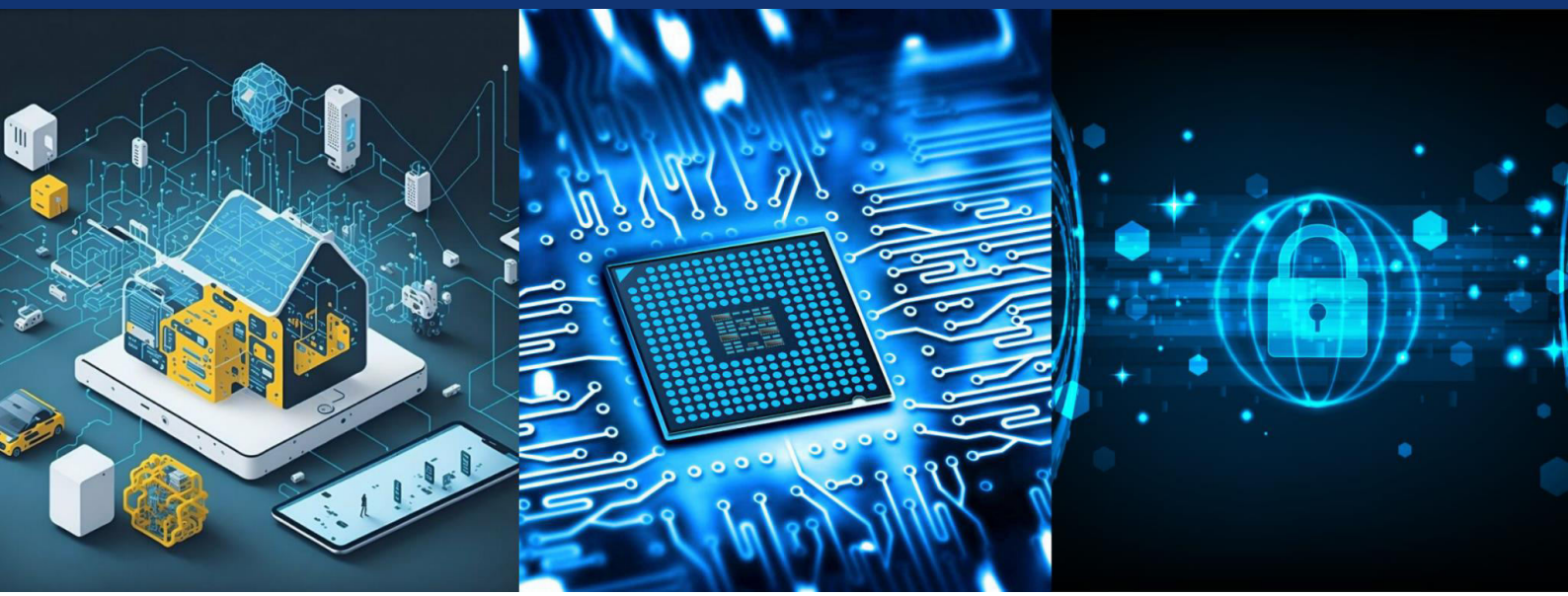
IJIRCCE

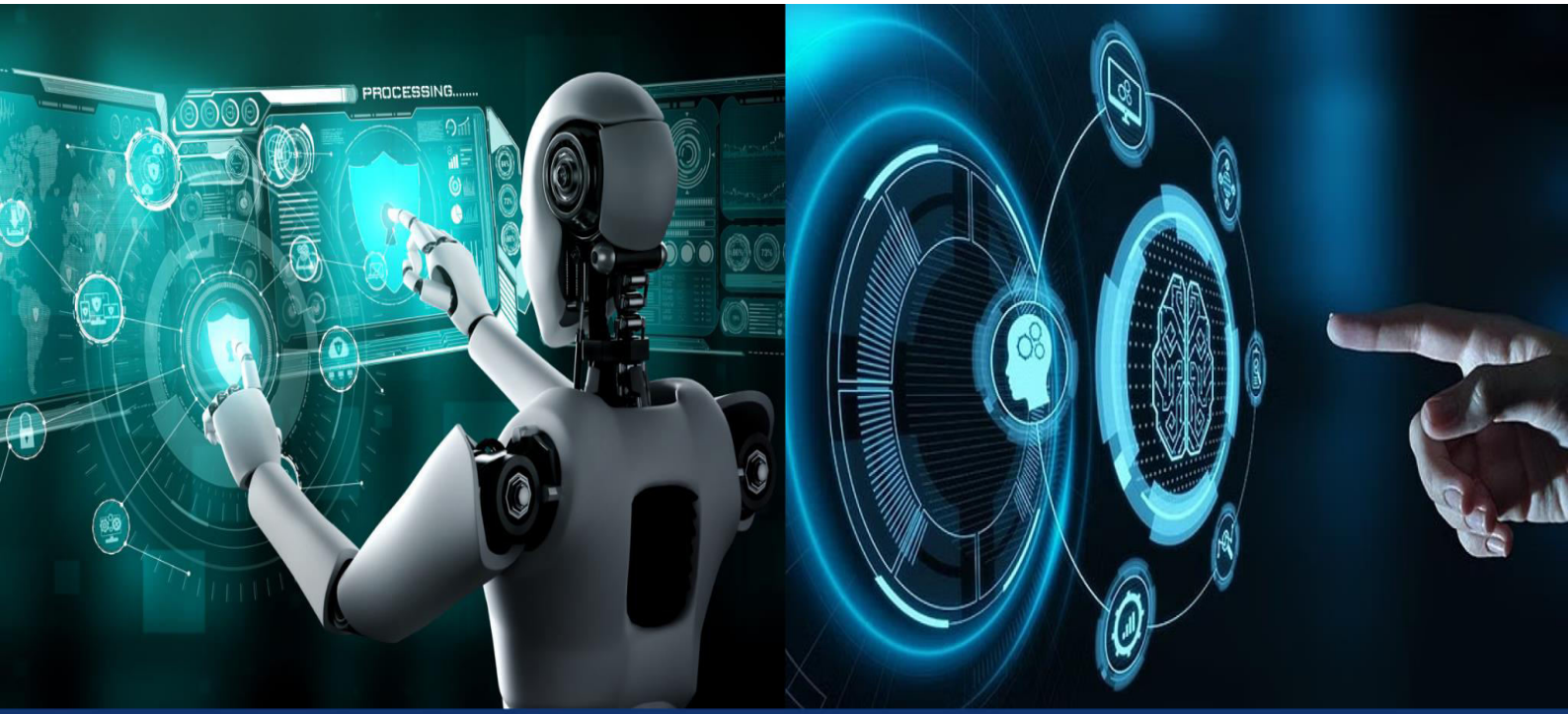
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Data Driven Customer Churn Prediction in E-Commerce: A Machine Learning Approach for Data Science

V Ravi Charan[1], K Vasuki[2],R V Sai Keerthana[3],D Neela Rao[4] ,Dr C.M.Suvarna Varma[5]

U G Students, Dept.of CSE-DS.,SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[1-4]

Professor, Dept.of CSE-DS, SRK Institute of Technology, Enikepadu, Vijayawada, Andhra Pradesh, India[5]

ABSTRACT: Customer churn in the e-commerce industry occurs when customers stop purchasing products or services from an online platform, significantly impacting both revenue and long-term business growth. Retaining customers has become increasingly crucial in today's highly competitive and rapidly evolving digital marketplace. A strong and loyal customer base not only ensures consistent sales but also fosters brand trust, organic referrals, and increased lifetime value. This study aims to analyze e-commerce user data with the goal of predicting which customers are likely to disengage from the platform. By employing a range of machine learning algorithms, the study evaluates customer behavior patterns and compares predictive performance across various accuracy and efficiency metrics.In addition to churn prediction, the analysis enables data-driven decision-making by uncovering actionable insights into customer dynamics. By identifying key churn indicators—such as reduced activity, order frequency decline, or dissatisfaction signals—businesses can implement timely interventions to re-engage users. Proactive strategies such as personalized marketing, loyalty programs, and improved customer support can be guided by these findings. Ultimately, understanding and addressing churn helps e-commerce platforms strengthen customer relationships, reduce acquisition costs, and sustain long-term profitability in a highly dynamic environment.

KEYWORDS: E-commerce attrition ,Churn prediction ,Machine learning ,XGboost ,Random forest

I. INTRODUCTION

Customer retention is crucial for e-commerce success, as customer churn leads to significant revenue losses. Predicting churn involves analyzing customer behavior, purchase history, and interaction patterns to identify at-risk customers. By leveraging machine learning techniques, businesses can uncover key indicators of churn and implement targeted interventions such as personalized offers and enhanced customer service to improve retention and loyalty.

