Predicting Databel Customer Churn Using Machine Learning Models

Created Date: October 2023

Author: Thanita Eiamvijit

Coding URL: https://github.com/thanita-evj/Data-Analysis.git

Table of Contents

1 Introduction 2
2 Definition of “Churn” 2
3 Objective 3
4 Setting Up the Environment 3
5 Data Exploration (Exploratory Data Analysis) 3
5.1 Overview of Dataset 3
5.2 Duplicate Rows 5
5.3 Missing Values 5
5.4 Remove Redundant Variables 6
5.5 Univariate Analysis (Descriptive Statistics) 6
5.5.1 Central Tendency and Dispersion 6
5.5.2 Distribution of Numerical Variables 7
5.5.3 Distribution of Categorical Variables 7
5.5.4 Outlier Detection 9
5.6 Bivariate Analysis (Correlation Analysis) 9
6 Data Preprocessing 11
6.1 Encoding Categorical Variables 11
6.2 Feature Scaling (Standardization) 11
6.3 Feature Selection 12
7 Independent and Dependent Variables 13
8 Imbalanced Data 13
8.1 Imbalanced Data 13
8.2 Resampling Technique 13
9 Data Splitting 14
10 Model Selection, Training, and Evaluation 14
10.1 Model Selection 14
10.2 Training the Model 14
10.3 Model Evaluation 14
10.3.1 Accuracy 15
10.3.2 Confusion Matrix 15
10.3.4 ROC Curve 17
11 Save and Test the Model using Sample Set 18
12 Conclusion 19

1. Introduction

In today's high business competition, customer retention is more importance than ever. The loss of customers, commonly referred to as "Customer Churn", becomes significant challenges for businesses, both in terms of lost revenue and increased marketing costs associated with acquiring new customers. Predicting and understanding the reasons behind churn can help implementing targeted strategies that increase customer satisfaction and loyalty.

In this way, machine learning becomes an invaluable tool with the ability to process large amounts of data and identify complex patterns. Machine learning models can predict potential churners with a level of accuracy. This project delves deep into the customer churn prediction based on Databel's dataset, employing various machine learning models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost.

2. Definition of "Churn"

According to Investopedia, the Churn Rate, also known as the rate of attrition or customer churn, is the rate at which customer stop doing business with an entity. It is commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given period.

The churn rate formula is the number of customers lost during the period divided by total number of customers. Churn rate is commonly expressed as a percentage used to evaluate the retention success and potential growth or decline of a business's customer base. A high churn rate may indicate customer dissatisfaction, while a low rate suggests customers are satisfied and remain loyal to the product or service. Minimizing the churn rate is crucial for sustained growth and profitability.

Churn Analysis is the process of evaluating and understanding the reasons and patterns behind customer choosing to end their relationship with a company or service over a specific period. Churn analysis is including:

- Descriptive analytics: It is used to understand the current churn rate and profile of churned customers.
- Predictive analytics: It is used to apply machine learning models to predict potential churn based on customer behavior and characteristics.
- Prescriptive analytics: It is used to recommend specific actions to prevent or reduce churn.

3. Objective

The objective of conducting customer churn prediction is to proactively identify customers who are most likely to discontinue using company's products or services in the near future, enabling businesses to implement targeted retention strategies. This predictive approach allows business to maximize customer lifetime value, optimize resource allocation, and increase overall profitability by focusing on maintaining existing customer relationships rather than solely investing in acquiring new ones. Through understanding and addressing specific reasons leading customer churn, business aims to improve customer satisfaction, loyalty, and long-term revenue generation.

4. Setting Up the Environment

Set up Python environment by importing necessary libraries and loading the dataset.

- Data analysis libraries: "numpy" for mathematical functions and "pandas" for data analysis.
- Visualization libraries: "matplotlib.pyplot" and "seaborn".
- Machine learning libraries: "sklearn" for data splitting, data preprocessing, matrices, various algorithms and "imblearn" for handling imbalanced dataset.
- Machine learning models: "LogisticRegression", "RandomForestClassifier", "DecisionTreeClassifier", "SVC", and "xgboost".
- Evaluation: "roc_curve" and "roc_auc_score" for plotting ROC curve and area under the curve.

5. Data Exploration (Exploratory Data Analysis)

5.1 Overview of Dataset

The dataset used in this project is a fictional churn dataset from a Telecom provider named Databel, the dataset consists of 29 columns or variables and 6687 rows of customer records with no time dimension. The description of variables is described in Table 1:

Table 1. The description of variables

No.	Column Name	Description	Data Type
1	Customer ID	Customer number with a unique identifier.	Categorical
2	Churn Label	Indicate whether a customer has left the company’s service or not. (Yes, No)	Categorical
3	Account Length (in month)	The duration in months that the account has been active.	Numeric
4	Local Calls	The number of local calls made by the customer.	Numeric
5	Local Mins	The total number of minutes spent on local calls.	Numeric
6	Intl Calls	The number of international calls made by the customer.	Numeric
7	Intl Mins	The total number of minutes spent on international calls.	Numeric
8	Intl Active	Indicate if the customer is active in making international calls. (Yes, No)	Categorical
9	Intl Plan	Specify if the customer has an international calling plan. (Yes, No)	Categorical
10	Extra International Charges	Additional charges incurred for international services beyond the plan.	Numeric
11	Customer Service Calls	The number of times a customer has contacted customer service.	Numeric
12	Avg Monthly GB Download	Average monthly data usage by the customer in gigabytes.	Numeric
13	Unlimited Data Plan	Indicate if the customer is on an unlimited data plan. (Yes, No)	Categorical
14	Extra Data Charges	Additional charges incurred for data usage beyond the plan.	Numeric
15	State	U.S. state in which the customer resides.	Categorical
16	Phone number	Customer’s phone number.	Categorical
17	Gender	The gender of the customer (Male, Female)	Categorical
18	Age	Age of the customer.	Numeric
19	Under 30	The customer is below 30 years of age. (Yes, No)	Categorical
20	Senior	The customer is categorized as a senior. (Yes, No)	Categorical

21	Group	The customer is part of any group or not. (Yes, No)	Categorical
22	Number of Customers in Group	The customer mentioned in the group.	Numeric
23	Device Protection & Online Backup	Indicate if the customer has both device protection and online backup services. (Yes, No)	Categorical
24	Contract Type	The type of contract the customer has with the company (Month-to-Month, One Year, and Two Year).	Categorical
25	Payment Method	The method by which the customer pays their bills (Credit Card, Direct Card, and Paper Check).	Categorical
26	Monthly Charge	The amount charged to the customer on a monthly basis.	Numeric
27	Total Charges	The total amount charged to the customer till date.	Numeric
28	Churn category	Attitude, Competitor, Dissatisfaction, Price, and Others.	Categorical
29	Churn Reason	Reasons by customers to stop using the product or service.	Categorical

5.2 Duplicate Rows

Examine the dataset for any potential duplicate entries to ensure that the data quality and integrity is accurate for analysis and model training. This dataset has no duplicate rows, ensuring that the subsequent prediction is based on unique customer records.

