iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50479

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Customer Churn Prediction on E-Commerce Using Machine Learning

Rohit Kumar Jaiswal[1], Amit Kori[2], Rohit Inkar[3], Chetan Adari[4], Samiksha Bansode[5]
[1, 2, 3, 4]*UG Scholar,* [5]*Assistant professor Computer Engineering University of Mumbai, Alamuri Ratnamala Institute of Engineering & Technology, India*

Abstract: ***For E-commerce businesses to produce successful marketing plans and customer retention tactics, client churn vaticination is pivotal. In order to handle the longitudinal timeframes and multiple data variables of B2Ce-commerce consumers' buying habits, the authors of this study present a loss vaticination model that integrates k- means client segmentation with support vector machine (SVM) vaticination. guests are divided into three groups according to the approach, which also defines the main customer groupings. In order to anticipate client development, the study analyses the efficacity of logistic retrogression and SVM vaticination. The findings show that client segmentation greatly increases each indicator's capability to read values, emphasizing the significance of k- means clustering segmentation. also, it's demonstrated that SVM vaticination is more accurate than logistic retrogression vaticination. The conclusions of this study have important ramifications for client relationship operation.***

Keywords: Churn prediction Machine learning techniques Boosting algorithm.

I. INTRODUCTION

Customers are a valuable asset for any business as they play a vital role in enhancing market competitiveness and performance. In today's fiercely competitive market, customers have a plethora of products and service providers to choose from. Research shows that the cost of acquiring a new customer is often higher than retaining an existing one. By maintaining a strong and long-lasting relationship with customers, a business can derive more profits from its existing customers. A mere 5% increase in customer retention can lead to a 25-95% increase in the net present value of the business. Similarly, reducing the customer churn rate by 5% can result in a 25-85% increase in the average profit margin of the enterprise. Therefore, it has become crucial for businesses to leverage their existing customer resources and prevent customer loss to maintain their market advantage. One effective approach to achieving this is through customer churn prediction techniques that can help identify customers who are at risk of leaving, enabling the business to take proactive measures to retain them. This is particularly important in the highly competitive telecommunications market, where companies must analyse customer behaviour to identify churn risks and take appropriate steps to retain customers. This involves examining customers' calling behaviour, their interactions with the operator, package subscriptions, account information, calling details, and demographic characteristics. In e-commerce, the significance of churn prediction and analysis lies in its ability to help companies anticipate and identify clients who may be at risk of leaving, allowing them to take necessary measures to reduce or prevent customer churn and minimize potential losses.

II. LITERATURE REVIEW

Forecasting customer churn can be classified into three categories: traditional statistical analysis, machine learning, and combinatorial classifiers. Traditional statistical methods like logistic regression and linear discriminant analysis are interpretable but may not perform well with large and complex data.

Machine learning methods, including support vector machines, decision trees, and artificial neural networks, have shown promising results in predicting customer churn across various industries. Combinatorial classifiers such as XGBoost and AdaBoost combine several weak classifiers to form a strong classifier and have been used to predict customer churn in datasets with time series characteristics.

While past studies have made valuable contributions to churn prediction of contractual customers in industries such as telecom, banking, and B2B e-commerce, predicting customer loss in B2C e-commerce requires personalized approaches as it is a multidimensional problem. Thus, this study will focus on predicting the loss of non-contractual customers in B2C e-commerce enterprises by analysing customer data characteristics.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

III. PROPOSED SYSTEM

We used Dash and machine learning models to create a single-page web application for customer churn analysis. We conducted exploratory data analysis to identify missing values, categorical and numerical variables, and columns that have a high impact on customer churn in recent years. Our dataset includes 5630 unique customer IDs, and all columns with n=5630 have no missing values. We then split the data into a 90% training dataset and a 10% test dataset. We trained four base learners - Decision Trees, Random Forests, Support Vector Machines, and KNN classifiers. These models' outputs were fed into the meta-classifier of the Stacking Classifier, which used logistic regression.

