

# Customer Churn Prediction Using Machine Learning



**Author:** Mostafa Mohamed Mostafa El-Gallad

**Institution:** Faculty of Computer & Information Sciences - Mansoura University

**Date:** 2024

**Supervisor:** Dr Haitham El-Gharib

## Table of Contents

1. **Introduction**
2. **Problem Statement**
3. **Objective**
4. **Literature Review**
5. **Dataset Description**
6. **Exploratory Data Analysis (EDA)**
7. **Methodology**
8. **Model Selection and Evaluation**
9. **Application Development**

10. **Conclusion**
11. **Future Work**
12. **References**

## 1. Introduction

Customer churn is a critical issue for businesses, especially those with subscription-based services or recurring customers. Predicting churn helps organizations understand customer behavior and implement effective retention strategies. This project explores the use of machine learning to build a customer churn prediction model that forecasts the likelihood of customers discontinuing their service.

## 2. Problem Statement

High customer churn can lead to significant revenue loss, making it vital for businesses to predict churn early. This project aims to develop a machine learning model to accurately predict customer churn and provide actionable insights to reduce it.

## 3. Objective

The objective of this project is to build a machine learning model that can predict whether a customer will churn based on historical data. The model will help businesses:

- Identify at-risk customers
- Implement retention strategies
- Increase customer satisfaction and loyalty

## 4. Literature Review

Several studies have demonstrated the importance of predicting customer churn using machine learning techniques. Key research areas include:

- **Churn prediction models:** Previous work has focused on decision trees, random forests, and logistic regression.
- **Feature engineering:** Effective features such as customer tenure, order frequency, and satisfaction scores have proven to enhance model accuracy.
- **Oversampling techniques:** Handling class imbalances with methods like SMOTE improves model performance for minority classes.

## 5. Dataset Description

The dataset used for this project is sourced from an e-commerce company. It contains historical data on customer behavior, including:

- **CustomerID:** Unique identifier
- **Tenure:** Length of time as a customer
- **PreferredLoginDevice:** Device used to access services
- **OrderCount:** Number of orders placed
- **SatisfactionScore:** Satisfaction rating provided by the customer
- **Churn:** Whether the customer churned (1 = churned, 0 = retained)

The dataset comprises 5,000 records with 18 features.

## 6. Exploratory Data Analysis (EDA)

Key insights from the EDA:

- **Churn rate:** 16.8% of customers churned, while 83.2% were retained.
- **Device usage:** Most customers prefer logging in via mobile devices.
- **Payment preferences:** Debit and credit cards are the most common payment methods.
- **City tiers:** Tier 2 cities have the lowest customer count, indicating potential opportunities for targeted campaigns.

## 7. Methodology

### Data Preprocessing

- **Handling Missing Data:** Used Random Forest Iterative Imputer for missing value imputation.
- **Class Imbalance:** The target variable 'Churn' was imbalanced. Applied Synthetic Minority Over-sampling Technique (SMOTE) to balance the data.
- **Feature Encoding:** Categorical features were converted using one-hot encoding.

### Modeling

- **Training and Testing Split:** The data was split into 80% training and 20% testing sets.
- **Oversampling:** SMOTE was applied to the training set to address class imbalance.

## 8. Model Selection and Evaluation

Several machine learning models were evaluated, including:

- **Logistic Regression**
- **Random Forest Classifier**
- **XGBoost Classifier**

The XGBoost model outperformed others with the highest accuracy and F1 score. Feature importance from the model revealed that the most significant factors contributing to churn

prediction were:

- **Tenure**
- **Satisfaction Score**
- **Order Count**
- **Cashback Amount**

Model accuracy on the test set: **92.4%**F1 Score: **0.89**

## 9. Application Development

A **Flask-based web application** was developed to allow businesses to predict customer churn by inputting customer data. The app features:

- **Input Interface:** For entering customer data (e.g., tenure, order count).
- **Output Interface:** Displays churn prediction results (churn/retain).
- **Model Deployment:** The trained XGBoost model is used for predictions.

## 10. Conclusion

The customer churn prediction model provides actionable insights that help businesses reduce churn and increase customer retention. The project demonstrated the effectiveness of machine learning techniques in analyzing customer data and predicting churn.

## 11. Future Work

To further enhance the project, the following improvements are suggested:

- **Feature Engineering:** Adding more features such as customer interaction history, browsing data, and social media engagement.
- **Model Tuning:** Fine-tuning the hyperparameters of the XGBoost model for improved accuracy.
- **Real-time Data:** Deploying the model to predict churn using real-time data streams.

## 12. References

1. Ouko, Allan. "Customer Churn Prediction Using Machine Learning." Medium. [Link](#)
2. Kaggle Dataset. [Link](#)
3. Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, 2019.
4. Frey, Kirill. [Building Recommender Systems with Machine Learning and AI](#)  
[Full Project Link in Github](#)