Customer Churn Prediction in the Telecom Industry

A Machine Learning Approach to Service Enhancement in Telecom Companies

Estarta AI & DS training

Nashat Alfarajat

Table of Contents

Page 1 – Introduction and Data Overview

- Introduction
- About the Data

Page 2 – Data Quality and Outliers

- Checking for Missing Values
- Duplicate Checks
- Handling Outliers

Page 3 – Numerical Features and Categorical Encoding

- Numerical Feature Correlation Analysis
- Encoding Categorical and Target Features
- Feature Importance Analysis
    - Chi-Square Test (State)
    - Information Gain
- Data Preparation for Modeling

Page 4 – Modeling: Logistic Regression

- Logistic Regression Model
- Classification Results

Page 5 – Modeling: Decision Tree and XGBoost

- Decision Tree Model
- XGBoost Model
- Classification Results

Page 6 – XGBoost Analysis

- Feature Importance for XGBoost
- ROC-AUC for XGBoost

Introduction

In today's highly competitive telecommunications market, retaining existing customers is just as important—if not more—than acquiring new ones. Customer churn, the phenomenon where subscribers discontinue their service and move to a competitor, poses a serious challenge for telecom companies. High churn rates can lead to revenue losses, increased marketing costs, and reduced customer loyalty.

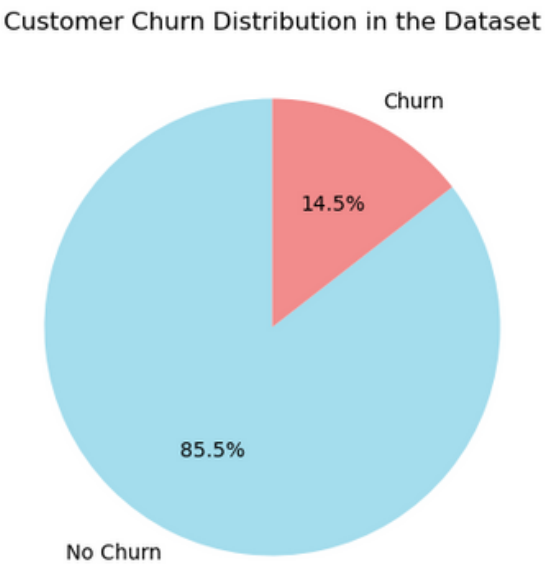
Machine learning provides a powerful solution to this challenge by enabling companies to predict customer churn before it happens. By analyzing customer behavior and usage of network services, predictive models can identify which subscribers are at a higher risk of leaving. This allows telecom providers to take proactive measures—such as offering tailored service plans, improving customer support, or addressing billing issues—before customers make the decision to switch.

In this report, I present a machine learning model designed to predict customer churn based on patterns of customer interaction with telecom services. The insights generated from such models can help telecom companies improve decision-making, design targeted retention strategies, and ultimately enhance the quality of their services.

About the Data

This public dataset is provided by the CrowdAnalytix community as part of their churn prediction competition. The real name of the telecom company is anonymized to protect confidentiality. It contains 20 predictor variables, mostly related to customer usage patterns.
The dataset includes 3,333 customer records, of which 483 customers are churners and 2,850 are non-churners, resulting in a churn ratio of approximately 14%.

```
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 20 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   State                   3333 non-null   object
 1   Account length          3333 non-null   int64
 2   Area code               3333 non-null   int64
 3   International plan      3333 non-null   object
 4   Voice mail plan         3333 non-null   object
 5   Number vmail messages   3333 non-null   int64
 6   Total day minutes       3333 non-null   float64
 7   Total day calls         3333 non-null   int64
 8   Total day charge        3333 non-null   float64
 9   Total eve minutes       3333 non-null   float64
 10  Total eve calls         3333 non-null   int64
 11  Total eve charge        3333 non-null   float64
 12  Total night minutes     3333 non-null   float64
 13  Total night calls       3333 non-null   int64
 14  Total night charge      3333 non-null   float64
 15  Total intl minutes      3333 non-null   float64
 16  Total intl calls        3333 non-null   int64
 17  Total intl charge       3333 non-null   float64
 18  Customer service calls  3333 non-null   int64
 19  Churn                   3333 non-null   bool
dtypes: bool(1), float64(8), int64(8), object(3)
memory usage: 498.1+ KB
```