5.3 Missing Values

Missing values were identified specifically in “Churn Category” and “Churn Reason” columns. Both columns show the same count of 4918 missing entries. Recognizing missing values in dataset is an important step to ensure the precision of subsequent analysis and prediction, as missing information can potentially skew results.

5.4 Remove Redundant Variables

The column “Customer ID”, “Phone Number”, and “Churn Reason” were excluded from the original dataset to enhance the efficacy of the modeling process. “Churn Reason” column comprises of unique textual reasons, which can introduce unnecessary complexity to the model. By omitting these columns, it aims to focus on features that directly contribute to a more efficient predictive analysis.

5.5 Univariate Analysis (Descriptive Statistics)

Univariate analysis focuses on examining a single data variable. Since it's a single variable, it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

5.5.1 Central Tendency (Mean, Median, Mode) and Dispersion (Range, Standard Deviation, Interquartile)

A summary of key findings of customer behaviors based on statistics values:

- **Account Length (in months):** Customers have been with the company for an average 32 months, with a range from 1 months to a maximum of 77 months. Half of the customers have been with the company for 29 months or less.
- **Local Calls & Local Mins:** On average, customers make about 131 local calls and spend about 323 minutes on these calls. There is a wide range in the number of calls made from 1 to 918 and the duration of these call from 4 to 1234 minutes.
- **International Calls & Mins:** Customers make an average of 51 international calls and spend about 130 minutes, but the standard deviation is high (around 103) indicating the variability. Some customers do not make international calls at all, as indicated by the minimum value of 0.
- **Extra International Charges:** The average additional charge for international services beyond the plan is approximately $34. However, 75% of customers incur charges of $16.4 or less, indicating that a smaller proportion of customers face higher charges.
- **Customer Service Calls:** Customers contact customer service on average less than once (0.92 times). A majority (75%) of customers calls two times or fewer.
- **Avg Monthly GB Download & Extra Data Charges:** Customers download an average of 6.7 GB of data monthly, incurring an average of $3.37 in extra data charges, but 75% of customers do not have any additional data charges.
- **Age:** The average age of customers is around 47 years with the youngest being 19 and the oldest 85. Half of the customer base us aged 47 or younger.
- **Number of Customers in Group:** 75% of data indicate 0, which show the customer did not have a group.
- **Monthly Charge & Total Charges:** Customers are charged an average of $31 monthly. Throughout their tenure with the company, the customers have been charged average total of $1084, with some customers being charged as much as $5574.

5.5.2 Distribution of Numerical Variables

The majority of histograms (see Figure 1) show a right-skewed distribution, indicating that most customers fall into the lower range for many of these services and charges. Moreover, certain variables such as “Number of Customers in Group” and “Customer Service Calls” are heavily skewed, with most of data points clustered at the lower end. The age distribution is the most normally distributed among all variables.

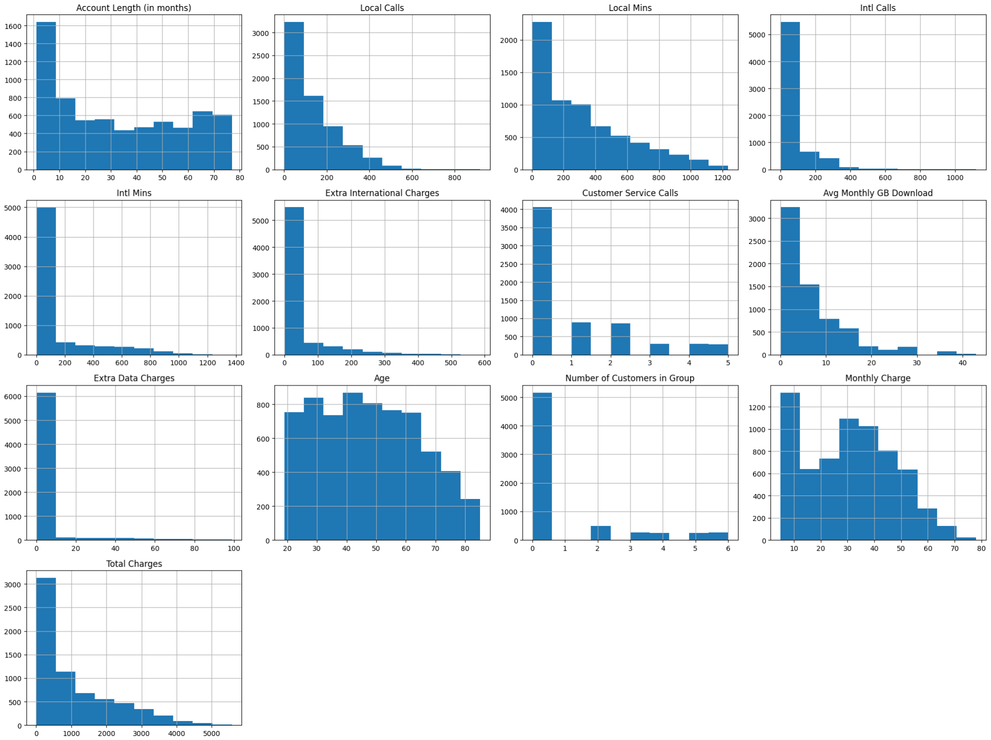


Figure 1. Distribution of numerical variables in histograms

5.5.3 Distribution of Categorical Variables

The distribution of categorical variables in Figure 2 shows that most customers have not churned and prefer month-to-month contracts. The top reasons for churn are unprovided, followed by competitors. The majority did not opt for additional services such as “Intl Plan”, “Device Protection & Online Backup”, or “Unlimited Data Plan”. Direct debit is the main method of payment. The gender distribution is balanced, while attributes such as “Senior” and “Under 30” indicate that most customers are not seniors and not under the age of 30. Furthermore, there is a diverse distribution of customers across states (see Figure 3) with “WV” having the highest number and “CA” having a notably smaller customer base. Most of states have a relatively even distribution of customers, clustered around 100-150 range.

