Customer Churn Prediction in Telecommunications

Candidate Details:
Name: Abhishek Kumar Singh
Email: yadavabhishekkumar41@gmail.com

Project Overview

This project aims to predict customer churn in a telecommunications company using machine learning. The process involves several steps: environment setup, data loading and preprocessing, exploratory data analysis (EDA), feature engineering, model building, evaluation, and result visualization.

1.Dataset Description

The dataset contains customer information from a telecommunications company, used to predict whether a customer will churn (leave the service) or not. Here are the details:

Dataset Characteristics

Characteristic	Description
Size	7,043 rows and 21 columns
Missing Values	'TotalCharges' column has some missing values
Categorical Features	Many features are categorical and need encoding

Features

Feature	Description	Type
customerID	Unique identifier for each customer	Categorical
gender	Gender of the customer (Male, Female)	Categorical
SeniorCitizen	Indicates if the customer is a senior citizen (1) or not (0)	Categorical
Partner	Indicates if the customer has a partner (Yes, No)	Categorical
Dependents	Indicates if the customer has dependents (Yes, No)	Categorical
tenure	Number of months the customer has been with the company	Numerical
PhoneService	Indicates if the customer has a phone service (Yes, No)	Categorical
MultipleLines	Indicates if the customer has multiple lines (Yes, No, No phone service)	Categorical
InternetService	Type of internet service the customer has (DSL, Fiber optic, No)	Categorical
OnlineSecurity	Indicates if the customer has online security (Yes, No, No internet service)	Categorical
OnlineBackup	Indicates if the customer has online backup (Yes, No, No internet service)	Categorical
DeviceProtection	Indicates if the customer has device protection (Yes, No, No internet service)	Categorical
TechSupport	Indicates if the customer has tech support (Yes, No, No internet service)	Categorical
StreamingTV	Indicates if the customer has streaming TV service (Yes, No, No internet service)	Categorical

StreamingMovies	Indicates if the customer has streaming movies service (Yes, No, No internet service)	Categorical
Contract	Type of contract the customer has (Month-to-month, One year, Two year)	Categorical
PaperlessBilling	Indicates if the customer has paperless billing (Yes, No)	Categorical
PaymentMethod	Payment method used by the customer (Electronic check, Mailed check, Bank transfer, Credit card)	Categorical
MonthlyCharges	The amount charged to the customer monthly	Numerical
TotalCharges	The total amount charged to the customer	Numerical
Churn	Indicates if the customer churned (Yes) or not (No)	Categorical

Target Variable

Target Variable	Description	Type
Churn	Indicates if the customer churned (Yes, No)	Categorical

This dataset is suitable for binary classification tasks, aiming to predict the likelihood of a customer churning based on their attributes. It includes both numerical and categorical features, providing a comprehensive set of information to build predictive models.

2.Environment Setup

We begin by setting up the necessary libraries for data manipulation, visualization, and machine learning. Each library is selected for specific reasons:

- **pandas** and **numpy** for data manipulation and numerical operations.
- **matplotlib** and **seaborn** for data visualization.
- **scikit-learn** for machine learning models and evaluation metrics.

```python
# Data manipulation
import pandas as pd
import numpy as np

# Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Machine learning
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
```

3.Data Preprocessing

Loading the Dataset

The dataset is loaded from a CSV file using pandas.

Data Cleaning

1. **Dropping Irrelevant Columns**: Drop the 'customerID' column as it does not contribute to the analysis.
2. **Handling Missing Values**: Convert 'TotalCharges' to numeric and drop rows with missing values.

Encoding Categorical Variables

Convert categorical variables to numerical using one-hot encoding.