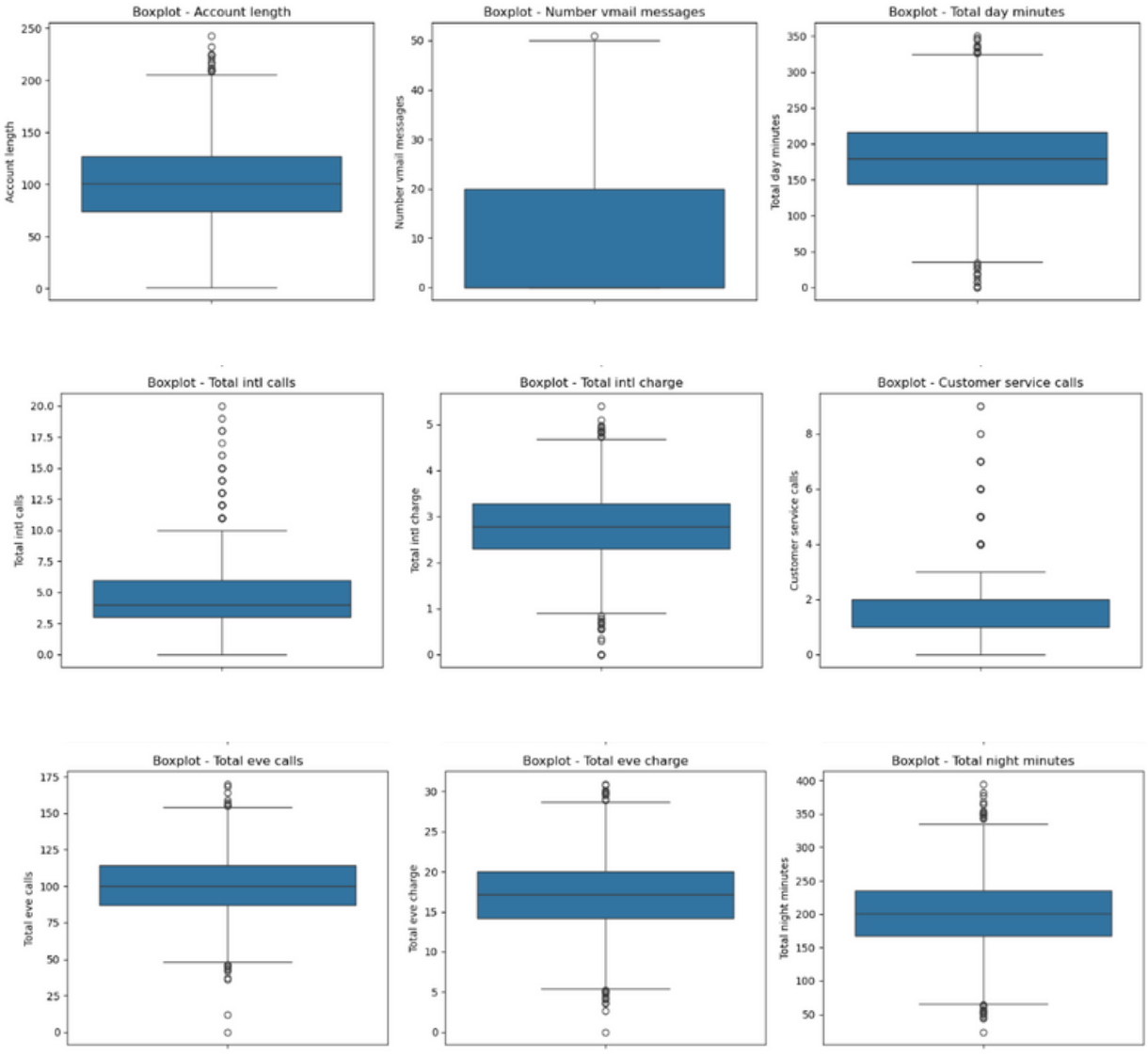


The dataset's structure provides a solid foundation to analyze customer behavior and develop machine learning models to predict churn effectively.

Data Quality and Outlier Analysis

Part 1 – Data Quality:
The dataset was carefully examined and found to be complete and consistent. There are no missing values among the 20 features, and no duplicate records were detected. This ensures that the data is ready for analysis and machine learning modeling without additional cleaning.