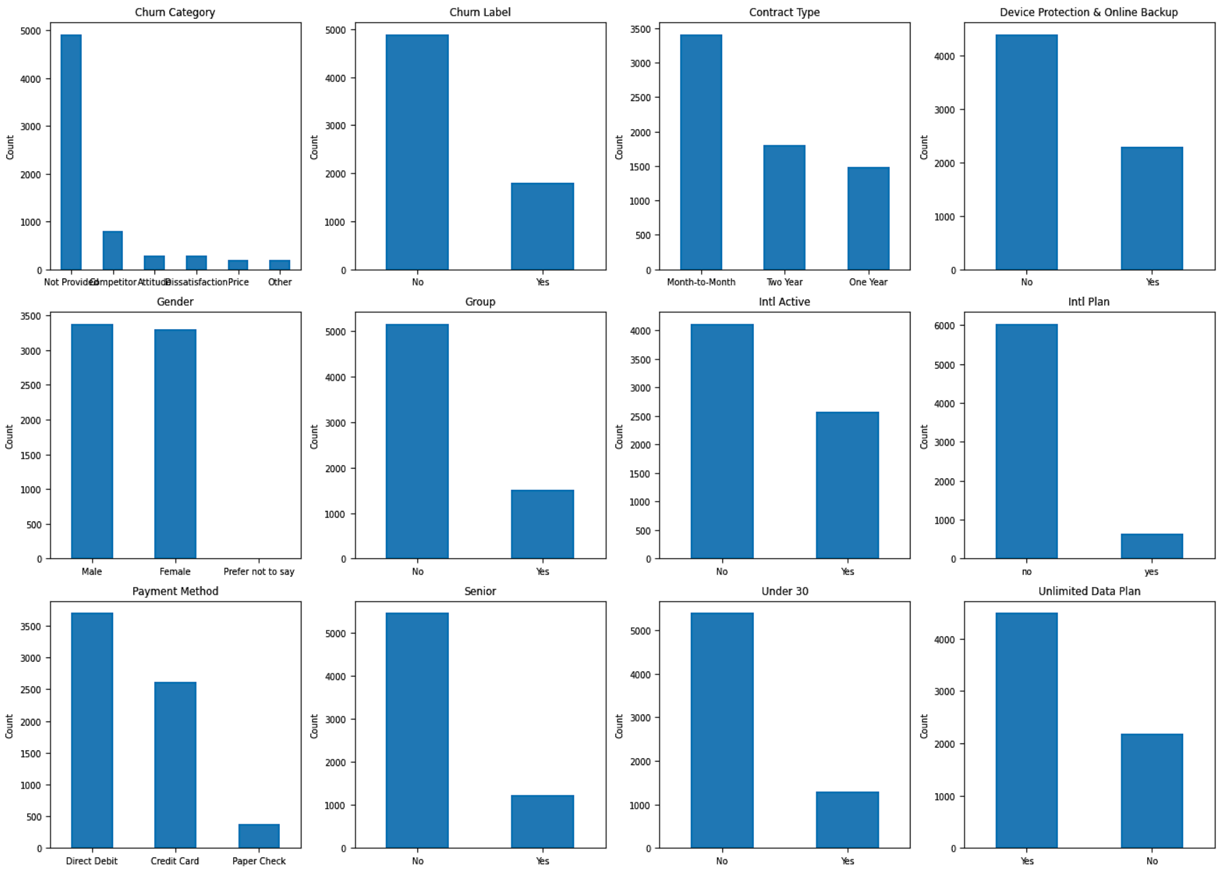


Figure 2. Distribution of categorical variables

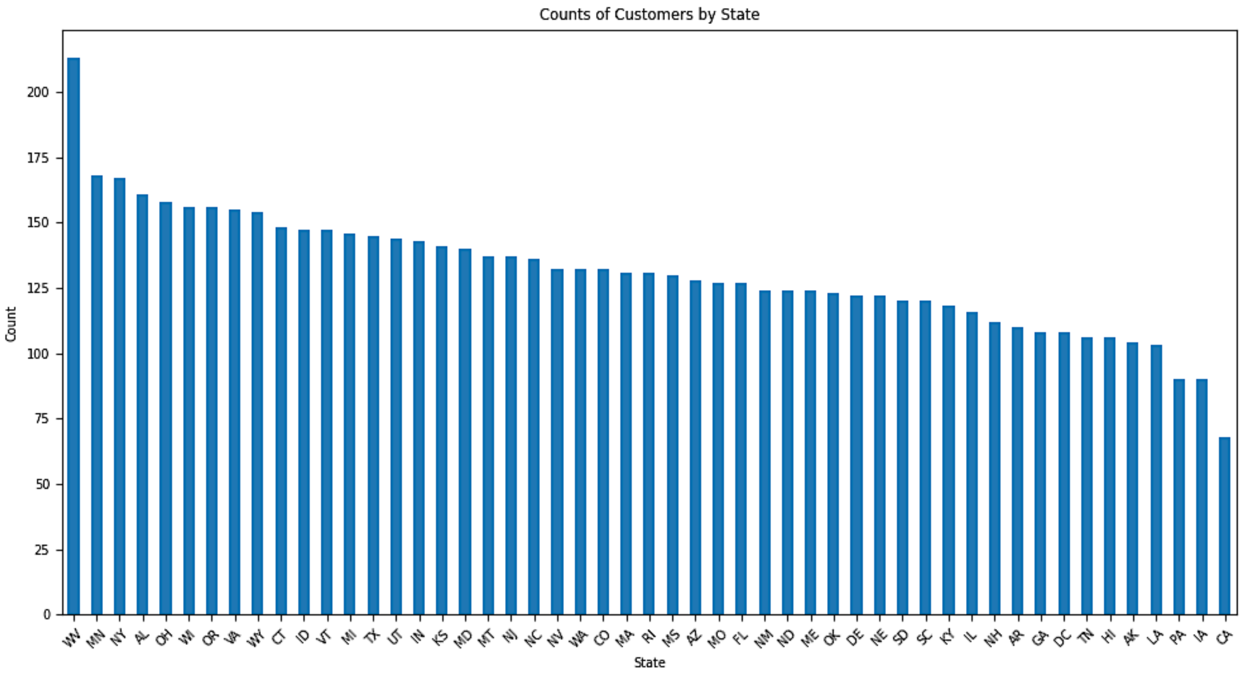


Figure 3. Distribution of customers by state

5.5.4 Outlier Detection

The highest percentage of outliers (see Table 2) are in “Number of Customers in Group”, “Extra International Charges”, “Intl Calls”, and “Intl Mins”, all with percentages exceeding 15% considering high percentage of outliers that outright removal can lead to a substantial loss of data. For churn prediction, understanding the behavior of outliers can be crucial. Sometimes, outliers might represent high-value customers or those with specific needs. Instead of deleting the outliers, considering applying scaling and using algorithms that are less sensitive to outliers might be the better option for this project.

Table 2. Percent of outliers for numerical features

Item	Feature Name	Percent of Outlier
1	Number of Customers in Group	22.75%
2	Extra International Charges	20.44%
3	Intl Calls	16.23%
4	Intl Mins	16.09%
5	Extra Data Charges	10.29%
6	Avg Monthly GB Download	5.46%
7	Local Calls	1.85%
8	Total Charges	1.84%
9	Local Mins	0.57%
10	Account Length (in months)	0.00%
11	Customer Service Calls	0.00%
12	Age	0.00%
13	Monthly Charge	0.00%

5.6 Bivariate Analysis (Correlation Analysis)

Bivariate analysis is a statistical method that used to determine the relationship between two variables such as correlation or scatter plot, studying whether a relationship exist and how strong it may be for a pair of data. In Figure 4 provides a heatmap of correlation coefficients between various features of a dataset. Correlation coefficient measures the strength and direction of linear relationship between two continuous variables, ranging between -1 (negative correlation) and 1 (positive correlation).