We compared the prediction performance of LR and SVM using three commonly used performance indicators - Accuracy, Recall, and Precision. However, we believe that customer data in e-commerce enterprises is unique and requires personalized approaches. These enterprises often update product information or upload various evaluation information for customer retention, which is different from financial and telecommunication customer information. Thus, we also considered the operational efficiency of the models, especially when predicting customers in real-time. We found that SVM's data training time is significantly shorter than that of LR when training customer data.

We used four base learners in our analysis. Decision Trees are unsupervised machine learning algorithms that can be used for classification or regression. Logistic regression models the probability of a binary outcome and is commonly used in classification problems. Random Forest is a machine learning algorithm that combines the outputs of multiple decision trees to overcome overfitting and bias issues. Finally, Support Vector Machines are supervised machine learning algorithms that are commonly used in classification problems and have been shown to have good performance in customer churn analysis.

IV. SYSTEM ARCHITECHTURE

The main aim of this research was to distinguish between customers who churned and those who stayed, and to determine the factors that contribute to churn. The findings of this study indicate that single male customers are slightly more likely to churn. In addition, customers who prefer the Mobile order category were found to be more prone to churn. Moreover, churned customers showed a slightly higher preference for using a phone or mobile device to log in, which could be due to the customer experience provided by the E-commerce platform's phone version. Additionally, the study identified that churned customers have a higher mean for complaints, city tier, number of addresses, and number of registered devices. However, surprisingly, churned customers had a higher satisfaction score compared to the retained customers. On the other hand, the tenure and the count of the number of orders were found to be lower for churned customers, which is expected.

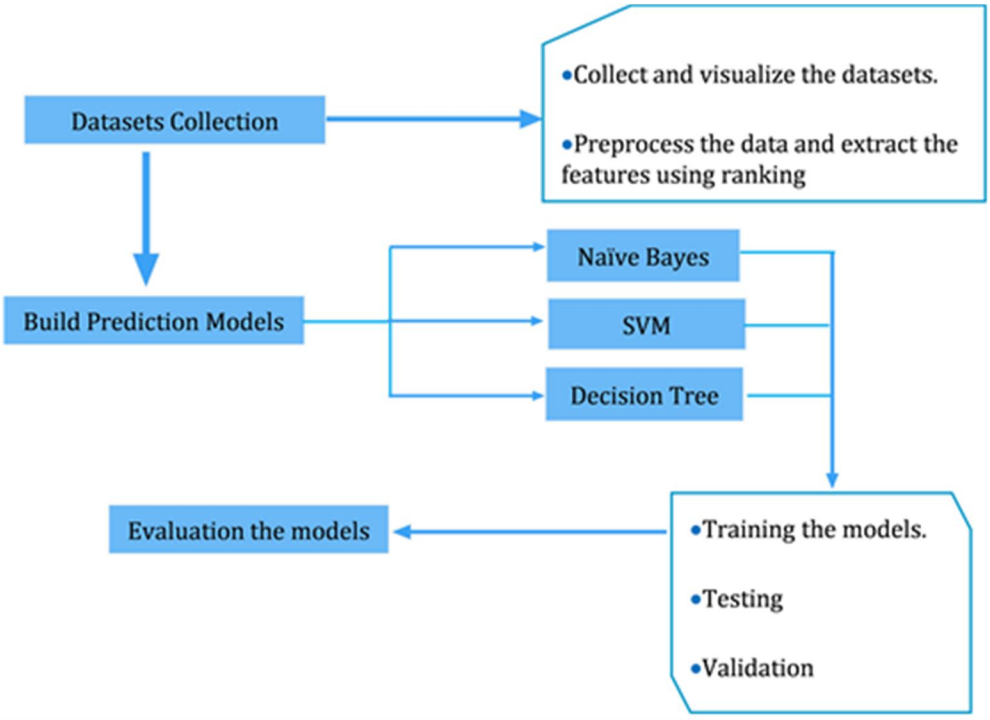


Fig -1: System Architecture

V. DATA FLOW DIAGRAM

The system diagram depicts an E-commerce platform as a rectangular shape, which includes various components such as Customer Data, Data Pre-processing, Model Building, Trained Model, and Churn Prediction. The input data for the system is represented by the Customer Data component. The Data Pre-processing component includes four sub-components, which are Data Cleaning, Data Integration, Feature Engineering, and Feature Selection. The Model Building component includes two sub-components, which are Algorithm Selection and Hyperparameter Tuning. The Trained Model component represents the output of the Model Building process. Finally, the Churn Prediction component uses the Trained Model to make predictions on the input data.