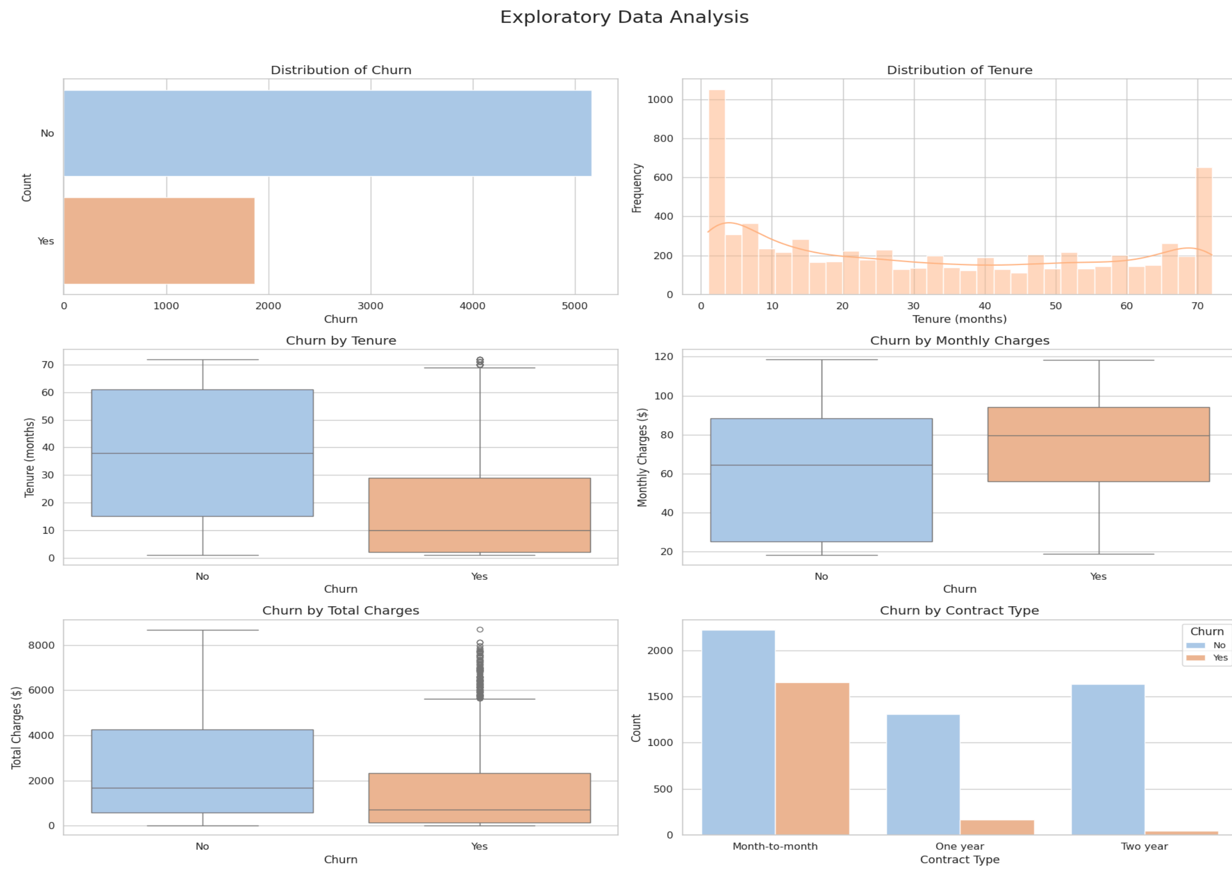
Splitting Features and Target Variable

Separate the features and the target variable.

```
# Drop irrelevant column
data.drop('customerID', axis=1, inplace=True)

# Handle missing values in TotalCharges column
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'], errors='coerce')
data.dropna(subset=['TotalCharges'], inplace=True)

# Encode categorical variables
data_encoded = pd.get_dummies(data, drop_first=True)

# Prepare data for analysis
X = data_encoded.drop('Churn_Yes', axis=1)  # Features
y = data_encoded['Churn_Yes']  # Target variable

# Display the processed data
print("Processed Data:")
print(X.head())
print("Target Variable:")
print(y.head())
```

4.Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and uncovering patterns, anomalies, and relationships within the data.



Distribution of Churn:

- This histogram shows the distribution of churn responses (Yes or No) based on two ranges of tenure (duration of service).
- The left bar represents customers who churned (‘Yes’), and the right bar represents those who did not (‘No’).
- **Observation:** Customers with shorter tenure (left range) tend to have a higher churn rate.