The dataset shows the strong positive correlation as follows:

- “Local Calls” is strongly correlated with “Local Mins” (0.96) and “Account Length” (0.86). Since “Local Calls” and “Local Mins” have a very high correlation, which make senses as more minutes likely equate to more calls. This might cause potential multicollinearity if used together in predictive models in which the models might struggle to differentiate the effect on the outcome since they move together.
- “Intl Calls” is closely associated with “Intl Mins” (0.86).

- “Account Length” is highly correlated with “Total Charges” (0.79) and “Local Mins” (0.84).

There is a lack of strong correlations between many features, suggesting that they provide unique information, which could be useful for modeling.

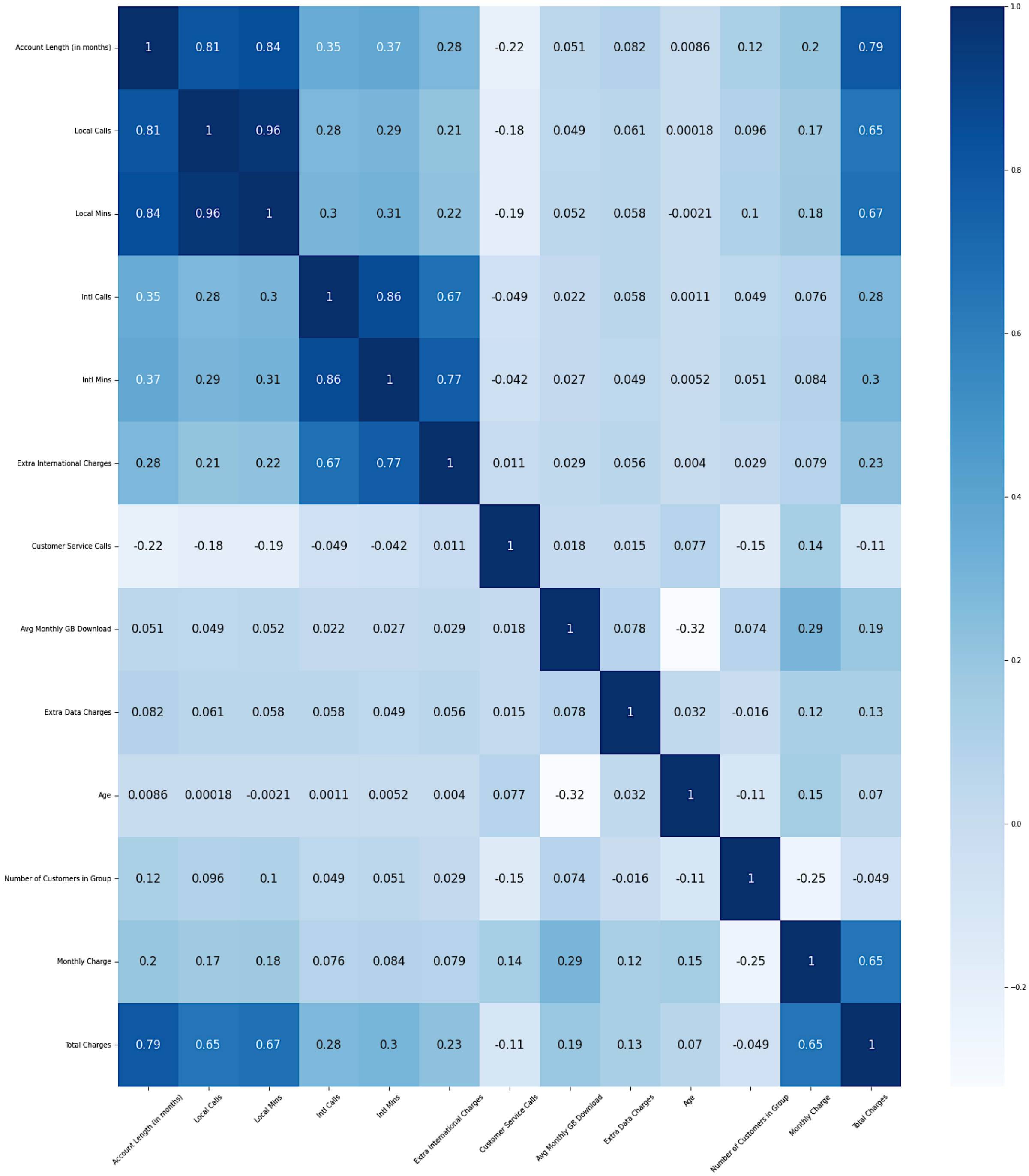


Figure 4. Heatmap of correlation coefficients

6. Data Preprocessing

6.1 Encoding Categorical Variables

Encoding is the process of converting categorical variables into a unique numeric format that can be provided to machine learning algorithms to improve predictions. The “LabelEncoder” from “sklearn.preprocessing” module is one method used for this purpose in the project. The categorical variables such as “Churn Label”, “Intl Active”, “Intl Plan”, “Unlimited Data Plan”, “State”, “Gender”, “Under 30”, “Senior”, “Group”, “Device Protection & Online Backup”, “Contract Type”, “Payment Method”, “Churn Category” have been encoded.