This project aims to develop a robust churn prediction model using machine learning algorithms like logistic regression, decision trees, random forests, and XGBoost. Effective feature engineering and data preprocessing will enhance model accuracy, while evaluation metrics such as precision, recall, and F1-score will ensure reliability. By accurately identifying potential churners, e-commerce businesses can optimize marketing strategies, improve customer experience, and gain a competitive edge through data-driven decision-making

In the paper, we explained the related works in section II , Background of which algorithms used and implemented in our churn prediction is explained in section III in detail. Our Proposed System section IV , CCP-Boost algorithm uses an ensemble method of classifiers where XGBoost concepts are the base for our algorithm and in section V the comparitive results are shown using graphical representations and we concluded and future works focused on section V1

II. RELATED WORKS

Customer churn prediction is crucial in e-commerce for understanding customer behavior and minimizing revenue loss. Researchers apply machine learning techniques like logistic regression, decision trees, and deep learning to analyze customer transactions, browsing history, and purchase frequency. Feature engineering, incorporating factors such as

order frequency, cart abandonment rates, and engagement levels, enhances predictive accuracy. Advanced methods like ensemble learning (Random Forest, XGBoost) and deep learning (LSTMs, neural networks) further improve performance. Customer segmentation using clustering techniques (K-Means, DBSCAN) helps identify behavior patterns, while NLP analyzes reviews and support tickets to detect early churn signals. Explainable AI (XAI) ensures transparency, fairness, and ethical considerations in churn models. Researchers emphasize privacy concerns as companies collect vast user data. Real-time analytics and reinforcement learning are emerging trends, enabling personalized retention strategies. These innovations help e-commerce platforms proactively reduce churn and enhance customer experience

The existing literature on customer churn primarily focuses on the financial sector, emphasizing the importance of understanding customer behavior to mitigate attrition. Studies such as Geiler et al. (2022) and Machado and Karray (2022) highlight behavioral insights but fall short in offering comparative analyses of machine learning (ML) techniques. Lemmens and Gupta (2020) explore the profit implications of churn without delving into predictive modeling. While Al-Mashraie et al. (2020) and De Lima Lemos et al. (2022) recognize the potential of ML in forecasting churn, they lack in-depth comparisons of model performance.

S.no	Paper _information	Description	Limitations / Inference
1.	Geiler et al. (2022); Machado and Karray (2022)[1]	It discusses customer churn in the financial sector and highlights the importance of understanding customer behavior to reduce attrition.	Does not provide a detailed comparative analysis of various ML techniques for churn prediction.
2	Lemmens and Gupta (2020)[2]	Focuses on managing churn to maximize profits, particularly by investigating the profit-loss ratio when customers stop using products.	Lacks a deep analysis of predictive modeling techniques for churn prevention.
3	Al-Mashraie et al. (2020)[3]	Provides a foundation for further research in predicting bank customer attrition.	Does not explore specific methodologies for improving churn prediction accuracy.
4	De Lima Lemos et al. (2022)[4]	Highlights how ML can help track and forecast customer churn by monitoring behavioral changes.	Lacks a detailed comparison of various ML models to guide banks in decision-making.
5	Dias et al. (2020)[5]	Discusses how ML-based models can predict customer attrition and allow banks to focus on high-risk customers.	Limited discussion on feature selection, model evaluation metrics, and real-world application challenges.
6	Schaeffer and Sanchez (2020)[6]	Discusses the importance of customer retention strategies in financial institutions and how churn impacts business profitability.	Lacks an empirical comparison of different machine learning models for churn prediction.
7	He et al. (2014)[7]	Examines the role of customer behavior analytics in predicting	The study is dated, and newer ML approaches might outperform the

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

		churn and emphasizes the importance of early intervention.	proposed methods.
8	Karvana et al. (2019)[8]	Investigates feature engineering techniques to enhance churn prediction accuracy in banking.	Focuses on feature selection but does not provide insights into deep learning models for churn prediction.

III. BACKGROUND

1.Machine Learning Models

1.1 XGBClassifier (Extreme Gradient Boosting Classifier)

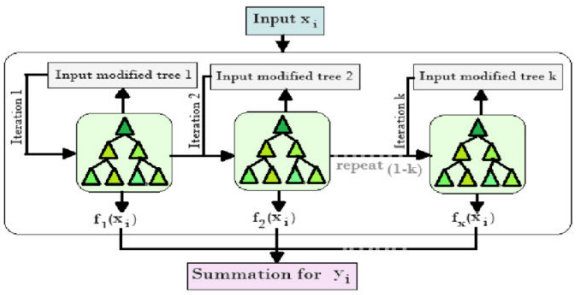


Fig1:XGBoost

In Fig1 we explain the XGBoost algorithm, as depicted in the image, iteratively refines predictions by sequentially building decision trees, where each tree focuses on correcting the errors made by the previous ones; beginning with an input data point xi, each tree fk(xi) is constructed based on the residuals of the preceding tree, effectively "boosting" the model's accuracy through successive iterations, and finally, the predictions from all trees are summed to produce the final output yi, resulting in a highly accurate and robust predictive model.