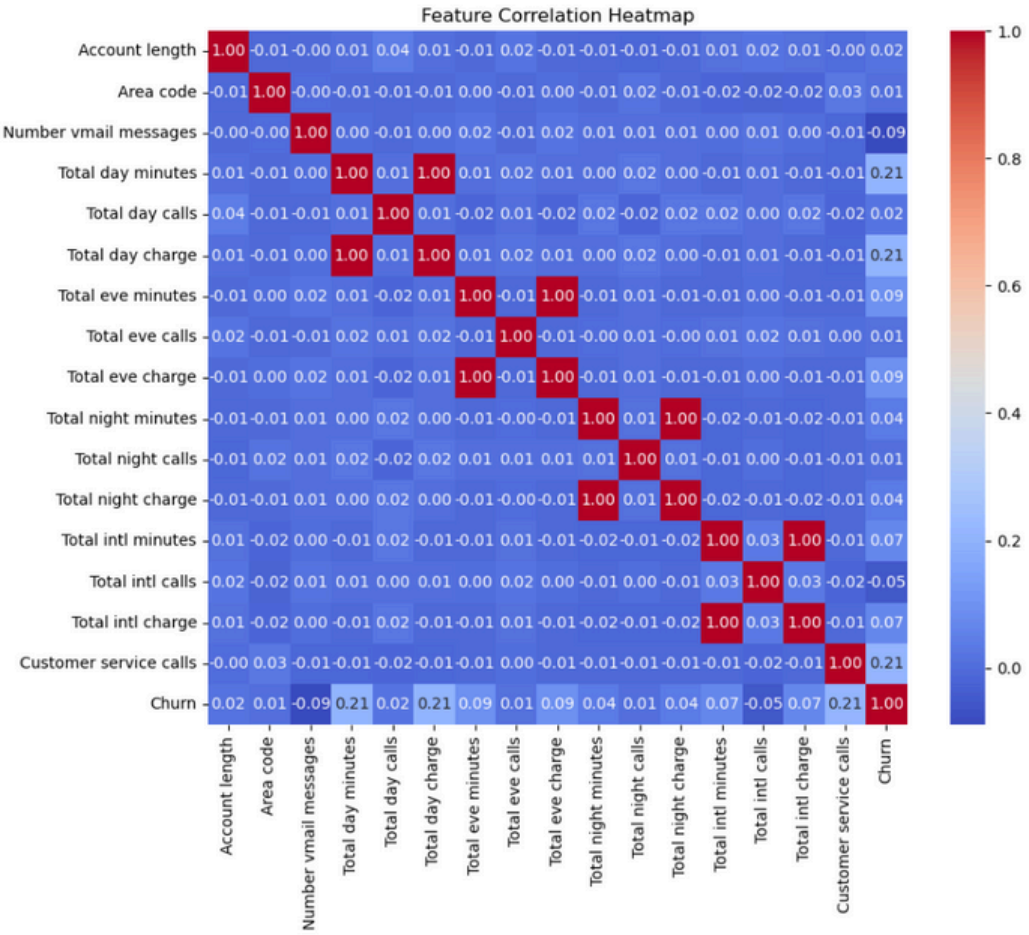
Part 2 – Outlier Analysis:
Numerical features were inspected for outliers using box plots. After review, all outliers were retained, as they represent valid customer behavior patterns. Removing them could have distorted the data and negatively affected the model's ability to learn from real-world usage variations.



Numerical Feature Correlation Analysis

The relationships between all numerical features were examined using a correlation heat map. This visualization helps identify which numerical variables are strongly correlated with one another. Strong correlations may indicate redundancy, while low correlations suggest that features provide independent information for the model.

Analyzing these correlations is important for feature selection and understanding how different usage patterns and charges relate to each other, which can improve the performance and interpretability of the churn prediction model.



To simplify the model and avoid redundancy, I dropped one feature from each pair of strongly correlated numerical variables. Specifically, the following features were removed: Total day minutes, Total night minutes, Total eve minutes, and Total intl minutes. This step helps reduce multicollinearity while preserving the essential information needed for accurate churn prediction.

Encoding Categorical and Target Features

The categorical features International plan and Voice mail plan were encoded into numerical format to make them compatible with machine learning models. Specifically, “Yes” was mapped to 1 and “No” to 0.

Additionally, the target variable Churn was converted from a boolean to an integer, with churners represented as 1 and non-churners as 0. This encoding ensures that all features are in a suitable format for model training.