6.2 Feature Scaling (Standardization)

Feature scaling is used to standardize the range of independent variables in the dataset. “StandardSacler” from “sklearn.preprocessing” is applied, which standardizes features by removing the mean and scaling them to unit variance. It is done by subtracting the mean of each feature and then dividing by its standard deviation. As a result, each feature will closely have a mean of 0 and standard deviation of 1 (see Table 3).

Table 3. Mean and standard deviation values of standardization

Item	Feature Name	Mean	Standard Deviation
1	Account Length (in months)	0.000000e+00	1.000075
2	Local Calls	1.700117e-17	1.000075
3	Local Mins	-2.890199e-16	1.000075
4	Intl Calls	-1.275088e-17	1.000075
5	Intl Mins	6.375440e-17	1.000075
6	Extra International Charges	8.075557e-17	1.000075
7	Customer Service Calls	3.400234e-17	1.000075
8	Avg Monthly GB Download	1.700117e-17	1.000075
9	Extra Data Charges	-2.550176e-17	1.000075
10	Age	-7.544270e-17	1.000075
11	Number of Customers in Group	1.360094e-16	1.000075
12	Monthly Charge	-8.925615e-17	1.000075
13	Total Charges	-1.020070e-16	1.000075

6.3 Feature Selection

Feature selection is the process of selecting the most important features from the original set of features in a dataset to improve model performance, overfitting, and simplify models for better interpretation. This project applied feature importance using Random Forest classifier.

In the Figure 5 shows that “Churn Category” has the most important feature value close to 0.6. However, churn category is directly related to the target variable “Churn Label” as it provides the cause for the churn, then it is not appropriate to include in the model training process, it would leak information about the target into the model. In the practice, it would not have access to the reason for churn before predicting whether a customer will churn or not. Therefore, it is recommended to remove it from the feature used to train the model.

There is a variation in importance among the other features underscores the significance of understanding the underlying reasons for customer churn.

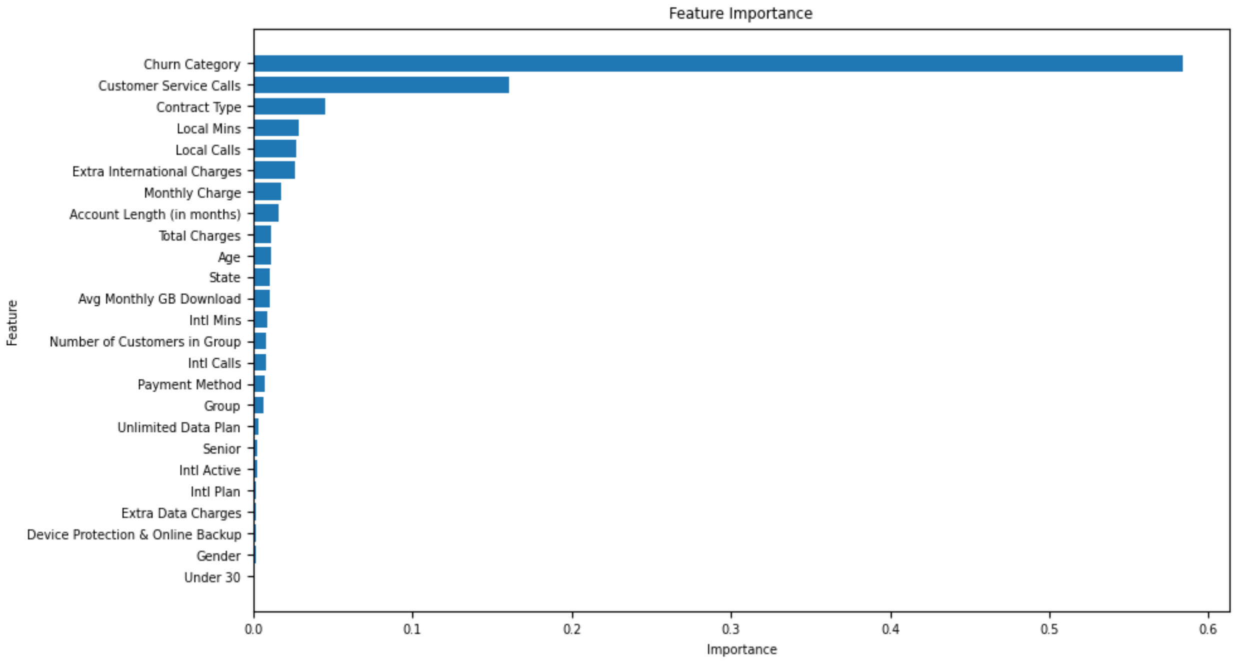


Figure 5. Feature importance

7. Independent and Dependent Variables

X is the independent variables by dropping columns “Churn Label”, “Churn Category”, “Local Calls”, “Number of Customers in Group”, and “Intl Mins”, leaving the remaining columns as the features to be used for training and prediction.

y is set to “Churn Label” column, which represents the target (dependent variable) or outcome for prediction.

8. Imbalanced Data

8.1 Imbalanced Data

Imbalanced data refers to a situation in classification where the number of observations for one class significantly exceeds the number of observations for the other classes, representing unequal classification. The imbalance in dataset can lead to biased models that perform well on the majority class but poorly on the minority class.

Here is the distribution of churn label in the dataset about 73% of the samples belong to “0” class (customer who did not churn) and remaining 27% of the samples belong to “1” class (customer who did churn). The ratio of 73:27 represents imbalanced data where the class label “0” is the majority class making up three-quarter of dataset, while the class label “1” is the minority class.