1.2 RandomForestClassifier

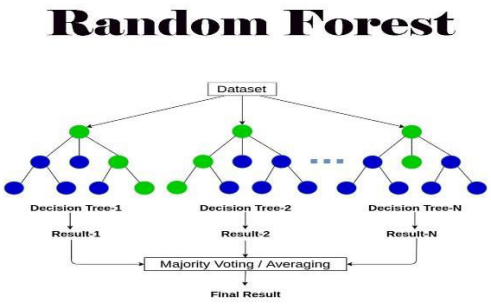


Fig2:Random Forest

Fig2 illustrates the core concept of Random Forest: it operates by constructing multiple decision trees. Each tree, as shown, is built upon a random subset of the data and features, leading to diverse structures. The branching within each tree, represented by nodes splitting into green and blue circles, signifies decisions based on feature values. Finally, the algorithm aggregates the predictions from all trees through a process like majority voting or averaging, as depicted by

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

the "Majority Voting / Averaging" box at the bottom. This ensemble approach, where the "Final Result" is derived from the collective wisdom of multiple trees, significantly reduces overfitting and enhances the model's robustness and accuracy.

1.3 LogisticRegressionCV (Logistic Regression with Cross-Validation):

In Fig3,Logistic Regression, as depicted, models the probability of a binary outcome (y=0 or y=1) based on input features (X). It applies a linear combination of features, then transforms this using the sigmoid function, producing an S-shaped curve that maps predictions to a range between 0 and 1, representing probabilities. A threshold is set to classify the outcome, often 0.5, determining whether the predicted probability is closer to 0 or 1. The algorithm optimizes coefficients to best fit the data, minimizing the error between predicted and actual outcomes. This makes it suitable for classification tasks where the goal is to predict the likelihood of an event occurring.Along with them we have used another classifiers like BaggingClassifier, AdaBoostClassifier, KNeighborsClassifier, RidgeClassifier.

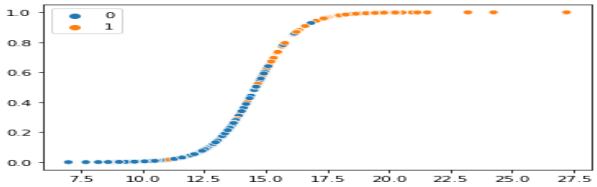


Fig3: Logistic Regression

2.Dataset

A e-commerce dataset with the following attributes are taken for analysis to perform the prediction as shown in Table 1.

Data	Variable	Description
E Comm	CustomerID	Unique customer ID
E Comm	Churn	Churn Flag
E Comm	Tenure	Tenure of customer in organization
E Comm	PreferredLoginDevice	Preferred login device of customer
E Comm	CityTier	City tier
E Comm	WarehouseToHome	Distance in between warehouse to home of customer
E Comm	PreferredPaymentMode	Preferred payment method of customer
E Comm	Gender	Gender of customer
E Comm	HourSpendOnApp	Number of hours spend on mobile application or website
E Comm	NumberOfDeviceRegistered	Total number of deceives is registered on particular customer
E Comm	PreferedOrderCat	Preferred order category of customer in last month
E Comm	SatisfactionScore	Satisfactory score of customer on service
E Comm	MaritalStatus	Marital status of customer
E Comm	NumberOfAddress	Total number of added added on particular customer
E Comm	Complain	Any complaint has been raised in last month
E Comm	OrderAmountHikeFromlastYear	Percentage increases in order from last year
E Comm	CouponUsed	Total number of coupon has been used in last month
E Comm	OrderCount	Total number of orders has been places in last month
E Comm	DaySinceLastOrder	Day Since last order by customer
E Comm	CashbackAmount	Average cashback in last month

Table 1. Dataset key attributes and its description

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Table1 presents a dataset focused on E-commerce customer behavior, outlining key variables like "CustomerID", "Churn" (indicating customer attrition), "Tenure", and demographic details such as "Gender" and "MaritalStatus". It also includes transactional information such as "OrderCount", "CouponUsed", "PreferredPaymentMode", and "CashbackAmount", alongside customer satisfaction metrics like "SatisfactionScore" and "Complain" status. The data aims to provide insights into customer engagement, purchasing patterns, and potential churn indicators for strategic business decisions.