Distribution of Tenure:

- The orange line graph displays the distribution of tenure (in months).
- Observation: Most customers have relatively short tenure (around 10-20 months), but there are also long-term customers.
- Churn by Monthly Charges (Box Plot):
- This plot compares the distribution of monthly charges for churned customers (‘Yes’) and non-churned customers (‘No’).
- **Observation:** Churned customers tend to have slightly higher monthly charges.

Churn by Tenure (Box Plot):

- This graph compares the distribution of tenure (in months) for customers who churned (‘Yes’) versus those who did not (‘No’).
- The box plot shows the following key features:

- Median: The line inside the box represents the median tenure for each group.
- Interquartile Range (IQR): The box represents the middle 50% of data.
- Whiskers: The lines extending from the box show the range of data within 1.5 times the IQR.
- Outliers: Individual points beyond the whiskers indicate extreme values.
- **Observation:** Churned customers tend to have shorter tenure compared to non-churned customers.

Churn by Monthly Charges (Box Plot):

- This graph compares the distribution of monthly charges (in dollars) for churned versus non-churned customers.
- Similar to the previous box plot, it shows the median, IQR, whiskers, and outliers.
- **Observation:** Churned customers have slightly higher monthly charges on average.

Churn by Total Charges (Box Plot):

- Similar to the previous box plot, this one compares total charges for churned versus non-churned customers.
- **Observation:** Total charges for churned customers vary widely, but non-churned customers tend to have more consistent total charges.

Churn by Total Charges (Histogram):

- This histogram incorrectly labeled as “Churn by Total Charges” actually shows the count of churn responses (‘Yes’ or ‘No’) across different ranges of total charges.
- **Observation:** Most customers fall into the lower total charges range.

Churn by Contract Type (Bar Chart):

- This bar chart displays counts for three contract types: Month-to-Month, One Year, and Two Year.
- It splits the counts by churn responses (‘Yes’ or ‘No’).
- **Observation**: Month-to-Month contracts have the highest churn rate, while Two-Year contracts have the lowest.

5.Feature Engineering

Create new features to improve model performance.

New Features

1. **Average Monthly Charges**: Calculate the average monthly charges.
2. **Tenure Groups**: Group tenure into categories.
3. **Has Multiple Services**: Indicate if the customer has multiple services.

Encoding the Data Again

Convert the newly created categorical features to numerical.

6.Model Building

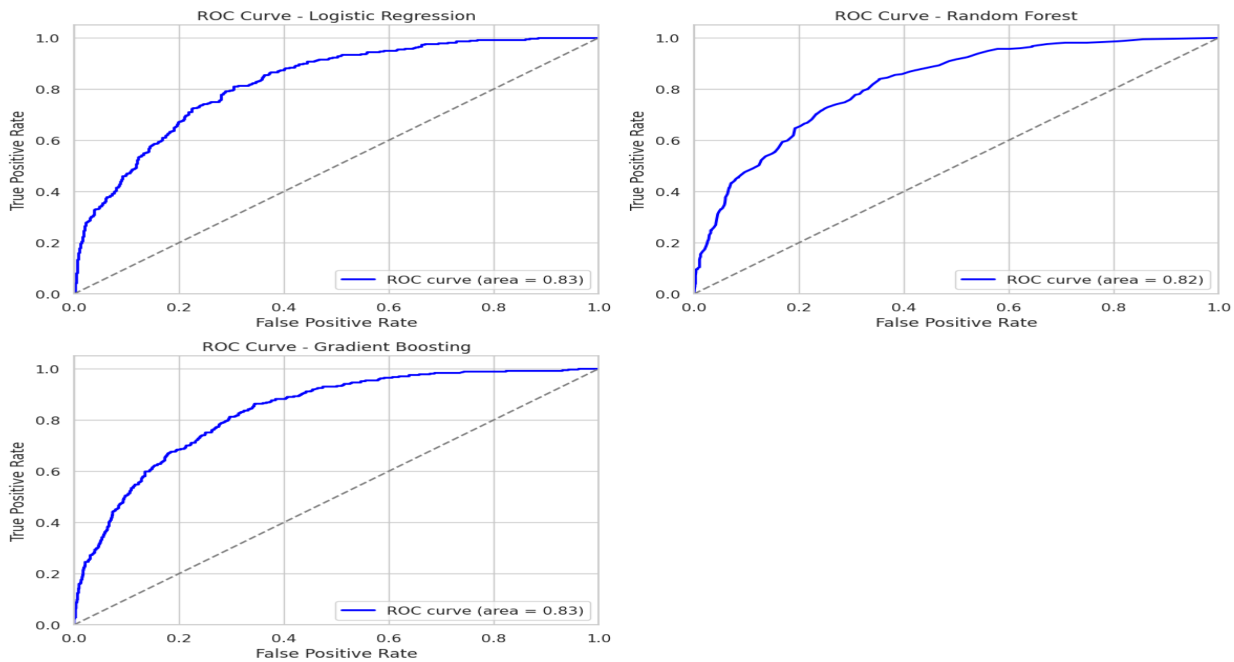
In the model building phase, three machine learning models are utilized for predicting customer churn: Logistic Regression, Random Forest, and Gradient Boosting. Let's discuss each model and why they are chosen:

Model	**Model type**	**Reason**
Logistic Regression	Classification	Simplicity and interpretability Well-suited for binary classification tasks Provides insights into feature impact
Random Forest	Ensemble Learning	Ability to handle complex relationships High accuracy and robustness against noisy data
Gradient Boosting	Ensemble Learning	Capability to capture complex relationships, Achieves high predictive accuracy and Useful for structured data and high accuracy tasks

7.Result and Graph analysis:

ROC of models:

ROC Curve: It’s a graphical representation of a classifier’s performance across different classification thresholds.
The x-axis represents the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR).

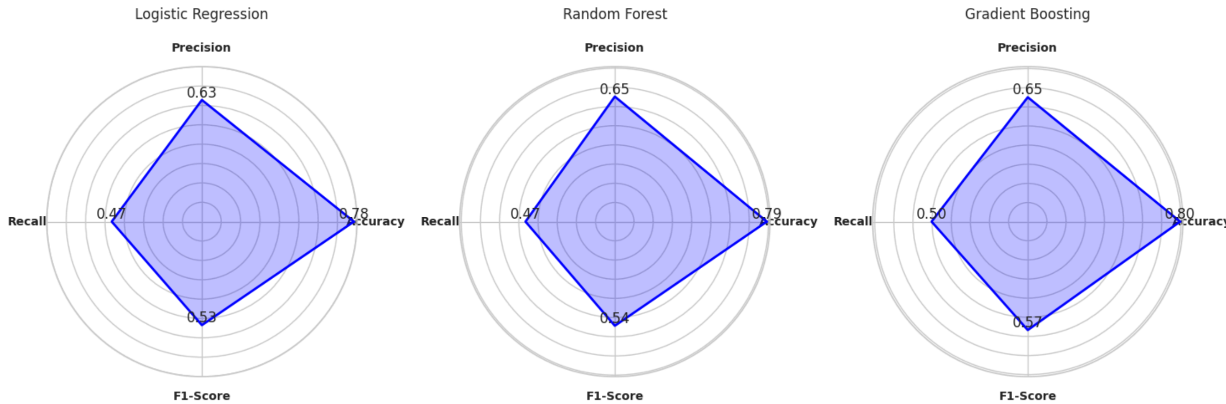


Logistic Regression:
The AUC (Area Under the Curve) is approximately **0.831**, indicating good discrimination ability.
The curve is above the random classifier line, suggesting better-than-random performance.

Random Forest:
The AUC is around **0.822**, slightly lower than logistic regression.
Again, the curve is above the random line, but not as close to the top-left corner.

Gradient Boosting:
Its AUC is also approximately **0.831**, matching logistic regression.
The curve is close to the top-left corner, indicating good performance.

Final result graph:



Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.78	0.63	0.47	0.53
Random Forest	0.79	0.65	0.47	0.54
Gradient Boosting	0.80	0.65	0.50	0.57