E-commerce Customers Churn Prediction Using Machine Learning

Shahmi A.J.A
Department of Electrical and Information Engineering
University of Ruhuna
Galle, Sri-Lanka
shahmiahamed0519@gmail.com

***Abstract*—Customer churn prediction is a crucial aspect of the e-commerce industry, helping businesses retain customers and maximize revenue. This study implements and evaluates various supervised machine learning models to predict customer churn based on behavioral and transactional data. The dataset consists of customer demographics, purchase history, and payment behavior. Three models-Logistic Regression, Random Forest, and Gradient Boosting-were evaluated for performance using accuracy, precision, recall, and F1-score. Results indicate that the Random Forest model outperformed the others, achieving the highest F1-score. This study highlights the effectiveness of machine learning in churn prediction and suggests strategies to mitigate customer attrition.**

Keywords—E-commerce, Customer Churn, Machine Learning, Random Forest, Gradient Boosting, Logistic Regression.

I. Introduction

Customer churn refers to the loss of customers who discontinue their relationship with a business. Predicting and reducing churn is critical for e-commerce companies, as acquiring new customers is more expensive than retaining existing ones.

Traditional methods rely on statistical analysis, but machine learning offers a more efficient way to identify churn patterns from large datasets.

The objective of this research is to develop a churn prediction model using various machine learning algorithms and evaluate their performance to determine the most effective approach.

II. Literature Review

Several studies have explored machine learning techniques for churn prediction across different industries. Previous research has demonstrated that ensemble methods, such as Random Forest and Gradient Boosting, often outperform traditional regression-based models due to their ability to capture complex interactions between features. Moreover, handling imbalanced datasets using techniques like **SMOTE** or cost-sensitive learning has proven to improve recall scores, which is crucial for identifying potential churners effectively.

III. Methodology

A. Dataset Description

The dataset used in this study is a synthesis "E-commerce Customer Behavior and Purchase Dataset" containing customer demographics, purchase transactions, and churn labels. The key features include:

- **Customer Age, Gender –** Demographic attributes.
- **Purchase Date, Product Category, and Total Purchase Amount –** Transaction history.
- **Payment Method, Returns –** Behavioral factors.
- **Churn –** Target variable (0=retained, 1=churned).

B. Data Pre-processing

To prepare the dataset for modeling, the following preprocessing steps were performed:

- **Handling Missing Values –** Imputed using the mode for categorical variables.
- **Feature Engineering –** Categorical variables (e.g., payment method, product category) were converted to numerical form using one-hot encoding.
- **Feature Scaling –** Standardization was applied to numerical features such as total purchase amount.
- **Handling Imbalanced Data –** The dataset had an imbalance in churned vs retained customers. To address this, the **SMOTE (Synthetic Minority Oversampling Technique)** was applied to balance the class distribution.

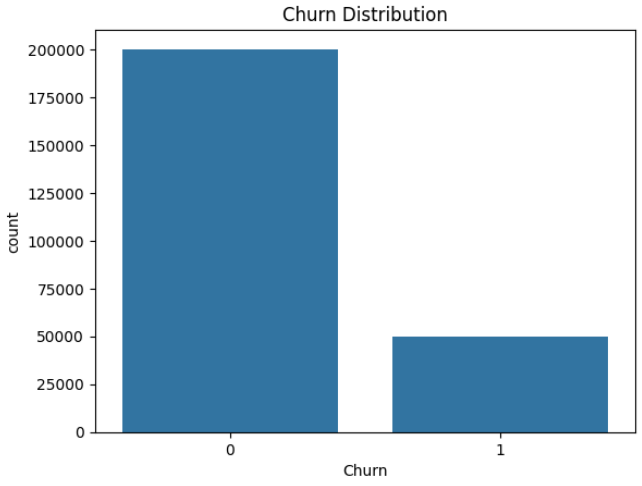

Fig. 1: Imbalanced Data of Churn Distribution

C. *Model Selection and Training*

Three machine learning models were trained and evaluated:

- **Logistic Regression –** A baseline model for classification problems.
- **Random Forest –** An ensemble method known for handling complex data patterns.
- **Gradient Boosting –** A boosting algorithm that optimizes model performance by reducing errors iteratively.

Each model was trained on 70% of the dataset, with 15% used for validation and 15% for testing. **K-fold cross-validation** was performed to ensure model generalization. Additionally, hyper-parameters tuning was conducted using **GridSearchCV** to optimize model parameters for better performance.

IV. Results and Discussion

The models were evaluated using four key metrics: **Accuracy, Precision, Recall, and F1-Score**. The results are summarized below:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.78	0.72	0.65	0.68
Random Forest	0.80	0.75	0.70	0.72
Gradient Boosting	0.79	0.74	0.68	0.71

From the results, the **Random Forest model** performed the best, achieving the highest accuracy and F1-score. Feature importance analysis showed that **total purchase amount, product category, and payment method** were the most influential factors in predicting churn, Furthermore, hyper-parameter tuning improved model performance by fine-tuning the number of estimators, tree depth, and minimum samples per split.

A comparative analysis with previous studies indicate that Random Forest remains a robust choice for prediction due to its interpretability and ability to capture complex relationships in the data.

V. Conclusion and Future Work

This study demonstrates the power of machine learning in predicting e-commerce customer churn. Among the models tested, the **Random Forest model proved to be the most effective,** balancing accuracy, precision, recall, and F1-score.

Future work can explore the use of deep learning techniques, such as neural networks, to further enhance prediction accuracy. Additionally, integrating real-time churn prediction into business decision making systems could provide immediate insights for customer retention strategies.

Furthermore, incorporating customer engagement metrics, such as website activity and time spent on platform, could enhance predictive power.

References

[1] Kaggle Dataset: E-commerce Customer Behavior and Purchase Dataset.
E-commerce Customer Data For Behavior Analysis

[2] Breiman, L. "Random Forests." Machine Learning, 2001.
Random Forests | Machine Learning

[3] Bishop, C. "Pattern Recognition and Machine Learning." Springer, 2006.
Pattern Recognition and Machine Learning | SpringerLink