IV. PROPOSED SYSTEM

1.Architecture

Fig4 illustrates the architecture of a customer churn prediction system of CCP Boost algorithm, outlining the flow of data and processes involved in identifying customers at risk of leaving. The system begins with the collection of **Customers Data**, encompassing various attributes such as demographics, usage patterns, and service interactions. This data serves as the foundation for the predictive model.

The core of the architecture is the **Data Driven Churn Prediction** module, where machine learning algorithms are applied to the customer data to forecast churn likelihood. This module analyzes historical patterns and correlations to identify key factors contributing to customer attrition. The output of this stage is **Churn Data**, which essentially represents the probability or score of each customer's likelihood to churn.

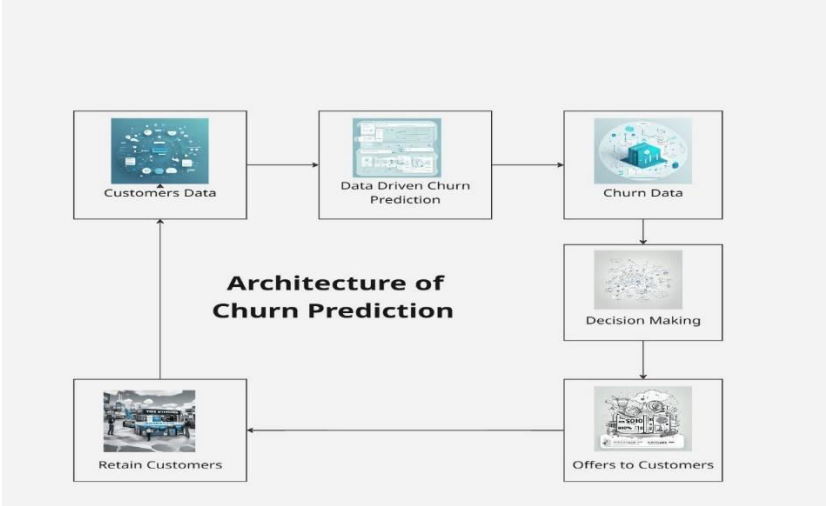


Fig4 : CCP Boost Architecture

Following the prediction, the system facilitates **Decision Making** based on the churn data. This involves segmenting customers based on their risk level and prioritizing those with a high propensity to churn. Subsequently, targeted **Offers to Customers** are generated, aiming to incentivize them to stay. These offers are tailored based on individual customer profiles and the factors influencing their potential churn.Finally, the system focuses on **Retaining Customers** through proactive engagement and personalized interventions. The effectiveness of these retention strategies is then fed back into the system, potentially influencing future data collection and model refinement, creating a closed-loop system for continuous improvement of churn prediction and customer retention.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2.Workflow

