Customer Churn Prediction Report Using Machine Learning

1. Introduction

Customer churn prediction is an essential task for businesses aiming to retain customers and reduce revenue losses. By predicting which customers are likely to leave, companies can implement targeted retention strategies to improve loyalty and increase profitability. Machine learning provides an effective method for analysing vast amounts of customer data and identifying patterns that indicate the likelihood of churn. This report outlines a customer churn prediction model using machine learning, implemented through the application of various algorithms, data preprocessing, and evaluation metrics.

2. Data Overview

The dataset used in this analysis is sourced from an e-commerce platform, containing customer information such as demographics, product usage, and churn status. The data was loaded and initially analysed to understand its structure. The following steps were taken to clean and preprocess the data before applying machine learning models:

- *Missing Values:* The dataset contained missing values, which were handled using imputation techniques. For numeric columns, the mean value was used for imputation, while for categorical columns, the most frequent value was used.
- *Duplicates:* Duplicate records were identified and removed to ensure the data's integrity.
- *Feature Encoding:* Categorical variables were encoded using one-hot encoding, transforming them into numerical features suitable for machine learning algorithms.
- *Data Scaling:* Numerical features were scaled using standardization (z-score normalization) to ensure that the model training is not biased by varying feature scales.

3. Data Preprocessing

Several important preprocessing steps were undertaken to prepare the data for machine learning modelling:

- *Missing Data Handling:* Numerical columns with missing values were filled with the mean, while categorical columns were filled with the most frequent value.
- *Categorical Encoding:* The categorical variables were converted into dummy variables using one-hot encoding, dropping the first column to avoid multicollinearity.
- *Scaling:* The numerical features were standardized to ensure all features are on the same scale, which helps improve the performance of many machine learning models.
- *Handling Imbalanced Data:* Since churned customers (those who left) are typically fewer than non-churned customers, the synthetic minority oversampling technique (SMOTE) was applied to balance the dataset by generating synthetic samples for the minority class.

4. Model Selection and Training

Three machine learning models were employed to predict customer churn:

- *Gradient Boosting Classifier:* A powerful ensemble method that builds multiple decision trees in a sequential manner, optimizing each tree to correct the errors of the previous ones.
- *Random Forest Classifier:* An ensemble method that constructs multiple decision trees and combines their predictions, reducing overfitting and improving model accuracy.
- *XGBoost Classifier:* An advanced implementation of gradient boosting, known for its speed and performance, often yielding higher accuracy in classification tasks.

The models were trained using the pre-processed training data, and their performance was evaluated on the test data.

5. Model Evaluation

The performance of each model was evaluated using the following metrics:

- *Accuracy:* The proportion of correctly classified instances.
- *Classification Report:* A detailed report showing precision, recall, and F1-score for both churned and non-churned classes.
- *ROC-AUC Score:* The area under the receiver operating characteristic curve, which indicates the model's ability to discriminate between churned and non-churned customers.
- *Confusion Matrix:* A matrix showing the true positives, true negatives, false positives, and false negatives, providing a clear view of the model's classification performance.
- *ROC Curve:* The plot of the false positive rate versus the true positive rate, used to evaluate the model's performance across various classification thresholds.

6. Results and Visualizations

The following results were observed for each model:

- *Gradient Boosting:* This model performed well, demonstrating a strong ability to predict customer churn with a high ROC-AUC score and good classification metrics.
- *Random Forest:* Random Forest showed comparable performance to Gradient Boosting, achieving a high accuracy and AUC score, with an interpretable confusion matrix.
- *XGBoost*: XGBoost provided the best performance in terms of accuracy and ROC-AUC score, effectively capturing complex patterns in the data.

Visualizations:

- *Churn Distribution:* Count plots were created to visualize the distribution of churn across different categorical features, revealing insights about factors contributing to churn.
- *Feature Histograms:* Histograms of numerical features were plotted to visualize their distributions, highlighting any skewness or outliers in the data.
- *Confusion Matrix:* Heatmaps of confusion matrices were displayed to assess the true and false positive rates for each model.

- *ROC Curves:* The ROC curves for each model were plotted, showing how well each model distinguishes between churned and non-churned customers.

7. Conclusion

The analysis demonstrates that machine learning techniques, particularly ensemble methods like Gradient Boosting, Random Forest, and XGBoost, can be highly effective in predicting customer churn. The models provided valuable insights into customer behaviour, enabling the identification of high-risk customers who are likely to churn.

By leveraging these predictions, businesses can implement targeted retention strategies, such as personalized offers or improved customer service, to enhance customer loyalty and reduce churn rates.

Future improvements could include further tuning of the hyperparameters of the models, exploring additional features, and incorporating more advanced techniques like deep learning to improve accuracy further. Additionally, real-time churn prediction models could be developed to allow businesses to take immediate action based on the predictions.