VI. REQUIREMENT ANALYSIS

A. *Hardware Requirements*

For Development we need a machine of following configuration:

1) CPU: Core i5 10th Gen, 1.2GHz.
2) RAM: DDR3 4GB.
3) HDD: 256 GB.
4) Systems: Monitor, Keyboard, Mouse.

B. *Software Requirements*

1) Operating System: Windows 8/10/11.
2) Programming Language: Python, JSON.
3) Development IDE: Visual Studio Code Version: 1.75
4) Other Software's: Google Collab, Jupiter Notebook.

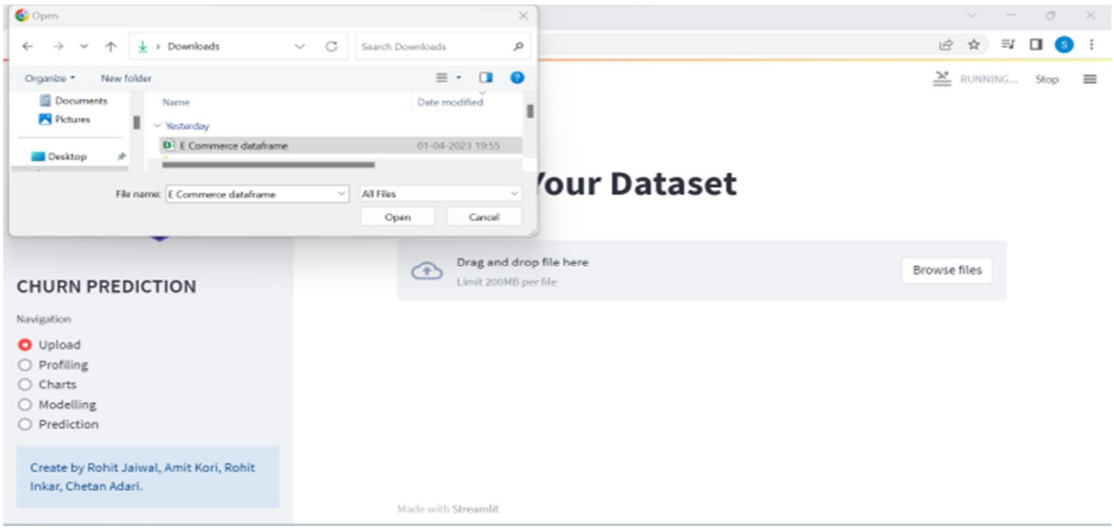
VII. RESEARCH AND METHODOLOGY

1) Google Collab is a free cloud based Jupyter notebook environment that is capable of running many popular machine learning libraries, which can be easily imported into the notebook for use.
2) Python is a programming language that has a simple and clear programming style and offers powerful features through various classes. It is also capable of easily integrating with other programming languages like C or C++.
3) NumPy is an open-source library used for analysing and calculating data in Python and is essential for implementing the array data type in Python. It is mainly used for matrix calculations.
4) Pandas and Matplotlib are Python libraries that are freely available and commonly used for analysing and visualizing data.
5) Their main goal is to provide users with efficient tools to perform quick iterations of data analysis, visualization, and debugging. However, more complex workflows may require more advanced integrated development environments (IDEs), such as Visual Studio IDE.