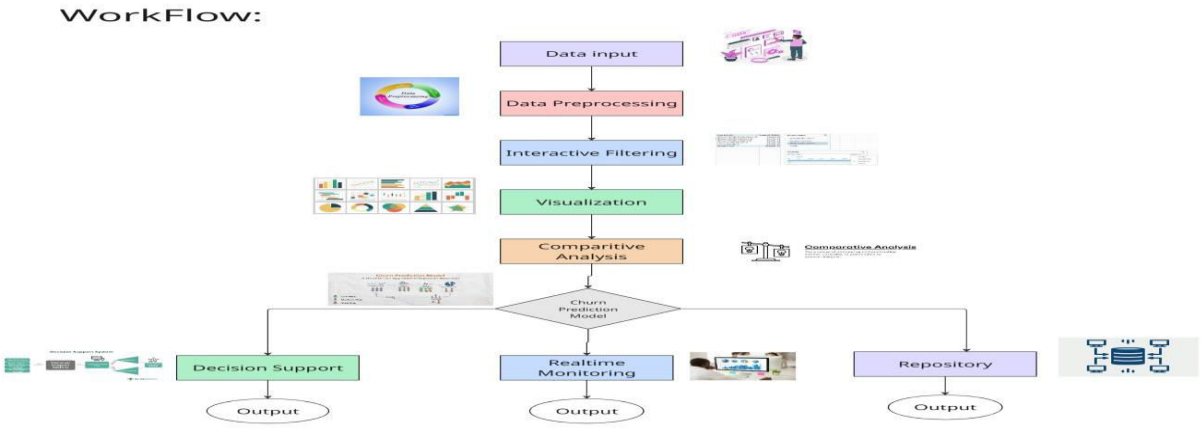


Fig 5:CCP Boost Workflow

Fig5 illustrates a comprehensive workflow for churn prediction and customer analysis, beginning with data input and preprocessing to refine raw data. Interactive filtering enables targeted data selection, followed by visualization for pattern identification and comparative analysis to understand customer behavior. The core of the process is a churn prediction model, which feeds into three critical outputs: decision support for strategic actions, real-time monitoring for immediate insights, and a repository for data storage and future analysis. This cyclical workflow ensures continuous improvement and informed decision-making based on evolving customer data

3.Data Collection and Preprocessing

3.1 Data Source:
The data typically comes from industries like telecom, banking, or e-commerce, where businesses want to predict whether a customer will churn (leave) or stay. This data is often stored in tabular formats like CSV or databases.

3.2 Handling Missing Values:

- Missing data in the dataset is common, and it's important to handle it. You can either use SimpleImputer, which replaces missing values with a specified strategy (e.g., mean, median, or mode), or IterativeImputer, which predicts missing values based on other data in the dataset.

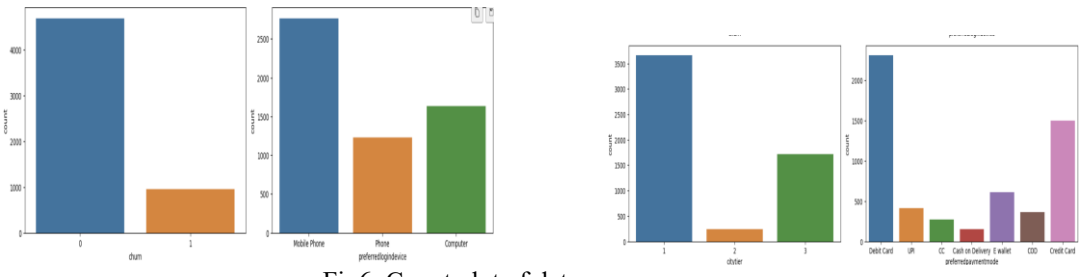


Fig6: Count plot of data

Fig6 displays a series of count plots, visually representing the distribution of categorical data across different variables, such as 'churn' and 'preferred login device', highlighting the frequency of each category within the dataset. These plots offer a quick understanding of the data's composition and potential imbalances, which can be crucial for further analysis and modeling.

3.3 Feature Selection and Encoding:
Feature Selection: Identifying which features (columns) are most relevant for predicting customer churn. For example, customer demographics, usage patterns, or contract details may be important.Encoding: Converting categorical

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

variables (like customer gender or subscription plan) into numerical values, so the machine learning model can understand them. This can be done with OneHotEncoder or LabelEncoder.

3.3.1
Splitting the Dataset:The data is split into a training set (usually 80%) to train the model and a testing set (usually 20%) to evaluate the model's performance. This ensures that the model is tested on unseen data to check its generalizability.

V. COMPARITIVE ANALYSIS AND RESULTS

1.Comparitive Analytic table

	model_name	test_accuracy	test_precision	test_recall	test_f1
7	CCP BoostClassifier	99.4	98.3	98.1	98.2
8	XGBoostClassifier	99.0	98.1	95.6	97.4
3	RandomForestClassifier	98.5	99.2	91.7	95.2
1	BaggingClassifier	98	97.6	90.4	93.9
2	GradientBoostingClasisifier	91.9	85.2	62.9	72.3
0	AdaBoostClassifier	89.8	75.8	58.2	65.8
4	LogisticRegressionCV	89.4	77.3	52.3	62.3
6	KNeighborsClassifier	87.5	70.4	44.8	54.8
5	RidgeClassifier	87.4	86.5	29.7	44.3

Fig7:Different model evalutions

Fig8 The table presents a comparative analysis of eight machine learning models, ranked by test accuracy, showcasing their performance across test precision, test recall, and test F1-score. CCPBoostClassifier leads with 99.4% accuracy and consistently high scores in all metrics, indicating robust performance. XGBoost Classifier,RandomForestClassifier and BaggingClassifier follow closely, demonstrating strong but slightly varied results.