Feature Importance Analysis

1. Categorical Feature – Chi-square Test:

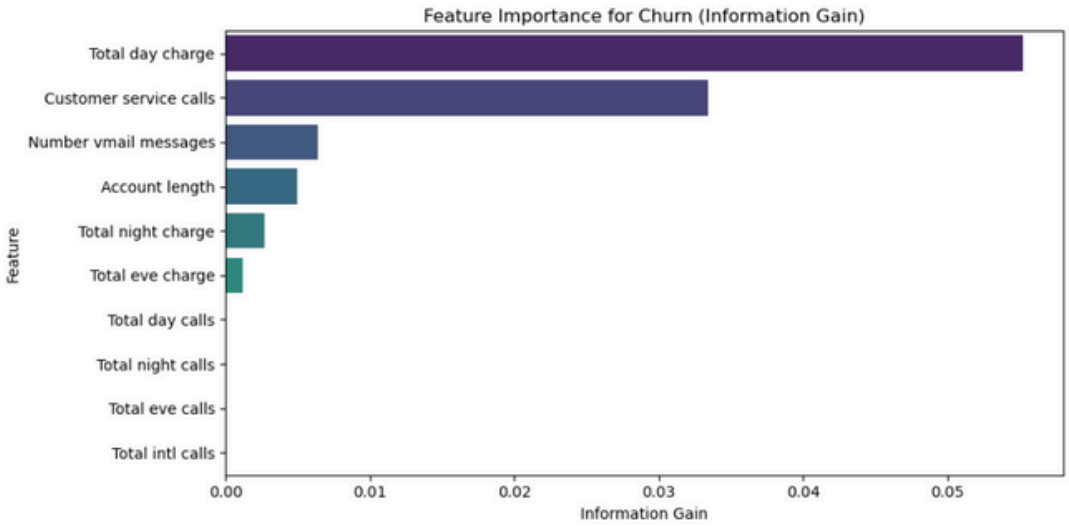
The categorical feature State was analyzed using a Chi-square test to measure its association with churn. A contingency table was created between State and Churn, and the test produced the following results:

- Chi-square statistic: 83.0438
- p-value: 0.0023
-

Since the p-value is less than 0.05, there is statistically significant evidence that a customer's state is associated with churn. This indicates that State is an important feature for predicting churn.

2. Numerical Features – Information Gain:

For the numerical features, information gain (or mutual information) was computed with respect to the target variable Churn. This measures how much knowing the value of a feature reduces uncertainty about churn. Features with higher information gain contribute more predictive power to the model.
Combining these analyses allows for better feature selection, ensuring that the most relevant variables—both categorical and numerical—are used in model training.



Data Preparation for Modeling

1. Train-Test Split: The dataset was split into training and testing sets using an 80/20 ratio.

2. Target Encoding for State: The categorical feature State was encoded based on the target variable (churn). Each state was assigned a value representing the proportion of churners in that state, allowing the model to capture the relationship between state and churn in numerical form.
This encoding was performed after the train-test split to prevent data leakage. If target encoding were applied before splitting, information from the test set could influence the encoding values in the training set, artificially inflating model performance. By encoding after splitting, we ensure that the model only has access to information from the training set during learning.

3. Standard Scaling of Numerical Features: All numerical features were standardized to have a mean of 0 and a standard deviation of 1. This ensures that features with larger scales do not dominate the model training and improves convergence and performance for many machine learning algorithms.

Modeling

Three machine learning models were trained to predict customer churn: Logistic Regression, Decision Tree, and XGBoost.
Since the dataset is imbalanced, model performance was evaluated using the confusion matrix and its associated metrics, including precision, recall, and F1-score, rather than relying solely on accuracy. This approach ensures that the models are assessed on their ability to correctly identify churners as well as non-churners.