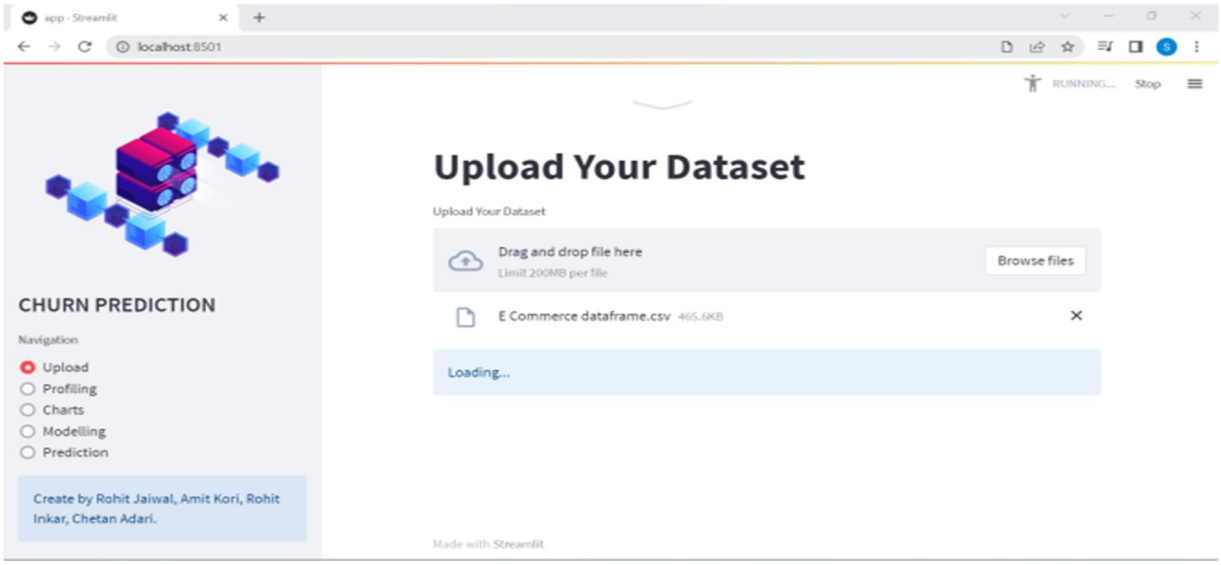
VIII. IMPLEMENTATION AND RESULTS

1) *Step 1:* To analyse and manipulate data, we will utilize two open-source libraries in Python - Pandas and NumPy. For data visualization, we will employ two libraries - Matplotlib and Seaborn. These libraries provide essential tools for data analysis, manipulation, and visualization.

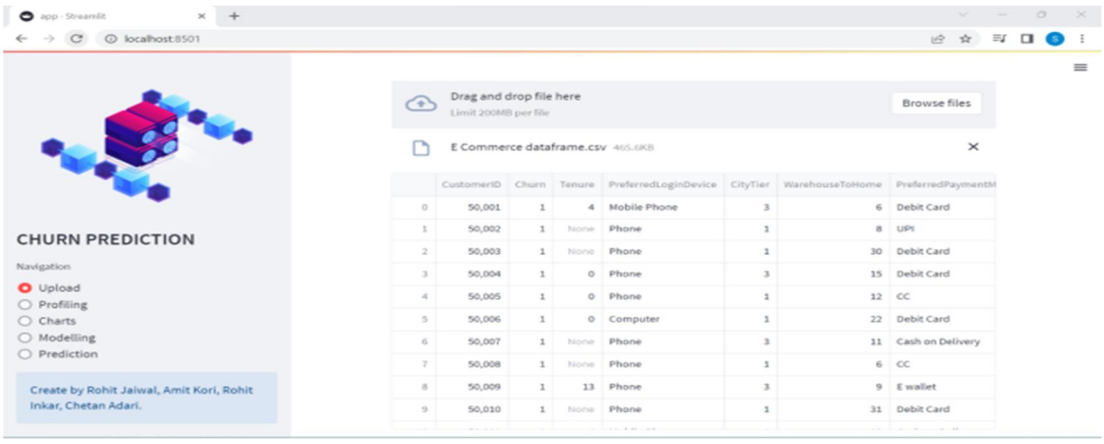
2) *Step 2:* Having imported the required libraries, we can now proceed to reading the data.