A performance drop is observed with GradientBoostingClassifier, exhibiting significantly lower accuracy and metrics compared to the top three. AdaBoostClassifier and LogisticRegressionCV further this trend, with accuracies below 90% and declining precision, recall, and F1-scores, suggesting potential classification challenges.

KNeighborsClassifier and RidgeClassifierCV are the bottom performers, showing low recall, especially RidgeClassifierCV's 29.7%, indicating a bias towards predicting negative outcomes. This highlights the need for model optimization or alternative selection for improved results.

In summary, CCP Boost Classifier, XGBClassifier, RandomForestClassifier, and BaggingClassifier are the top performers, with CCP Boost Classifier being the most effective. The remaining models indicate a need for further optimization or alternative model selection for improved results.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2.Results:

2.1 Accuracy

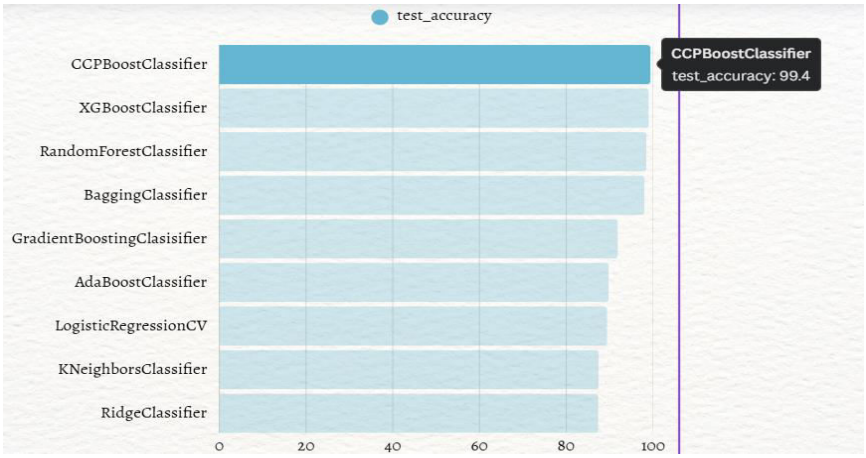


Fig8: Accuracy % for various Classifiers

Fig8 presents a horizontal bar chart comparing the test accuracy of various classification models, with CCPBoostClassifier exhibiting the highest accuracy at 99.4%, clearly outperforming the other models. The chart visually ranks the models, showing a significant drop in accuracy for models like GradientBoostingClassifier, AdaBoostClassifier, and LogisticRegressionCV, highlighting the superior performance of CCPBoostClassifier in this context. This visualization emphasizes the importance of model selection and the impact of different algorithms on predictive accuracy.

2.2 APP.py

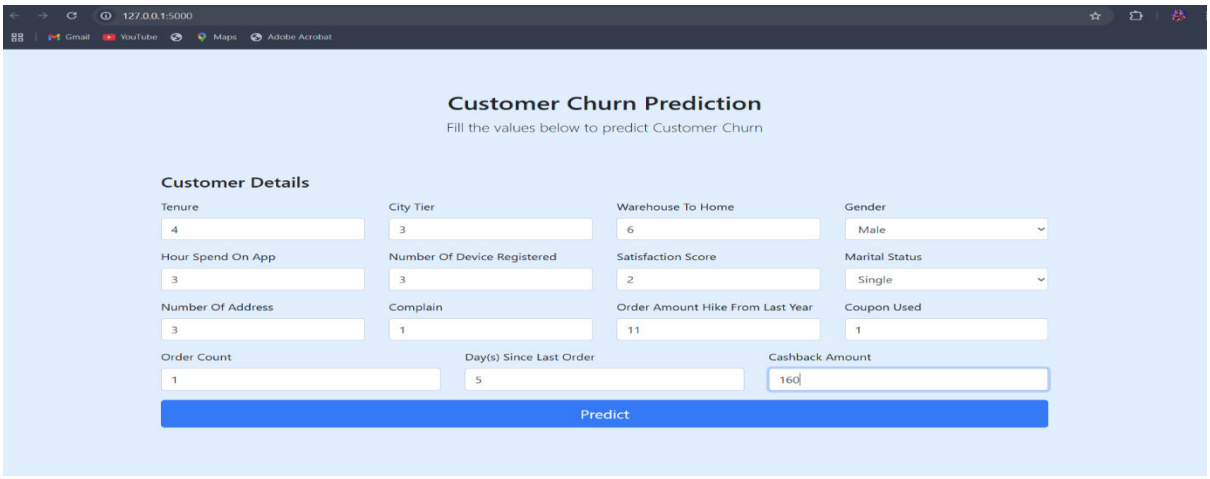


Fig9:Input page

Fig9 shows what i/p data we should collect and used for prediction.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304138