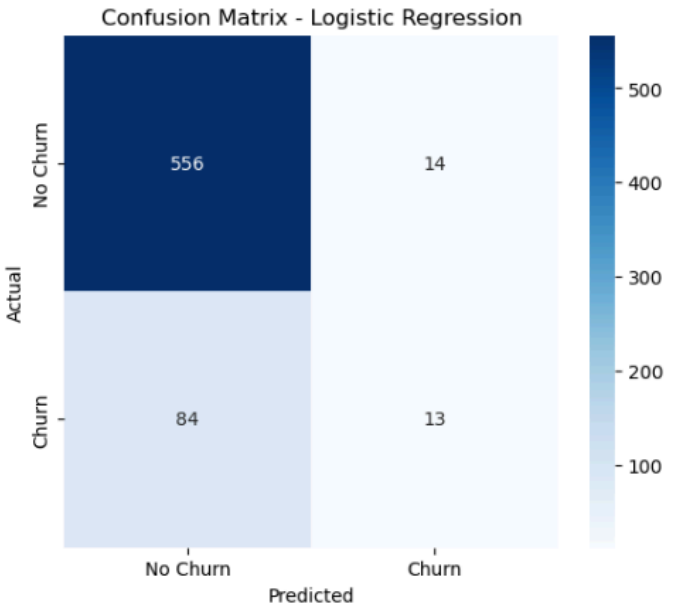
Logistic Regression Results

The Logistic Regression model achieved an accuracy of 85.31% on the test set. However, due to the imbalanced dataset, additional metrics from the confusion matrix provide more insight:

Accuracy: 0.8531

Classification Report:

	precision	recall	f1-score	support
No Churn	0.87	0.98	0.92	570
Churn	0.48	0.13	0.21	97
accuracy			0.85	667
macro avg	0.68	0.55	0.56	667
weighted avg	0.81	0.85	0.82	667



- Macro average: Precision 0.68, Recall 0.55, F1-score 0.56
- Weighted average: Precision 0.81, Recall 0.85, F1-score 0.82

Interpretation: While the model performs well in identifying non-churners, it struggles to correctly predict churners, highlighting the impact of data imbalance.

Decision Tree Results

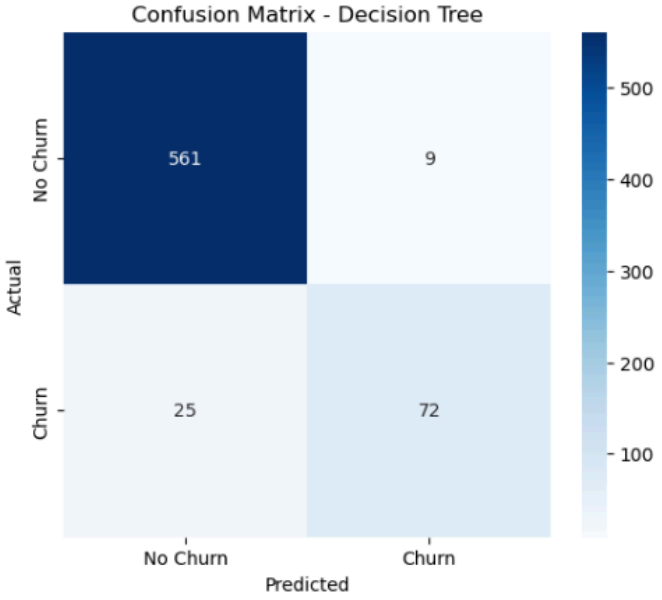
The Decision Tree model achieved an accuracy of 94.90% on the test set. Performance metrics from the confusion matrix are:

```
Accuracy: 0.9490

Classification Report:
              precision    recall  f1-score   support

    No Churn       0.96      0.98      0.97       570
       Churn       0.89      0.74      0.81        97

    accuracy                           0.95       667
   macro avg       0.92      0.86      0.89       667
weighted avg       0.95      0.95      0.95       667
```



- Macro average: Precision 0.92, Recall 0.86, F1-score 0.89
- Weighted average: Precision 0.95, Recall 0.95, F1-score 0.95

Interpretation: The Decision Tree performs very well, correctly identifying most churners and non-churners, and significantly outperforms Logistic Regression in handling the imbalanced dataset.

