good overall accuracy providing detailed insights into the model's performance for both the churned and non-churned consumers and compared to different models [21].

Then, statistical analysis for consumer survival is estimated through Kaplan-Meier curves shown in Fig.10.

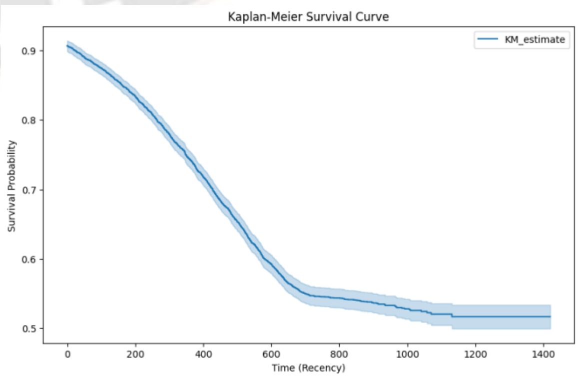

Figure 10. Kaplan-Meier curves denoting consumer churn probability.

The x-axis of the above graph represents time intervals in weeks and the y-axis denotes the survival probability. Initially, the curve starts at 1 representing consumers are active, and slowly decreases over time when the consumers are churned. Further separate curves can also be generated to identify consumer churn and tailor specific retention mechanisms.

V. CONCLUSION

This proposed work scrutinizes the techniques and data-driven strategies used to predict client attrition in the e-commerce region, offering insights into customer retention tactics. Predictive models and approaches must continually be enhanced due to the business's quick development and changing consumer preferences. A fascinating strategy is exposed by combining statistics and machine learning techniques. The trade-offs between accuracy and interpretability is highlighted by the comparative examination of the developed model. Understanding customer turnover depends greatly on the model's interpretability and explain ability. E-commerce businesses can make use of interpretable models to obtain perceptive information and make data-driven decisions to facilitate and efficiently retain consumers. The proposed research recommends churn prevention techniques as well as individualized marketing, targeted incentives, and improved customer service to increase customer loyalty and profitability. Future studies may examine real-time data analytics, sophisticated deep learning methods, and models that adjust to shifting consumer behavior as the e-commerce industry changes.

REFERENCES

[1] Li, L. & Zhang, J. (2021). Research and Analysis of an Enterprise E-Commerce Marketing System Under the Big Data Environment. *Journal of Organizational and End User Computing (JOEUC), 33*(6), 1-19. http://doi.org/10.4018/JOEUC.20211101.oa15

[2] JWenli, Z., & Liang, H. (2023). Research on optimization design of fresh e-Commerce APP based on KANO model. *International Conference on Cyber Security, Artificial Intelligence, and Digital Economy (CSAIDE 2023)*. https://doi.org/10.1117/12.2681544

[3] Šaković Jovanović J, Vujadinović R, Mitreva E, Fragassa C, Vujović A. The Relationship between E-Commerce and Firm Performance: The Mediating Role of Internet Sales Channels. *Sustainability*. 2020; 12(17):6993. https://doi.org/10.3390/su12176993

[4] Yue Hongqiang,Research on E-Commerce Data Standard System in the Era of Digital Economy From the Perspective of Organizational Psychology,Frontiers in Psychology,VOLUME=13,YEAR=2022,DOI=10.3389/fpsyg.2022.900698, ISSN=1664-1078

[5] Awasthi, S. (2022). Customer churn prediction on e-Commerce data using stacking classifier. https://doi.org/10.36227/techrxiv.20291694

[6] Demir, B., & Ergün, Ö. Ö. (2023). Customer churn prediction with machine learning methods in telecommunication industry. https://doi.org/10.21203/rs.3.rs-3343217/v1

[7] Gopal, P., & MohdNawi, N. B. (2021). A survey on customer churn prediction using machine learning and data mining techniques in e-Commerce. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. https://doi.org/10.1109/csde53843.2021.9718460

[8] Karamollaoglu, H., Yucedag, I., & Dogru, I. A. (2021). Customer churn prediction using machine learning methods: A comparative analysis. *2021 6th International Conference on Computer Science andEngineering(UBMK)*. https://doi.org/10.1109/ubmk52708.2021.9558876

[9] Lubis, A. R., Prayudani, S., Julham, Nugroho, O., Lase, Y. Y., & Lubis, M. (2022). Comparison of model in predicting customer churn based on users' habits on e-Commerce. *2022 5th International Seminar on Research of Information Technology and IntelligentSystems(ISRITI)*. https://doi.org/10.1109/isriti56927.2022.10052834

[10] 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. https://doi.org/10.3390/jtaer17010009

[11] Xiahou X, Harada Y. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*. 2022; 17(2):458-475. https://doi.org/10.3390/jtaer17020024

[12] Sahinkaya, G. O., Erek, D., Yaman, H., & Aktas, M. S. (2021). On the data analysis workflow for predicting customer churn behavior in cargo and logistics sectors: Case study. *2021 International Conference on Electrical, Communication, and Computer Engineering(ICECCE)*. https://doi.org/10.1109/icecce52056.2021.9514117

[13] Shobana J, Ch. Gangadhar, Rakesh Kumar Arora, P.N. Renjith, J. Bamini, Yugendra devidas Chincholkar,E-commerce customer churn prevention using machine learning-based business intelligence strategy, Measurement: Sensors,Volume 27,2023,100728,ISSN 2665-9174,https://doi.org/10.1016/j.measen.2023.100728.

[14] Baxani, R., & Edinburgh, M. (2022). Heart disease prediction using machine learning algorithms logistic regression, support vector machine and random forest classification techniques. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4151423

[15] Edlitz, Y., & Segal, E. (2022). Author response: Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards. https://doi.org/10.7554/elife.71862.sa2

[16] Harshini, P. S., Naresh, K., Pamulapati, S. R., & Lavanya, A. (2023). Diagnosis of liver diseases using machine learning algorithms and their prediction using logistic regression and ANN. *2023 3rd International Conference on Intelligent Technologies (CONIT)*. https://doi.org/10.1109/conit59222.2023.10205819

[17] Kadam, E., Gupta, A., Jagtap, S., Dubey, I., & Tawde, G. (2023). Loan approval prediction system using logistic regression and CIBIL score. *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. https://doi.org/10.1109/icesc57686.2023.10193150

[18] Manglani, R., & Bokhare, A. (2021). Logistic regression model for loan prediction: A machine learning approach. *2021 Emerging Trends in Industry 4.0 (ETI 4.0)*. https://doi.org/10.1109/eti4.051663.2021.9619201

[19] Rajasekaran, V., & Priyadarshini, R. (2021). An e-Commerce prototype for predicting the product return phenomenon using optimization and regression techniques. *Communications in Computer and Information Science*, 230-240. https://doi.org/10.1007/978-3-030-88244-0_22

[20] Reddy, D. J., Gunasekaran, M., & Sundari, K. S. (2022). undefined. *2022 International Conference on Cyber Resilience (ICCR)*. https://doi.org/10.1109/iccr56254.2022.9995969

[21] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simul. Model. Pract. Theory, vol. 55, no. 10, pp. 1–9, 2015, doi: 10.1016/j.simpat.2015.03.003.

[22] Retail Case Study Analysis | Kaggle