8.2 Resampling Technique

Since the dataset is highly imbalanced and most of machine learning algorithms work best when the number of samples in each class are about equal. Resampling is applied to adjust the distribution of data samples in the dataset.

Synthetic Minority Over-sampling Technique (SMOTE) is used in this project for resampling method to balance dataset with uneven class distributions. SMOTE works by generating new, synthetic samples of the minority class. For every data point in the minority class, SMOTE picks a few of its closest neighbors and then creates new points somewhere between the chosen point and its neighbors. These new, synthetic points are added to the dataset increasing the count of the minority class. This helps in reducing the risk of overfitting. After applying SMOTE, the class is balanced with 50% or 4891 of samples belong to “0” class and another 50% belong to “1” class.

9. Data Splitting

Data splitting using “train_test_split” function, is to split the data into training and testing sets for assessing the performance of machine learning models accurately. Training set is used to train the machine learning models, it is the largest portion of the dataset where the model learns patterns and relationships in the data. Testing set is the smaller portion of the data used for evaluating the model’s performance after training. The dataset in this project is divided using ratio of 70:20 for training and testing sets.

10. Model Selection, Training, and Evaluation

10.1 Model Selection

Model selection for customer churn prediction is the process of choosing the most suitable machine learning algorithm to predict which customers are most likely to terminate their services in order to identify and retain these customers and reduce the overall churn. This project applies five different machine learning algorithms for prediction as described below:

- **Logistic Regression:** It is used for modeling the probability of a binary outcome based on one or more predictive variables. For churn prediction, it would model the probability of a customer churn or not.
- **Decision Tree:** It breaks down a dataset into smaller subsets, making a decision at every level forming a tree-like model. For churn prediction, the tree would split customers based on the features.
- **Random Forest:** It is an ensemble method that builds multiple decision trees or a forest of decision trees. Each tree is trained on a random subset of the data and makes its own predictions. The Random Forest aggregates these predictions to produce a final result, which makes it more robust and often has higher accuracy than a single decision tree.
- **Support Vector Machine (SVM):** It tries to find a hyperplane that best divides a dataset into classes (churning or non-churning).
- **eXtreme Gradient Boosting (XGBoost):** It is an ensemble method and implementation of gradient boosting trees algorithm, but instead of fitting multiple trees in parallel like Random Forest, it builds one tree at a time, where each tree corrects the errors of the previous one.

10.2 Training the Models

Each of the selected models is trained using a training dataset allowing the models to learn the patterns and relationships.

10.3 Model Evaluation

After training, the models are tested on a separated test set to evaluate their performance.

10.3.1 Accuracy

Accuracy is the ratio of correctly predicted instances to the total instances to provide how well the model is performing. In Figure 6 shows that ensemble methods, particularly XGBoost (0.92) and Random Forest (0.91) have outperformed other models indicating their ability to capture complex relationships in the data. There is not a large gap between the highest (XGBoost) and lowest (SVM) accuracies, suggesting that the data is relatively good, and most models can capture its underlying patterns. However, SVM has the worst performance, it might require further hyperparameter tuning to perform better.

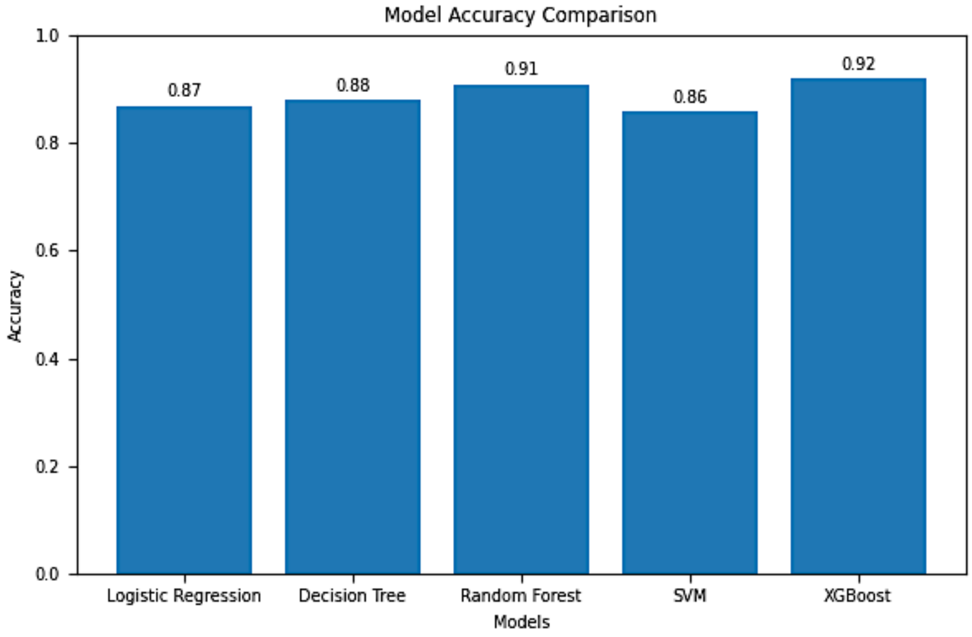


Figure 6. Accuracy of predictive models

10.3.2 Confusion Matrix

Confusion matrix is used to evaluate the performance of a classification model by presenting the actual and predicted classification. Confusion matrix consists of four values:

- **True Positive (TP):** The number of customers correctly predicted to churn. The model correctly identified as the customers going to churn so the company can target this group of customers with retention strategies before leaving.
- **True Negative (TN):** The number of customers correctly predicted to not churn. These customers are loyal, and the model correctly identified them.
- **False Positive (FP):** The number of customers incorrectly predicted to churn. The model wrongly tagged these customers as potential churners, which might lead to unnecessary retention offers.
- **False Negative (FN):** The number of customers incorrectly predicted to not churn. This is the riskiest group in the context of churn prediction. The customers churned,

but the model failed to predict it. This causes missed of opportunities for intervention.

