UML501- Machine Learning Lab

E-commerce Customer Churn Prediction

UML 501 Machine Learning Project Report

Submitted by:

Ria Goyal (102203069)
Aarav Mahajan (102203020)

BE Third Year, COE

Group No: 3CO2

Submitted to:

Dr. Raman Goyal

November 2024

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

TIET, Patiala

1.Introduction

Customer churn prediction is a critical aspect of e-commerce businesses aiming to retain customers and minimize revenue loss. This project focuses on building a machine learning model to predict whether a customer will churn based on various features related to customer behavior, demographics, and purchasing patterns.

2. Data Overview

The dataset used for this project consists of customer-related features such as:

- **Demographic Attributes**: Gender, Marital Status, and City Tier.
- **Behavioral Attributes**: Preferred login device, number of devices registered, satisfaction score, and number of complaints.
- **Transaction Attributes**: Order count, order amount hike from last year, and cashback amount.

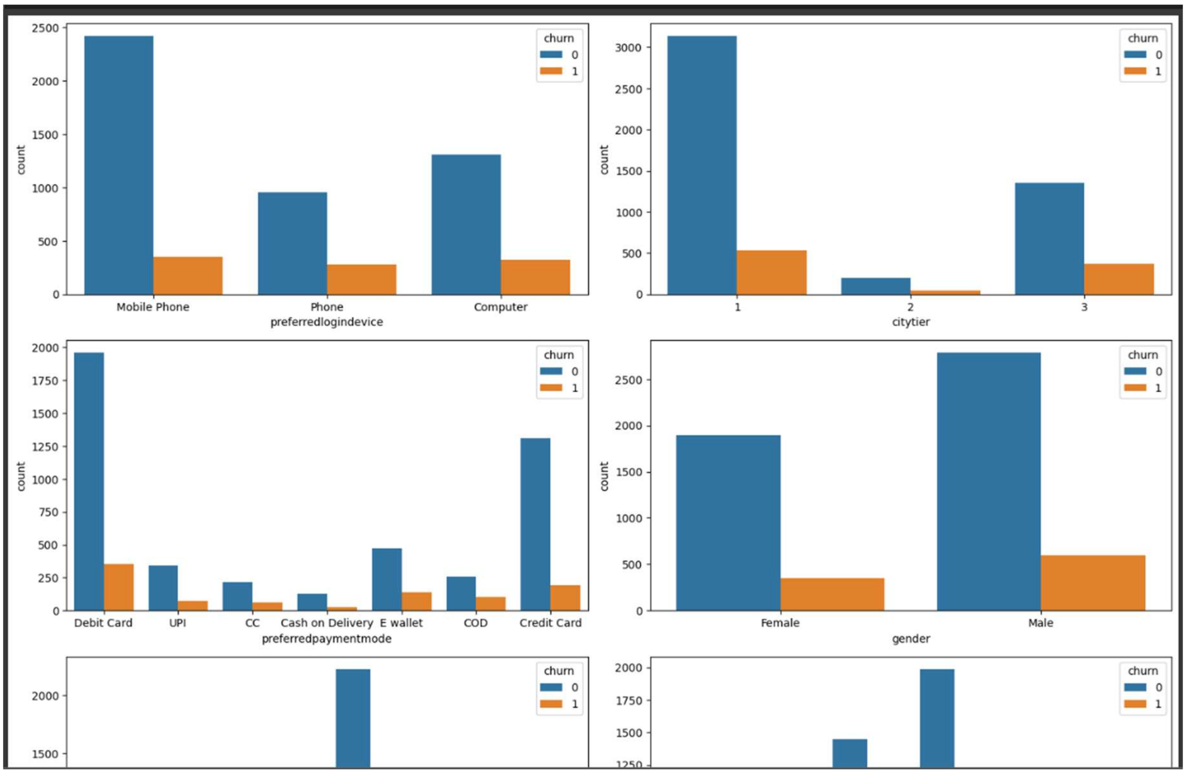

Fig: Plot the Churn distribution for each categorical variable

Missing Values Analysis: The dataset had missing values in several key columns, such as Tenure, HourSpendOnApp, and DaySinceLastOrder. Missing values were handled using the following imputation strategies:

- **Iterative Imputer**: For numerical columns like Tenure and OrderAmountHikeFromlastYear.
- **Simple Imputer**: For categorical columns using the most frequent value.

3. Data Preprocessing and Feature Engineering

Data preprocessing included:

- **Handling Missing Values**: Imputation techniques were applied to fill missing data.
- **Feature Scaling**: StandardScaler was used to normalize numerical features.
- **Encoding Categorical Variables**: One-Hot Encoding was applied to convert categorical features like PreferredPaymentMode and PreferedOrderCat into numerical format.

Feature Selection: Key features considered for model training included Tenure, HourSpendOnApp, OrderCount, and SatisfactionScore, which showed a strong correlation with the target variable (Churn).

4. Model Building

Three models were used to predict customer churn:

- **Logistic Regression**: A baseline model to understand the impact of each feature.
- **Random Forest Classifier**: An ensemble method to capture non-linear relationships in the data.
- **XGBoost Classifier**: A gradient boosting algorithm, which was tuned for optimal performance using GridSearchCV.