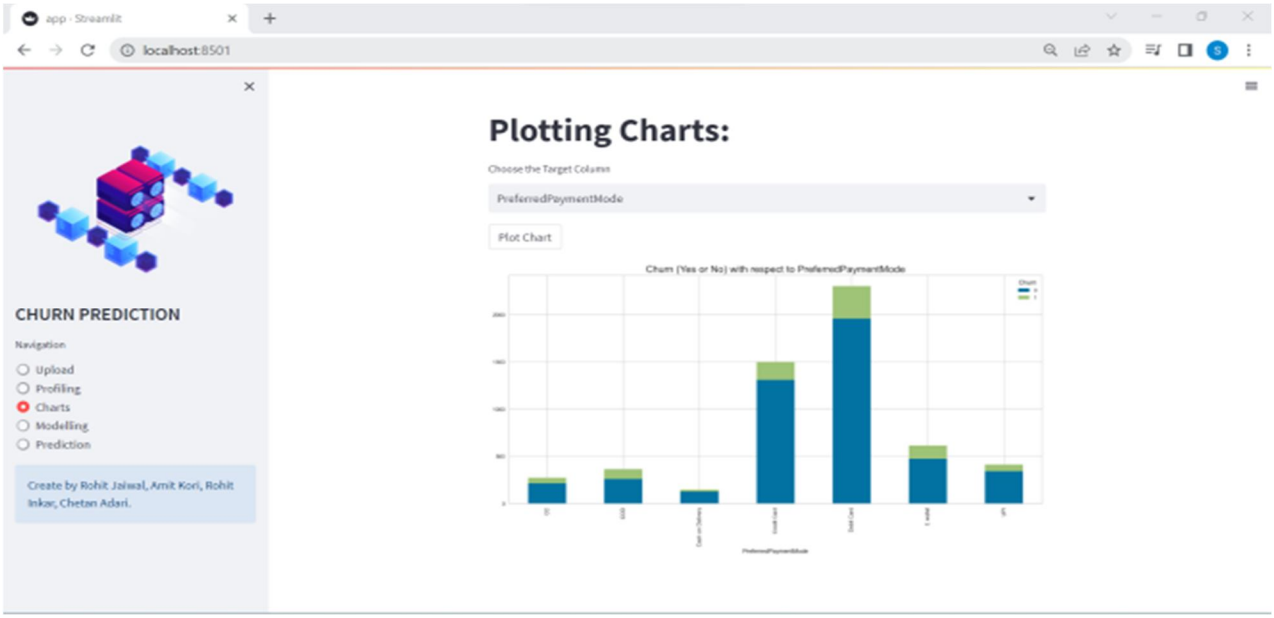
3) *Step 3:* This is the loading phase where your file is being uploaded.