In churn prediction, minimizing False Negative is crucial due to the association with revenue implication. At the same time, False Positive should be focused to ensure that retention resources are utilized effectively. In Table 4 provides a representation of confusion matrices for the fived evaluated models and summarizes as follows:

- **Logistic Regression:** It has almost equal TP (849) and TN (848), indicating that it predicts both classes (churn and no churn) with similar efficiency. FP (131) and FN (129) are closely equal, suggesting some misclassification but with balanced errors.
- **Decision Tree:** Higher TP (872) compared to Logistic Regression, meaning it has better at correctly predicting those customers who will churn. Slight decrease in FN (106) and FP (123) compared to Logistic Regression.
- **Random Forest:** It offers higher TN (895) compared to other models except XGBoost, meaning it is excellent at correctly predicting customers who will not churn. TP (886) is also strong, making it efficient at predicting both classes.
- **SVM:** It is comparable TN (850) to Logistic Regression. Highest FN (138) among all models, indicating it tends to miss more actual churners.
- **XGBoost:** It presents the highest TP (889) and TN (902), making it the best at correctly predicting customers who will churn and will not churn. Also it has the lowest FP (77) and FN (89), showing fewer overall misclassification.

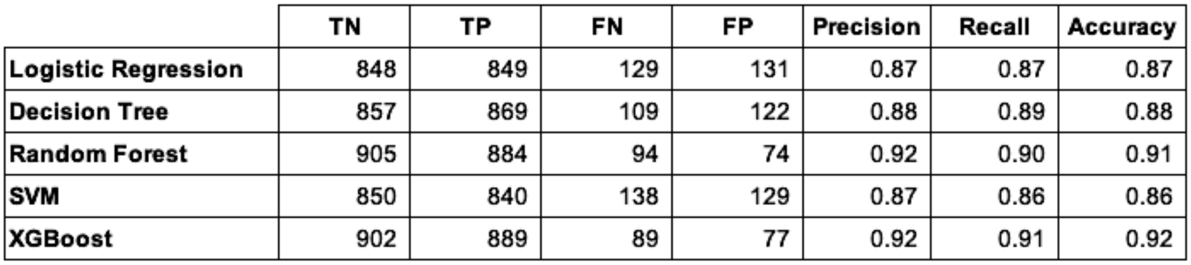
XGBoost and Random Forest stands out as the most efficient models showing strong TP and TN values, indicating the effectiveness of churners and non-churners. In term of minimizing misclassifications, Logistic Regression and Decision Tree have balanced results, but they do not outshine Random Forest and XGBoost.

Metrics derived from Confusion Matrix:

- Precision: It measures the accuracy of positive prediction.
- Recall (or Sensitivity or True Positive Rate): It measures the proportion of actual positives that were identified correctly.

Table 4 provides a representation of precision and recall for the fived evaluated models. It shows that XGBoost has the highest precision at 0.92, meaning that the model is right about 92% of the time when it says a customer will churn. It suggests that XGBoost is the best among other models at making correct positive prediction. Furthermore, XGBoost and Decision Tree have almost the same recall at about 0.91, spotting 91% of the customers who churn correctly. They are better than other models at identifying all potential positive instances.

Table 4. Summary of model results



	TN	TP	FN	FP	Precision	Recall	Accuracy
Logistic Regression	848	849	129	131	0.87	0.87	0.87
Decision Tree	857	869	109	122	0.88	0.89	0.88
Random Forest	905	884	94	74	0.92	0.90	0.91
SVM	850	840	138	129	0.87	0.86	0.86
XGBoost	902	889	89	77	0.92	0.91	0.92

10.3.3 ROC Curve

ROC curve is used to evaluate the performance of a binary classification model. The curve helps to show the model’s capability to distinguish between the positive and negative classes. ROC curve consists of two parameters:

- True Positive Rate (or recall): It is on y-axis used to measure the proportion of actual positives that are correctly identified.
- False Positive Rate: It is on x-axis used to measure the proportion of actual negatives that are incorrectly identified as positives.

A curve that is closer to the top-left corner (0,1) of the plot indicates a better performing model and a curve that is closer to the random guessing line indicates a less effective model.

Area Under Curve (AUC) is the overall ability of the model to differentiate between the positive and negative classes. The higher value of AUC is the better on classification with less random guessing.

From Figure 7 shows that XGBoost and Random Forest are the standout performers with their ROC curves close to the top-left corner and high AUC values at about 0.98 and 0.97, respectively.

