# **CUSTOMER CHURN PREDICTION**

## **1. Introduction**

Customer churn prediction is crucial for businesses to retain customers by identifying those at risk of leaving. This report details the machine learning process used to build a churn prediction model, analyzing the dataset, training the model, evaluating its performance, and visualizing the results.

## **2. Data Loading and Preprocessing**

### **Step 1: Importing Required Libraries**

The necessary libraries, such as pandas, numpy, sklearn, and matplotlib, were imported to handle data processing, model training, evaluation, and visualization.

### **Step 2: Loading the Built-in Dataset**

A built-in dataset was used for training and testing. This dataset contained various customer attributes, including demographic details, account information, and usage patterns, which were essential for predicting churn.

### **Step 3: Handling Missing Values and Encoding Categorical Data**

* Missing values were checked and handled using appropriate imputation techniques.
* Categorical variables were encoded using techniques such as **one-hot encoding** to convert them into numerical format.
* The dataset was then split into features (**X**) and target labels (**y**), where 0 represented non-churn customers and 1 represented churn customers.

## **3. Model Training and Evaluation**

### **Step 4: Splitting Data into Training and Testing Sets**

The dataset was divided into **training (80%)** and **testing (20%)** sets to ensure the model was trained on a majority of the data and tested on unseen data for evaluation.

### **Step 5: Choosing a Machine Learning Model**

The **Random Forest Classifier** was chosen for this task due to its ability to handle complex relationships and feature interactions while reducing overfitting.

### **Step 6: Model Training**

The model was trained using the training dataset, learning patterns and relationships between the input features and the target variable (churn status).

### **Step 7: Predictions on Test Data**

Once trained, the model was used to make predictions on the test dataset to evaluate its effectiveness in identifying customer churn.

## **4. Model Performance Evaluation**

### **Step 8: Generating the Confusion Matrix**

A **confusion matrix** was created to assess the model's performance, displaying the counts of correct and incorrect predictions:

|  |  |  |
| --- | --- | --- |
| **Actual / Predicted** | **Predicted 0 (Not Churn)** | **Predicted 1 (Churn)** |
| **Actual 0 (Not Churn)** | 9 | 4 |
| **Actual 1 (Churn)** | 3 | 21 |

### **Step 9: Analyzing the Confusion Matrix**

* **True Positives (TP):** 21 (Correctly predicted churn cases)
* **True Negatives (TN):** 9 (Correctly predicted non-churn cases)
* **False Positives (FP):** 4 (Incorrectly predicted churn cases that were actually non-churn)
* **False Negatives (FN):** 3 (Incorrectly predicted non-churn cases that were actually churn)

### **Step 10: Key Performance Metrics**

* **Accuracy:** 81% – The overall correctness of the model.
* **Precision:** 84% – The proportion of actual churn cases among all predicted churn cases.
* **Recall:** 87.5% – The proportion of correctly predicted churn cases out of all actual churn cases.
* **F1-score:** 85.6% – A balance between precision and recall, indicating strong overall model performance.

## **5. Visualization and Insights**

### **Step 11: Confusion Matrix Plot**

A **heatmap of the confusion matrix** was plotted using seaborn to visually represent correct and incorrect predictions. Darker colors indicate higher values, showing that most predictions were correct.

### **Step 12: Feature Importance Plot**

The **feature importance plot** was generated to show which customer attributes contributed the most to churn prediction. This helps businesses focus on key factors influencing churn.

### **Step 13: Additional Graphs for Performance Analysis**

* **ROC Curve & AUC Score** were plotted to evaluate how well the model differentiates between churn and non-churn customers.
* **Precision-Recall Curve** was used to understand the trade-off between precision and recall, especially in handling imbalanced data.

## **6. Insights and Recommendations**

### **Step 14: Key Observations**

* The model performed well in identifying churn customers, with high recall and accuracy.
* Some misclassifications (false positives and false negatives) were observed, which could be minimized with further improvements.
* Feature importance analysis revealed that customer engagement metrics had the most significant impact on churn prediction.

### **Step 15: Potential Improvements**

* **Hyperparameter Tuning:** Further optimization using Grid Search or Random Search can improve performance.
* **Feature Engineering:** Adding more relevant features, such as transaction history or support interactions, can enhance predictions.
* **Balancing the Dataset:** If churn cases are underrepresented, techniques like **SMOTE (Synthetic Minority Over-sampling Technique)** can help balance the dataset.
* **Threshold Optimization:** Adjusting the classification threshold can help minimize false positives or false negatives, depending on business needs.

## **7. Conclusion**

This customer churn prediction model effectively identifies customers at risk of leaving with an **81% accuracy and 87.5% recall**. The model provides valuable insights into key factors influencing churn, allowing businesses to take proactive measures for customer retention. Further improvements through hyperparameter tuning, feature engineering, and dataset balancing could enhance its reliability.