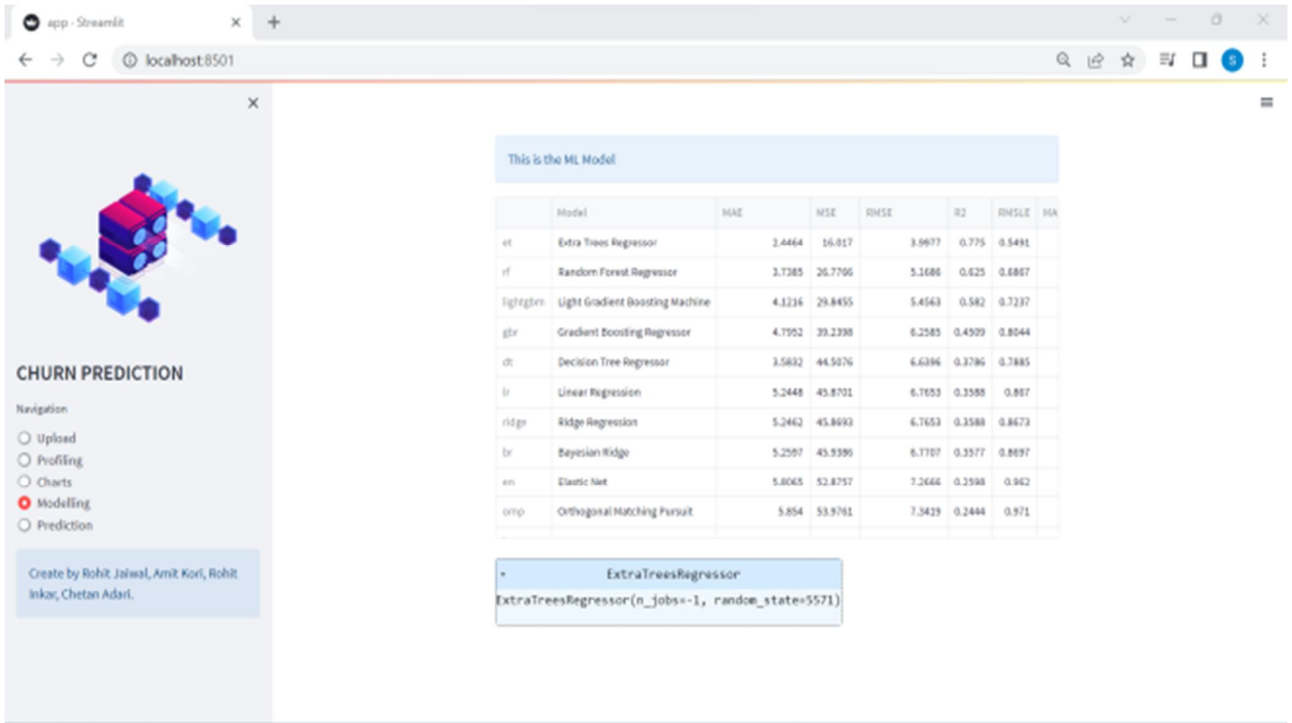
4) *Step 4:* Data Exploration



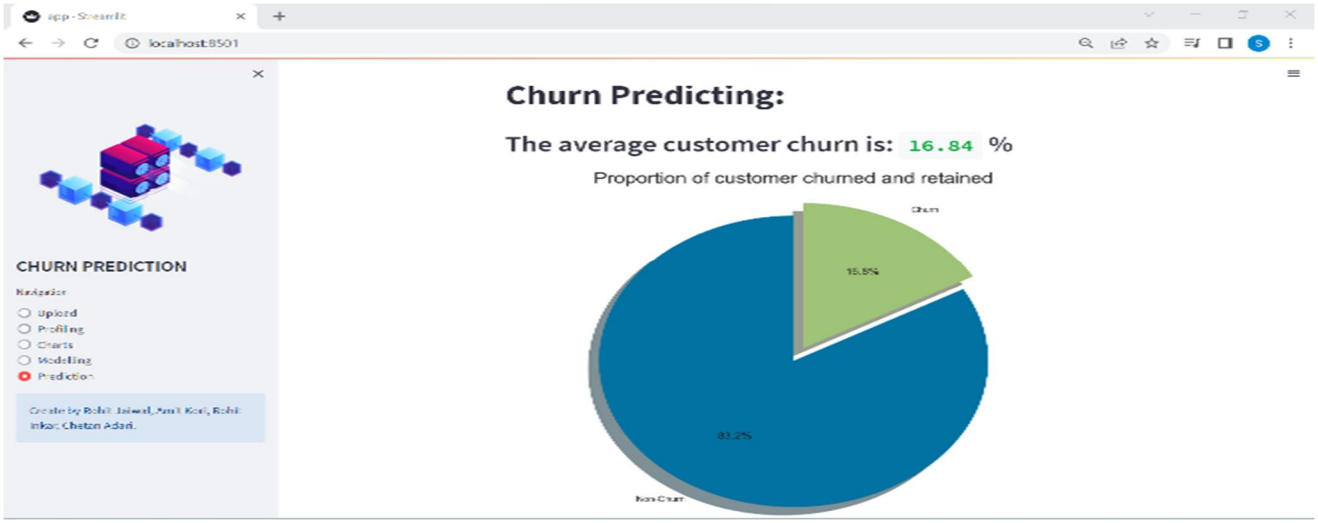
5) *Step 5:* After importing the necessary libraries, we have read the data and discovered that there are 5630 observations with missing values in some of the features. We will remove the irrelevant CustomerID column before proceeding further. Moving on to handling outliers, we will now explore if there are any outliers in our feature columns.