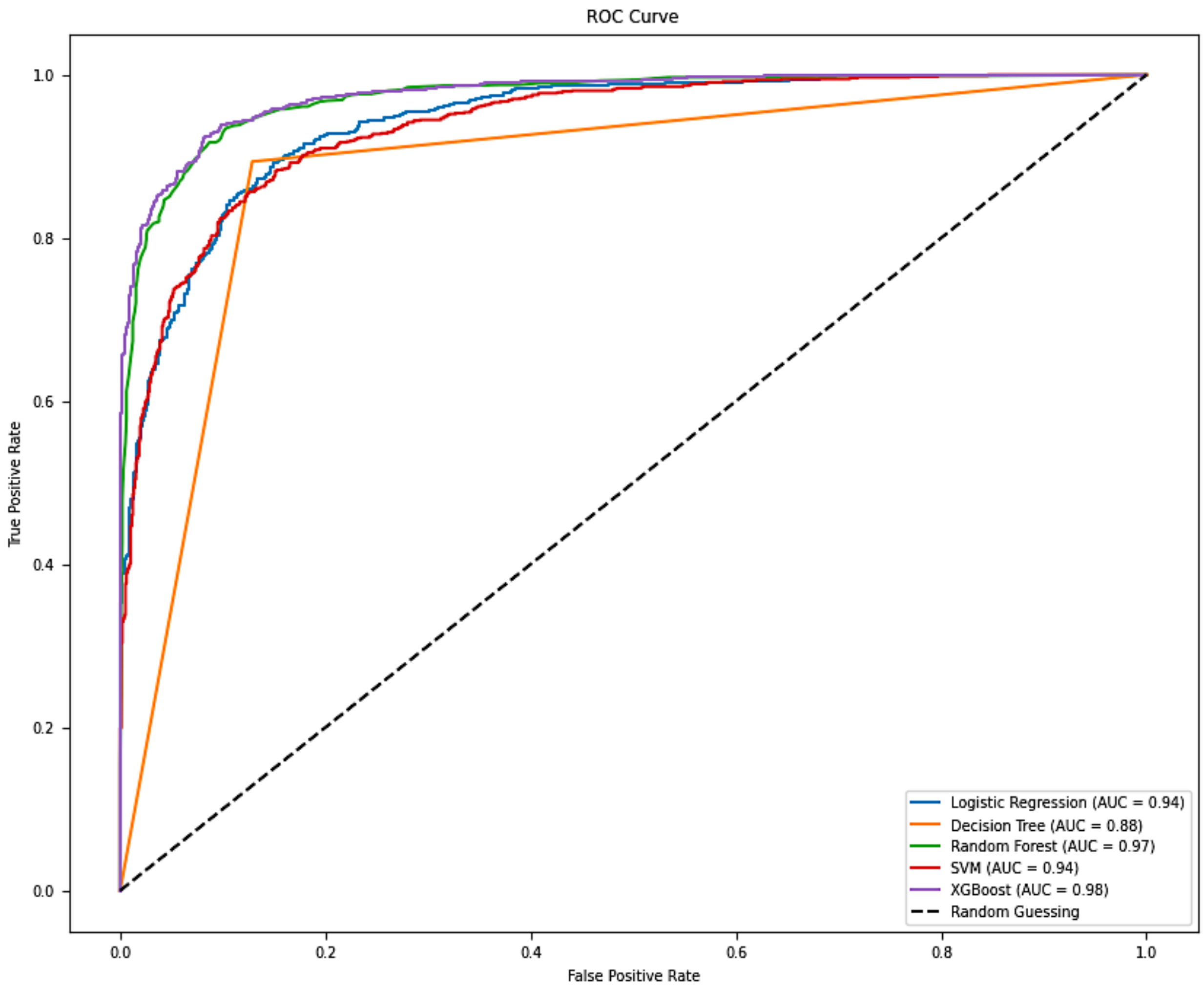
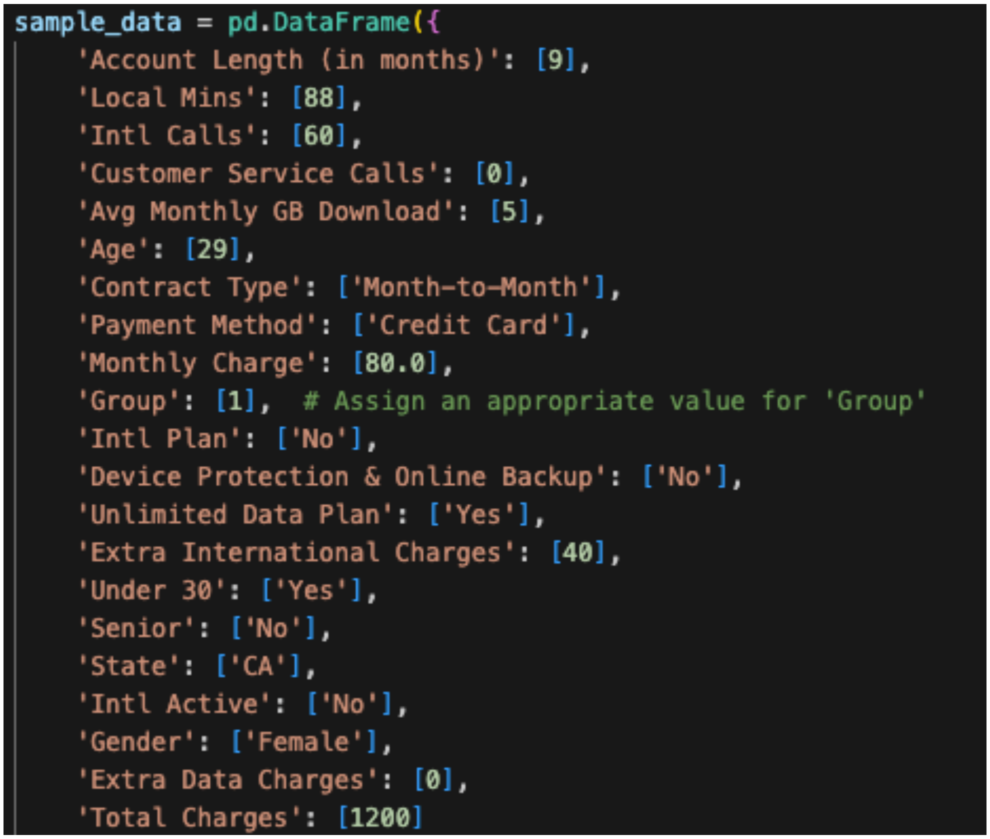


Figure 7. ROC curve

11. Save and Test the Model using Sample Set

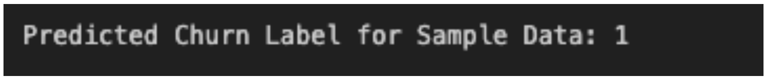
XGBoost performs the highest performance in predicting customer, the model is saved using the “pickle” for further use. The saved XGBoost model is reloading to predict the churn label for the sample data. The sample data is set as shown in Figure 8.



```python
sample_data = pd.DataFrame({
    'Account Length (in months)': [9],
    'Local Mins': [88],
    'Intl Calls': [60],
    'Customer Service Calls': [0],
    'Avg Monthly GB Download': [5],
    'Age': [29],
    'Contract Type': ['Month-to-Month'],
    'Payment Method': ['Credit Card'],
    'Monthly Charge': [80.0],
    'Group': [1],  # Assign an appropriate value for 'Group'
    'Intl Plan': ['No'],
    'Device Protection & Online Backup': ['No'],
    'Unlimited Data Plan': ['Yes'],
    'Extra International Charges': [40],
    'Under 30': ['Yes'],
    'Senior': ['No'],
    'State': ['CA'],
    'Intl Active': ['No'],
    'Gender': ['Female'],
    'Extra Data Charges': [0],
    'Total Charges': [1200]
```

Figure 8. Snapshot of sample data

A predicted result based on the sample data in Figure 9. shows a predicted churn label of "1" indicates that, according to the XGBoost model, the customer is likely to churn based on the provided information. Therefore, business can use this prediction as a tool for customer churn retention strategies.



```
Predicted Churn Label for Sample Data: 1
```

Figure 9. Snapshot of predicted result of sample data

12. Conclusion

In an attempt to predict and understand customer churn, this project has gone through from data collection to analytics. This project has identified key variables influencing churn, addressed challenges such as imbalanced dataset, and employed a variety of machine learning models to evaluate prediction accuracy. Through the model evaluations, the models have provided both reliability and interpretability, especially XGBoost and Random Forest, which outperformed and more robust in term of accuracy (about 92%) and correctness compared to Logistic Regression, Decision Tree, and Support Vector Machine.