XGBoost Results

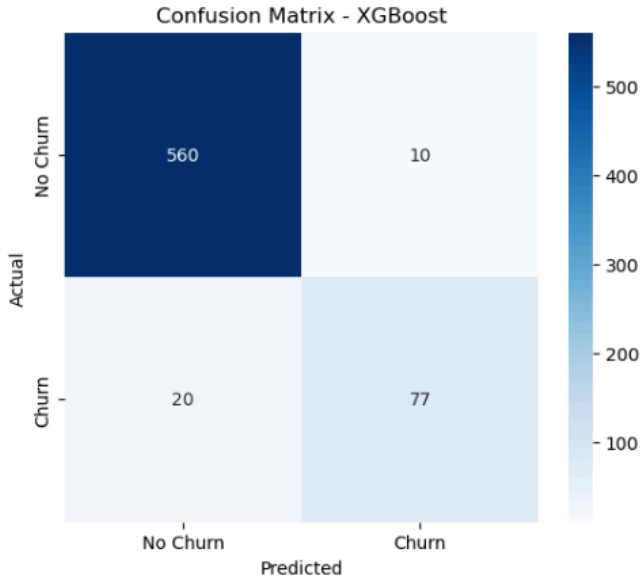
The XGBoost model achieved the highest performance among the three models, with an accuracy of 95.50% on the test set. The confusion matrix metrics are:

```
Accuracy: 0.9550

Classification Report:
              precision    recall  f1-score   support

    No Churn       0.97      0.98      0.97       570
       Churn       0.89      0.79      0.84        97

    accuracy                           0.96       667
   macro avg       0.93      0.89      0.91       667
weighted avg       0.95      0.96      0.95       667
```

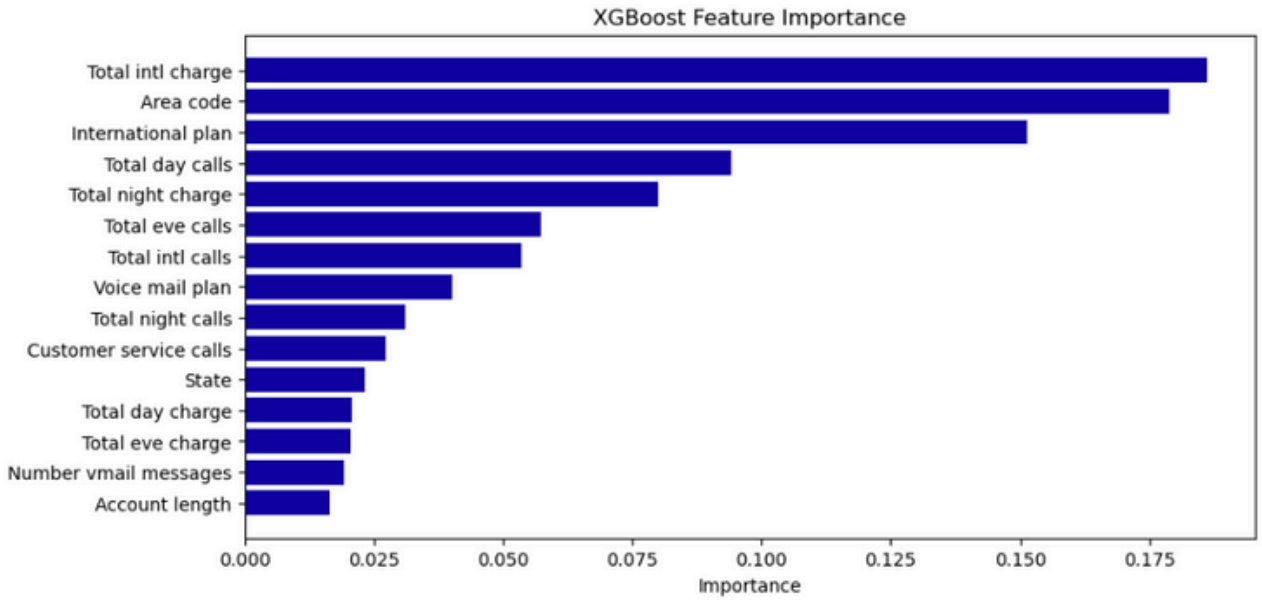


- Macro average: Precision 0.93, Recall 0.89, F1-score 0.91
- Weighted average: Precision 0.95, Recall 0.96, F1-score 0.95

Interpretation: XGBoost demonstrates strong predictive ability for both churners and non-churners, effectively handling the class imbalance. It outperforms both Logistic Regression and Decision Tree models, making it the most suitable model for this churn prediction task.

Feature Importance – XGBoost

The importance of each feature in predicting churn was evaluated using the XGBoost model. A visual representation of feature importance is included in the report.



ROC-AUC – XGBoost

The XGBoost model achieved a ROC-AUC score of 0.9146, indicating strong discrimination between churners and non-churners. A ROC-AUC value close to 1 means the model is highly effective at distinguishing positive cases (churners) from negative cases (non-churners).

This metric, along with the confusion matrix and F1-scores, confirms that XGBoost is performing well on the imbalanced dataset and is the most suitable model for predicting customer churn.