6) *Step 6:* We will now create visualizations for each variable in the dataset and their corresponding churn value. This will help us understand the relationship between each variable and churn. After visualizing the data, we will pre-process it by handling missing values, encoding categorical variables, and scaling the numerical features. Once the data is pre-processed, we will split it into training and testing sets and then train our models. We will train four base learners - Decision Trees, Random Forests, Support Vector Machines, and KNN classifiers. The outputs of these models will be fed into the Stacking Classifier's meta-classifier using logistic regression. Finally, we will evaluate the performance of our models using various metrics such as accuracy, precision, and recall.



7) *Step 7:* To gain further perceptivity, we can calculate the chance of churn contributed by each order for each variable. This information can be presented in the form of pie maps, showing the average client churn for each order.



IX. CONCLUSION

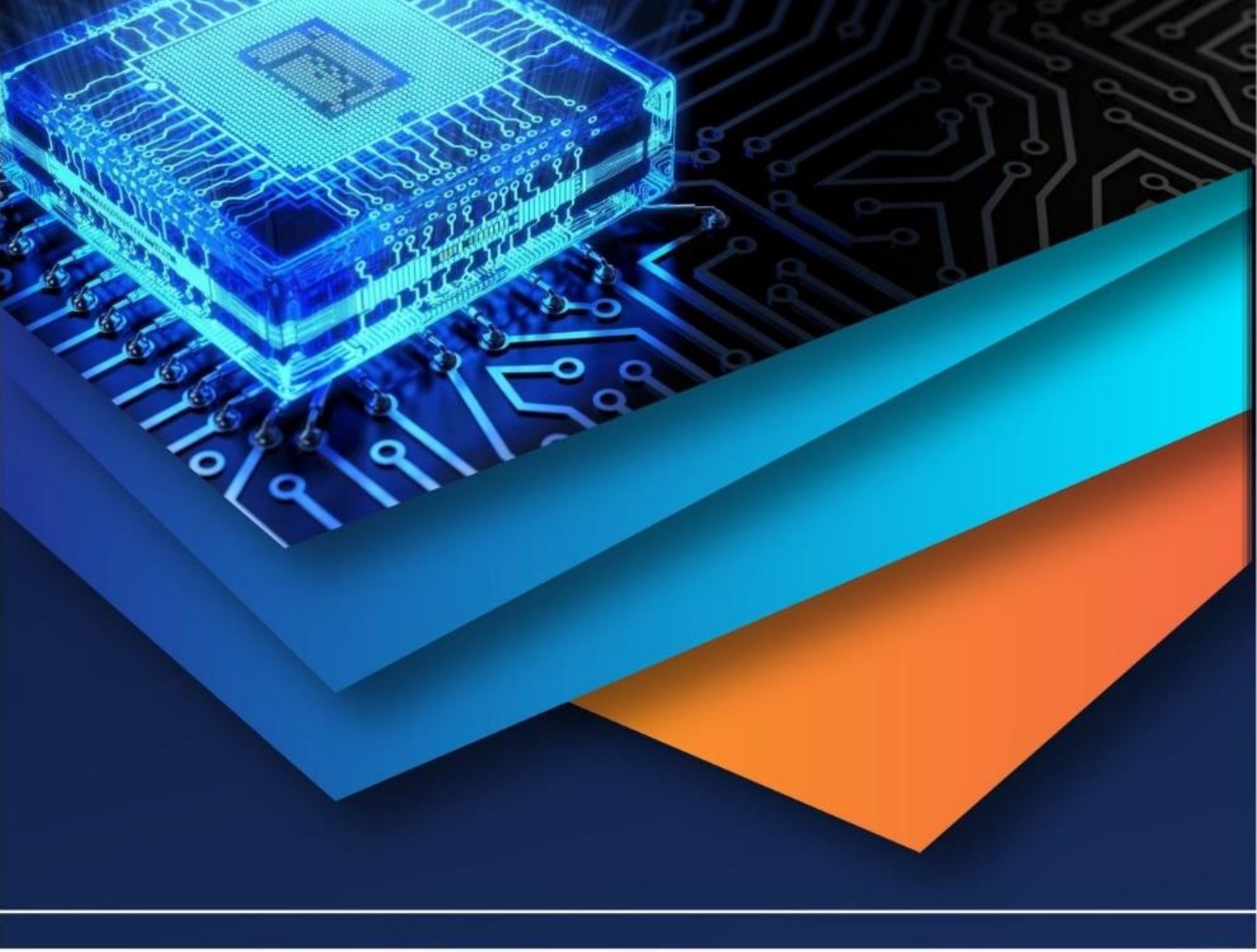
The ability to predict customer churn is crucial for e-commerce companies to remain competitive. Employing machine learning techniques in customer relationship management can aid companies in forecasting potential customer loss and devising effective marketing and retention strategies. This study aimed to evaluate the predictive ability of SVM and LR models using customer behaviour data from a B2C e-commerce enterprise. The k-means algorithm was employed for clustering subdivision to classify customers into three categories, and predictions were made for each category. The performance of the models was evaluated using accuracy, recall, precision, and AUC metrics.