5. Model Evaluation

The models were evaluated using various metrics:

- **Accuracy**: Proportion of correct predictions out of total predictions made.
- **Confusion Matrix**: Visualization of true positives, true negatives, false positives, and false negatives.
- **Classification Report**: Detailed report showing precision, recall, and F1-score.

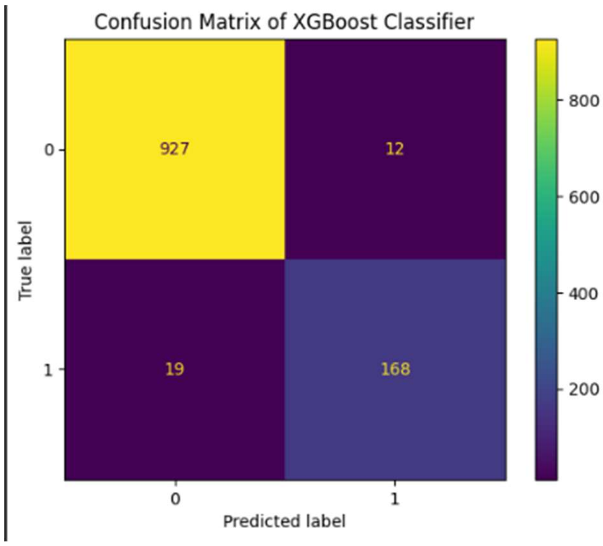

Fig: Confusion Matrix for XGBoost Classifier

Best Performing Model:

- The **XGBoost Classifier** outperformed the other models with the highest accuracy and F1-score, indicating its effectiveness in capturing complex patterns in customer data.

6. Results and Insights

- **Customer Tenure** and **Satisfaction Score** were strong indicators of churn. Customers with lower satisfaction scores and shorter tenure were more likely to churn.
- **Preferred Order Category** and **Preferred Payment Mode** also had significant predictive power, suggesting that customers with specific preferences had varying churn rates.
-

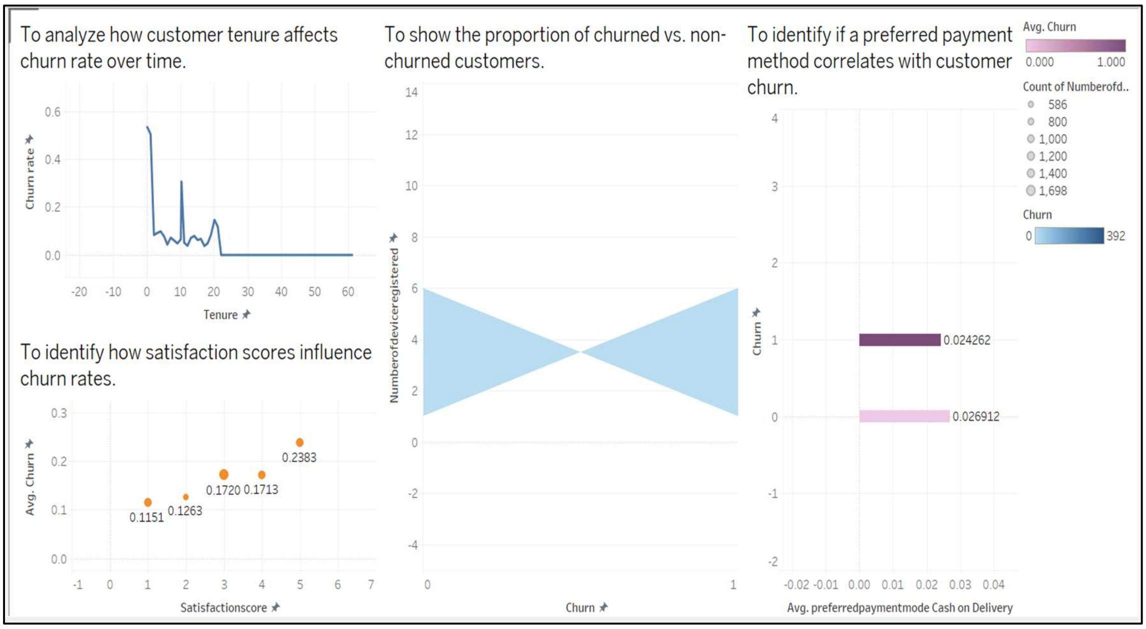

Fig: Tableau Dashboard for visualization

	model_name	test_accuracy	test_precision	test_recall	test_f1
1	XGBClassifier	99.300000	98.200000	97.600000	97.900000
2	RandomForestClassifier	98.600000	99.100000	92.400000	95.600000
0	LogisticRegressionCV	89.300000	77.800000	51.100000	61.600000

Fig: This shows the accuracy of various models used

7. Conclusion

This project demonstrated the effectiveness of using machine learning models, particularly the XGBoost Classifier, in predicting customer churn in the e-commerce sector. By identifying customers at risk of churning, businesses can implement targeted retention strategies to enhance customer satisfaction and reduce churn rates.

8. Future Work

Future enhancements could include:

- Incorporating additional features like customer reviews or social media engagement.
- Using more advanced deep learning models for further improvements in predictive accuracy.
- Conducting A/B testing to validate the impact of targeted interventions on reducing churn.

Appendix

The dataset used for this analysis can be accessed https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction/data.