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

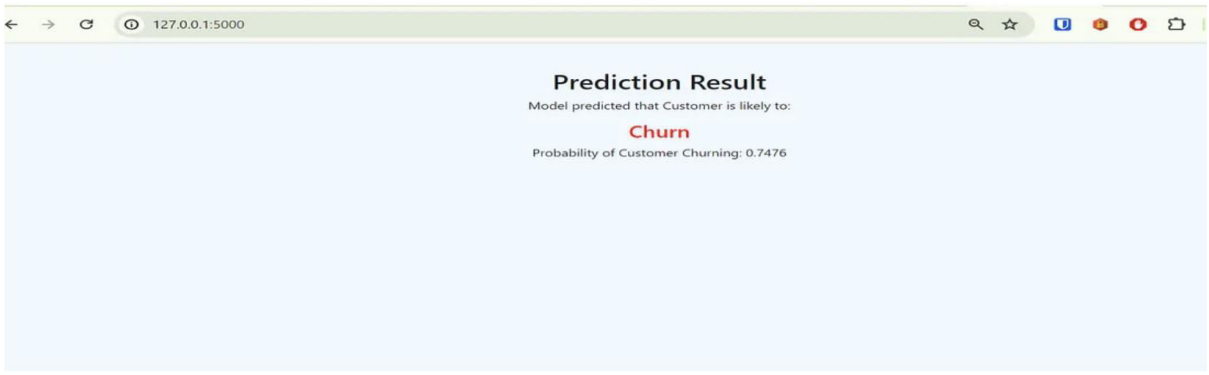


Fig10: Prediction Result

Fig10 Shows the predicted result of give data.

VI. CONCLUSION

This project demonstrates the power of data-driven churn prediction in enhancing customer retention for e-commerce businesses. By predicting churn with machine learning models, businesses can reduce customer attrition, improve retention rates, and ultimately increase revenue. However, there are opportunities for further research to refine these models and integrate them more seamlessly into real-time systems for even greater impact.The future of churn prediction lies in leveraging more granular and dynamic data, exploring advanced machine learning techniques, and ensuring that retention strategies evolve to meet the changing needs of customers. As e-commerce continues to grow, the ability to predict and prevent churn will remain a key differentiator for businesses seeking long-term success in a competitive landscape.

REFERENCES

1. Geiler, Louis, Séverine Affeldt, and Mohamed Nadif. "A survey on machine learning methods for churn prediction." International Journal of Data Science and Analytics 14.3 (2022): 217-242.
2.Lemmens, Aurélie, and Sunil Gupta. "Managing churn to maximize profits." Marketing Science 39.5 (2020): 956-973.
3. Al-Mashraie, Mohammed, Sung Hoon Chung, and Hyun Woo Jeon. "Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach." Computers & Industrial Engineering 144 (2020): 106476.
4. de Lima Lemos, Renato Alexandre, Thiago Christiano Silva, and Benjamin Miranda Tabak. "Propension to customer churn in a financial institution: a machine learning approach." Neural Computing and Applications 34.14 (2022): 11751-11768.
5. Dias, Joana, Pedro Godinho, and Pedro Torres. "Machine learning for customer churn prediction in retail banking." International Conference on Computational Science and Its Applications. Cham: Springer International Publishing, 2020.
6. Schaeffer, Satu Elisa, and Sara Veronica Rodriguez Sanchez. "Forecasting client retention—A machine-learning approach." Journal of Retailing and Consumer Services 52 (2020): 101918.
7. He, Benlan, et al. "Prediction of customer attrition of commercial banks based on SVM model." Procedia computer science 31 (2014): 423-430.
8. Karvana, Ketut Gde Manik, et al. "Customer churn analysis and prediction using data mining models in banking industry." 2019 international workshop on big data and information security (IWBIS). IEEE, 2019.
9. Lee, In, and Yong Jae Shin. "Machine learning for enterprises: Applications, algorithm selection, and challenges." Business Horizons 63.2 (2020): 157-170.
10. De Caigny, Arno, et al. "Incorporating textual information in customer churn prediction models based on a convolutional neural network." International Journal of Forecasting 36.4 (2020): 1563-1578.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details