The study had two primary objectives. Firstly, to assess the efficacy of customer segmentation and the predictive power of the model before and after segmentation based on customer shopping behaviour. The results indicated a substantial improvement in prediction accuracy after implementing k-means clustering segmentation. Secondly, to compare the performance of traditional statistical LR model prediction with machine learning-based SVM model prediction. The SVM model outperformed the LR model in terms of accuracy.

In conclusion, the research findings offer valuable insights for B2C e-commerce companies' customer relationship management efforts

REFERENCES

[1] Bi, Q.Q. Cultivating loyal customers through online customer communities: A psychological contract perspective. J. Bus. Res. 2019, 103, 34–44.
[2] Maria, O.; Bravo, C.; Verbeke, W.; Sarraute, C.; Baesens, B.; Vanthienen, J. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. Expert. Syst. Appl. 2017, 85, 204–220.
[3] Roberts, J.H. Developing new rules for new markets. J. Acad. Market. Sci. 2000, 8, 31–44.
[4] Reichheld, F.F.; Sasser, W.E. Zero defeofions: Quoliiy comes to services. Harvard. Bus. Rev. 1990, 68, 105–111.
[5] Jones, T.O.; Sasser, W.E., Jr. Why satisfied customer's defect. IEEE Eng. Manag. Rev. 1998, 26, 16–26.
[6] Nie, G.; Rowe, W.; Zhang, L.; Tian, Y.; Shi, Y. Credit card chum forecasting by logistic regression and decision tree. Expert. Syst. Appl. 2011, 38, 15273–15285.
[7] Gordini, N.; Veglio, V. Customers churn prediction and marketing retention strategies: An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind. Market. Manag. 2017, 62, 100–107.
[8] Zorn, S.; Jarvis, W.; Bellman, S. Attitudinal perspectives for predicting churn. J. Res. Interact. Mark. 2010, 4, 157–169.
[9] Datta, P.; Masand, B Automated cellular modeling and prediction on a large scale. Artif. Intell. Rev. 2000, 14, 485–502.
[10] Jain, H.; Khunteta, A.; Srivastava, S. Churn prediction in telecommunication using logistic regression and logit boost. Procdia Compute. Sci. 2020, 167.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY