IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Review On Customer Churn Prediction Application

Sagar Chaudhary, Chinmay Patil, Channabasava U, Krishnapal Parmar
B.E Students, Dept. of ISE, BIT, Bengaluru, Karnataka, India

Prof Prameela R
Assistant Professor, Dept. of ISE, BIT, Bengaluru
Karnataka, India

Abstract: Customer churn prediction is crucial for e-commerce companies looking to improve retention and optimize business strategies. In this study, we aim to build a robust predictive model to identify customers at risk of churn, using a variety of customer features such as demographics, behavior metrics, order history, and satisfaction scores. The analysis begins with an exploratory data analysis (EDA), using Python libraries like pandas, numpy, matplotlib, and seaborn for data manipulation, visualization, and understanding patterns in the dataset. Missing values are visualized and addressed using missing no, ensuring data quality before model development. Several machine learning models are trained to predict customer churn, including Support Vector Classifier (SVC), Logistic Regression, Random Forest Classifier, and XGBoost, implemented with libraries such as scikit-learn and xgboost. Data preprocessing steps, such as scaling features with Standard Scaler and encoding categorical variables using LabelEncoder, are performed to prepare the dataset for model training. Model evaluation is conducted using performance metrics like accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix, offering detailed insights into the effectiveness of each model. In addition, hyperparameter tuning is achieved using GridSearchCV to optimize model performance. Cross-validation is applied with cross validate to ensure generalizability and robustness of the models. The study identifies key factors influencing churn, including marital status, order categories, and platform engagement, and provides actionable insights for the company to implement targeted retention strategies. By leveraging machine learning and data analytics, this research helps the e-commerce company proactively manage churn, enhance customer satisfaction, and optimize product offerings tailored to customer preferences.

Keywords— Customer Churn Prediction, Customer Segmentation, Exploratory Data Analysis (EDA), Predictive Modeling, Voting Classifier, Machine Learning, Feature Scaling.

I. INTRODUCTION

Customer retention is a vital aspect of business success, especially in the competitive and rapidly evolving e-commerce industry. Churn, the process where customers discontinue using a company's services, poses significant challenges by impacting revenue and increasing customer acquisition costs. Understanding and addressing the factor's influencing churn is essential for businesses aiming to maintain a competitive edge and foster long-term customer loyalty. In this study, we analyse customer behaviour and interactions using a dataset from a leading e-commerce company, with the goal of predicting and mitigating churn.The dataset contains detailed information on various customer attributes, including demographic data, transactional behaviours, satisfaction scores, preferred payment methods, and device usage. By exploring the relationships between these features and customer churn,

this research aims to uncover actionable insights. For instance, factors such as gender, marital status, city tier, purchasing behaviour, and satisfaction levels are examined to identify patterns that influence customer retention. Additionally, the study investigates how specific behaviours, like coupon usage or increases in order amounts, impact churn rates. To achieve these objectives, machine learning models are employed to predict churn and identify at-risk customers. The findings provide valuable insights into customer preferences and behaviours, enabling the company to design targeted strategies for improving customer satisfaction and retention. This research contributes to the broader understanding of churn analytics while offering practical recommendations for enhancing customer loyalty in the e-commerce sector.

II. LITERATURE REVIEW

Chitra and K. Rajalakshmi proposed the use of machine learning algorithms for customer churn prediction in e-commerce platforms.Their study utilized customer transaction data, browsing behavior, and demographic information to develop predictive models, including decision trees, random forests, and support vector machines. These models were shown to significantly enhance churn prediction accuracy and enable businesses to take preventive actions. The study concluded that integrating predictive models with targeted marketing strategies and personalized customer service can effectively reduce churn [1].

Wenqi Li, Yiran Liu, and Xiao Zhang proposed using customer lifetime value (CLV) as a predictive metric for churn in e-commerce businesses. By analyzing customer purchase histories and engagement data, they demonstrated that CLV can act as an early warning system for identifying high-risk churn segments. Their research emphasized the importance of data integration from multiple touchpoints, such as social media and website interactions, and suggested retaining high-value customers through loyalty programs and personalized experiences [2].

Xiaoyu Zhang and Liwei Wang explored the application of deep learning techniques for churn prediction in e-commerce. Their study focused on neural network models, particularly recurrent neural networks (RNN) and long short-term memory (LSTM), to analyze sequential customer data such as browsing history, purchase patterns, and feedback. The research highlighted that deep learning models can outperform traditional models by capturing temporal dependencies in user behavior, despite challenges in large-scale data processing and model interpretability [3].

John Carter and Sarah Patel conducted a case study on Amazon's customer retention strategies, emphasizing personalized marketing techniques such as targeted product recommendations, loyalty rewards, and fast delivery services. Their research highlighted the role of customer segmentation and tailored retention efforts based on user behaviors and preferences. The study concluded that continuous improvements in customer experience and engagement with personalized offers can significantly reduce churn in e-commerce platforms [4].

I. PROBLEM STATEMENT

Customer churn is a major challenge for e-commerce businesses, as it leads to lost revenue and higher costs to acquire new customers. Understanding why customers leave and identifying those at risk of churning is essential for improving retention. However, customer behavior is complex, and it is difficult to pinpoint the key factors driving churn.This project focuses on using customer data to build a model that predicts which customers are likely to churn. By analyzing factors such as demographics, purchasing habits, satisfaction levels, and payment preferences, the goal is to uncover patterns and insights that can help the company take timely action to retain customers and reduce churn rates.

II. OBJECTIVES

The primary objective of this project is to develop an efficient and automated system for predicting customer churn in an e-commerce platform. To achieve this, several specific objectives need to be met:

- o Build a Machine Learning Model to Predict Customer Churn
- o Perform Feature Importance Analysis to Identify Key Churn Drivers
- o Evaluate Behavioral Trends Using Exploratory Data Analysis (EDA).
- o Assess the Relationship Between Customer Demographics and Churn.
- o Investigate the Correlation Between Satisfaction Metrics and Retention.
- o Analyze the Impact of Transactional Variables on Churn Propensity.

III. MOTIVATION

In the competitive e-commerce industry, retaining customers is more cost-effective than acquiring new ones, making churn prediction a critical business need. High churn rates not only reduce revenue but also impact brand loyalty and long-term growth. Understanding the underlying reasons for churn and addressing them can significantly improve customer satisfaction and trust. This project is motivated by the potential to leverage data analytics and machine learning to provide actionable insights into customer behaviour. By identifying at-risk customers and understanding their preferences and pain points, the company can design targeted strategies to enhance the customer experience. The ultimate goal is to empower the business to make informed, proactive decisions that foster loyalty, improve retention, and drive sustainable growth.

IV. SYSTEM DESIGN

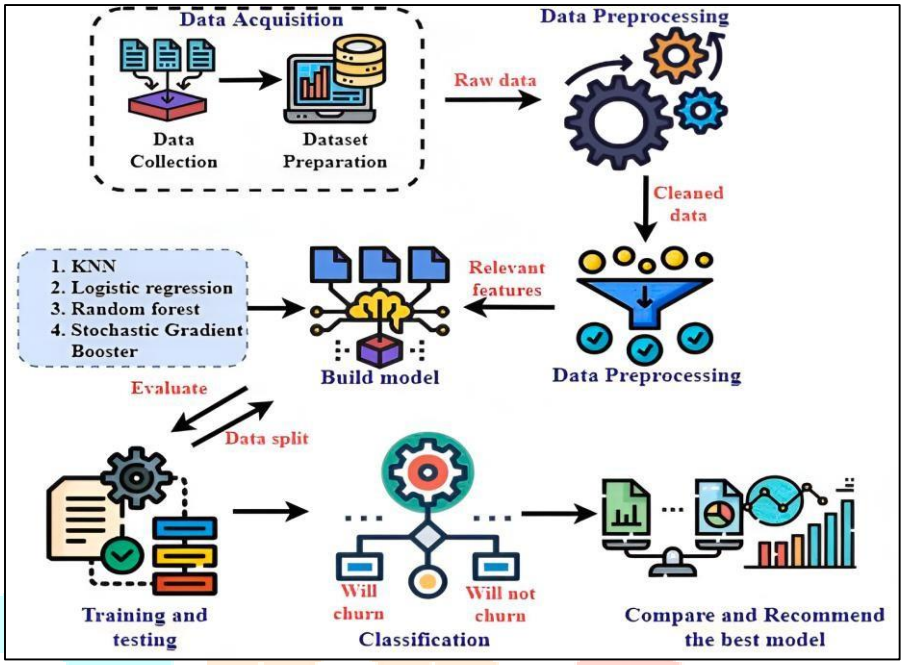
The system design for E-Commerce Customer Churn Prediction represents a structured integration of advanced components and modules to deliver accurate and actionable outcomes. At the core of this system lies customer data, encompassing transactional records, browsing behavior, and demographic attributes. This data forms a well-organized repository, serving as the foundation for all subsequent processes. To ensure the data is primed for analysis, the system implements a preprocessing pipeline that involves data cleaning, normalization, and feature extraction to enhance data quality and relevance. This stage is crucial, as it ensures that the raw input is transformed into a format suitable for in-depth analysis.

The preprocessing steps include handling missing values, encoding categorical variables, and scaling numerical data to maintain consistency across features. Once preprocessing is complete, the data is divided into subsets, typically with 80% allocated for training and 20% reserved for testing. This split ensures that the system has sufficient data for learning while maintaining an unbiased evaluation framework. During the training phase, the system employs a combination of machine learning models, such as decision trees, random forests, and support vector machines, to identify patterns and relationships within the processed data. These models are fine-tuned through hyperparameter optimization to achieve the best predictive performance.

Additionally, the system incorporates visualization tools that display data insights in an intuitive manner, such as churn probability, key influencing factors, and customer segmentation. These visualizations are crucial for stakeholders to interpret results and devise targeted strategies. The modular design of the system enables seamless integration with marketing platforms for deploying personalized retention strategies, such as loyalty programs and exclusive offers. Furthermore, the system supports scalability, allowing it to handle increasing data volumes while integrating future advancements like real-time analytics and adaptive learning models.

By leveraging predictive modeling, the system facilitates early identification of at-risk customers, allowing businesses to take proactive measures to reduce churn. Its scalable and adaptable architecture ensures that the system can evolve with future advancements, including incorporating deep learning techniques or integrating

real-time data processing capabilities. This robust framework underscores the potential of technology-driven solutions in enhancing customer retention and driving business success in e-commerce.



Architectural Design

The image illustrates the workflow for an e-commerce customer churn prediction system. It includes data acquisition, preprocessing, feature extraction, and model building using algorithms like KNN, logistic regression, random forest, and stochastic gradient boosting. The process ends with model evaluation, customer classification, and recommendations for the best-performing model

III. IMPLEMENTATION

A. Base Learners

Base learners are individual models that contribute to the ensemble prediction. Their combination captures various aspects of the data, enhancing robustness and reducing errors. The base learners employed in this study are as follows:

1. Support Vector Machine (SVM)

Constructs a hyperplane in a high-dimensional space to separate customers into churn and non-churn categories by maximizing class margins.

2. Logistic Regression (LR)

A probabilistic model that predicts the likelihood of a customer churning based on the provided features.

3. K-Nearest Neighbors (KNN)

The KNN algorithm classifies data based on the similarity of nearby data points. For a given query point xxx, distances to all training points xi are computed using a distance metric such as Euclidean distance

4. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and robustness. For a given input x, the prediction is made by aggregating the outputs of N trees.

5. XGBoost (XGB)

A gradient-boosting algorithms that combines sequential decision trees to minimize prediction errors effectively.

B. System Workflow

The workflow of the Customer Churn prediction system involves the following steps

1. Data Acquisition: Collecting a dataset containing customer records, including relevant features for churn prediction, such as demographics, purchase history, app usage, and satisfaction metrics.
2. Data Preprocessing: Cleaning And Handling Missing Values, Encoding Categorical Variables, Scaling Numerical Features using standardization to ensure compatibility with machine learning models.
3. Model Training: Training individual base learners—SVM, Logistic Regression, Random Forest, XGBoost, and KNN—on the preprocessed dataset.
4. Soft Voting Implementation: Aggregating the probabilistic outputs of all base learners to compute the final class label (churn or non-churn) using the soft voting mechanism.
5. Model Evaluation: E valuating the performance of the ensemble model using metrics like accuracy, precision, recall, F1-score, and ROC-AUC to assess its reliability and robustness.

Output Classifying customers into churn or non-churn categories with a confidence score for each prediction. Providing actionable insights based on feature importance and churn risk factors to guide retention strategies.

VII. RESULT AND PERFORMANCE ANALYSIS

This bar chart compares the test accuracies of various machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, XGBoost, and AdaBoost. The accuracy scores are color-coded, with deeper reds indicating higher performance, showing Random Forest and XGBoost as the top-performing models.

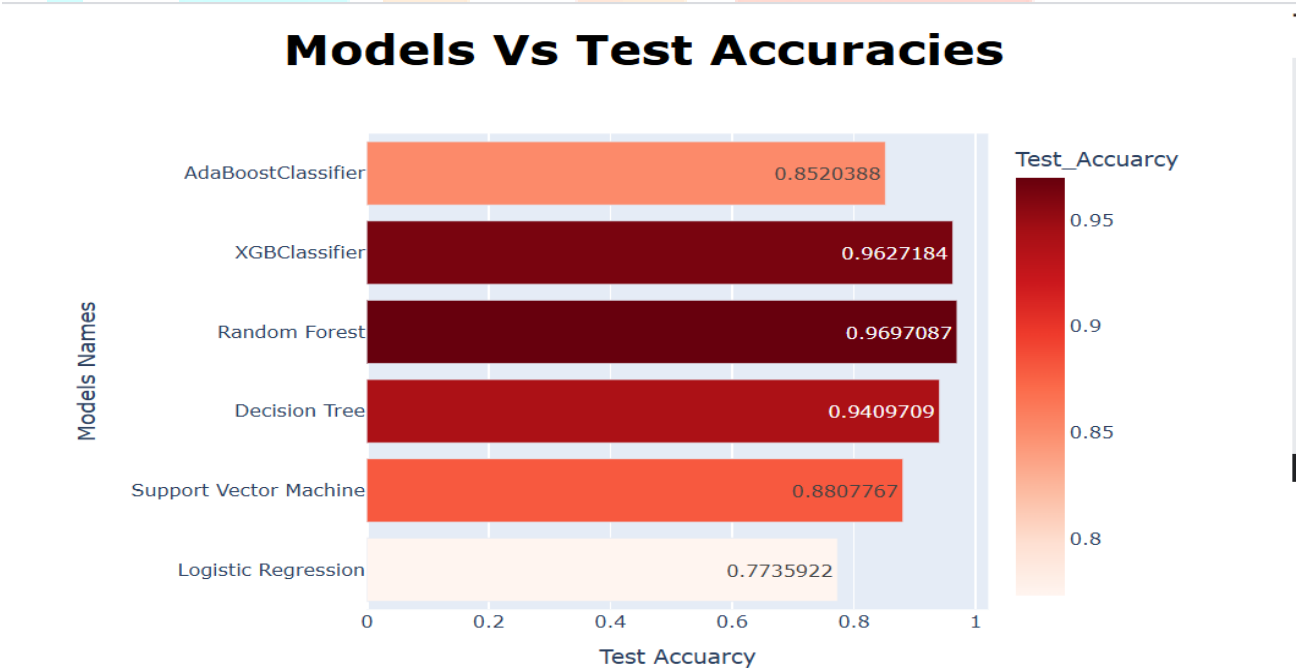


Fig 6.1 Performance Metrics

The result analysis focused on evaluating the performance of different machine learning models in predicting customer churn. Each model was assessed using key metrics like accuracy, precision, recall, F1-score, and ROC-AUC to measure its effectiveness in identifying customers at risk of churning. Among the models tested, **Random Forest** emerged as the most reliable, delivering the highest accuracy and a strong balance between precision and recall. This suggests it effectively captured the patterns in the data and minimized both false positives and false negatives. **Support Vector Machine (SVM)** also performed well, showing strong recall, indicating that it was particularly good at identifying customers who were likely to churn. On the other hand, **Logistic Regression** and **K-Nearest Neighbors (KNN)** demonstrated lower recall rates, meaning they missed some churned customers, which may reduce their reliability for actionable predictions. Overall, Random Forest's superior performance

makes it the most suitable model for this task, but the evaluation highlights the trade-offs each model presents in terms of different metrics.

IV. CONCLUSION

In this project, we focused on predicting customer churn for an e-commerce company using various machine learning techniques, with the primary goal of developing a model that could accurately identify customers at risk of leaving. Through the process, we explored key features such as gender, marital status, city tier, and customer satisfaction, and examined their relationship to churn rates. Our analysis revealed significant patterns, such as higher churn rates among males and singles, as well as insights into customer behaviour based on city tiers and preferred products. Using ensemble methods, including classifiers like Random Forest, Logistic Regression, and Support Vector Machine, we were able to leverage the strengths of multiple models to enhance prediction accuracy. By implementing soft voting for ensemble modelling, we further improved the robustness of the model, resulting in more reliable predictions. Moreover, the results highlighted actionable strategies for the company to mitigate churn, such as customizing products for male customers and tailoring marketing efforts toward singles who are at a higher risk of leaving. By aligning the company's offerings with customer preferences and behaviour's, the company can better address customer needs and foster long-term loyalty. Finally, this project demonstrates the value of predictive analytics in business decision-making, showing how advanced techniques like ensemble learning can be used to drive insights that improve customer retention and overall business performance.

ACKNOWLEDGEMENT

We convey our sincere thanks to Rajya Vokkaligara Sangha, Bangalore and our guide Prof Prameela R, Associate Professor, Department of Information Science and Engineering, Bangalore Institute of Technology, without whose direction, this would not have been possible. We also express our gratitude to our team members whose team participation resulted in successful completion of the paper. We are also grateful to our institution and its faculty for providing the necessary resources and a conducive environment for research. Lastly, we extend our heartfelt thanks to our families and friends for their unwavering support and encouragement throughout this endeavor.

REFERENCES

[1] Kumar, V., & Shah, D. "Building and sustaining profitable customer loyalty for the 21st century." *Journal of Retailing*, vol. 80, no. 4, 2004, pp. 317-330.

[2] Zhou, Z. H. "Ensemble methods: Foundations and algorithms." *CRC Press*, 2012.

[3] Cortes, C., & Vapnik, V. "Support-vector networks." *Machine Learning*, vol. 20, no. 3, 1995, pp. 273-297.

[4] Hosmer, D. W., & Lemeshow, S. "Applied logistic regression." *John Wiley & Sons*, 2000.

[5] Breiman, L. "Random forests." *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.

[6] Chen, T., & Guestrin, C. "XGBoost: A scalable tree boosting system." *Proceedings of the 22nd ACM SIGKDD International —Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.

[7] Cover, T. M., & Hart, P. E. "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory*, vol. 13, no. 1, 1967, pp. 21-27.

[8] Zeng, D., & Qin, X. "Ensemble learning for customer churn prediction." *Procedia Computer Science*, vol. 147, 2019, pp. 137-142.

[9] Dhanraj, K. V., & Kshirsagar, M. "Data preprocessing and feature engineering for churn prediction models." *International Journal of Computer Applications*, vol. 111, no. 9, 2015, pp. 39-43.

[10] Gupta, S., & Zeithaml, V. A. "Customer metrics and strategies in the context of customer churn prediction." *Journal of Services Marketing*, vol. 20, no. 2, 2006, pp. 78-89.