**CUSTOMER CHURN PREDICTION USING MACHINE LEARNING**

**Abstract:**

Customer churn is when customers stop using a company's services. In the telecom industry, reducing churn is important to keep customers and increase revenue. This project aims to **predict customer churn** using **machine learning models** like **Logistic Regression, Decision Tree, Random Forest, and LightGBM**.

I analyze customer data, including **tenure, monthly charges, contract type, and internet service**, to find key factors that influence churn. The dataset is cleaned, processed, and visualized before applying machine learning models to predict churn.

The results help telecom companies identify at-risk customers and take actions like offering discounts or better services to retain them. The **Random Forest and LightGBM models** performed best in prediction accuracy.

This project provides a **data-driven solution** to reduce churn and improve customer retention strategies.

**1. Introduction**

Customer churn happens when customers stop using a service. In the telecom industry, churn is a big problem because losing customers affects revenue and business growth. Understanding why customers leave and predicting churn can help companies **improve customer retention**.

This project aims to **predict customer churn** using **machine learning**. By analyzing customer data such as **tenure, monthly charges, contract type, and internet service**, I can identify patterns that indicate whether a customer is likely to leave.

To achieve this, I:

1. **Analyze and preprocess the dataset** (handle missing values, remove outliers, and encode categorical data).
2. **Perform Exploratory Data Analysis (EDA)** to find trends and important features.
3. **Use machine learning models** like **Logistic Regression, Decision Tree, Random Forest, and LightGBM** to predict churn.
4. **Evaluate model performance** and determine the best model for accurate predictions.
5. The insights from this project can help telecom companies **take proactive steps** like offering discounts, improving services, or addressing customer concerns to reduce churn and retain valuable customers.

## ****2. System Analysis****

### **Existing System:**

In the traditional approach, telecom companies use basic statistical methods and customer service feedback to analyze churn. However, these methods have limitations:

Limitations of the Existing System:

* Relies on **historical customer feedback** rather than predictive insights.
* Uses **rule-based approaches**, which may not generalize well to new data.
* Lacks the ability to **identify hidden patterns** in customer behavior.
* Cannot provide **real-time churn predictions** for proactive decision-making.

### **Proposed System**

To overcome the limitations of the existing system, this project introduces a **machine learning-based churn prediction model** that uses historical customer data to predict which customers are likely to leave.

**Features of the Proposed System:**

1. **Data-Driven Approach** – Uses machine learning to analyze customer behavior patterns.
2. **Predictive Modeling** – Identifies customers at risk of churning before they leave.
3. **Multiple Algorithms** – Compares different machine learning models to find the most effective one.
4. **Feature Importance Analysis** – Determines which factors contribute most to churn.
5. **Automated Decision Support** – Helps businesses take proactive actions, such as offering discounts or improving services.

### **Objectives**

* Analyze the dataset to understand the distribution and relationships of features.
* Preprocess the data to handle missing values, outliers, and categorical variables.
* Develop and evaluate multiple machine learning models to predict churn.
* Identify key factors influencing customer churn.

## 3. System Specification

### **3.1 Software Requirements**

* Programming Language: Python
* Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, lightgbm
* IDE: VS Code

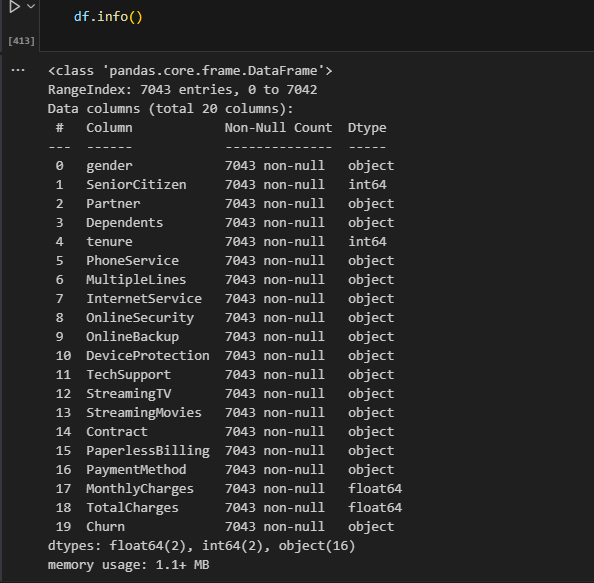
## ****Methodology****

## ****Dataset Overview****

### The dataset comprises customer information from a telecom company, including demographics, account details, and service usage. It contains 7,043 entries and 21 features.

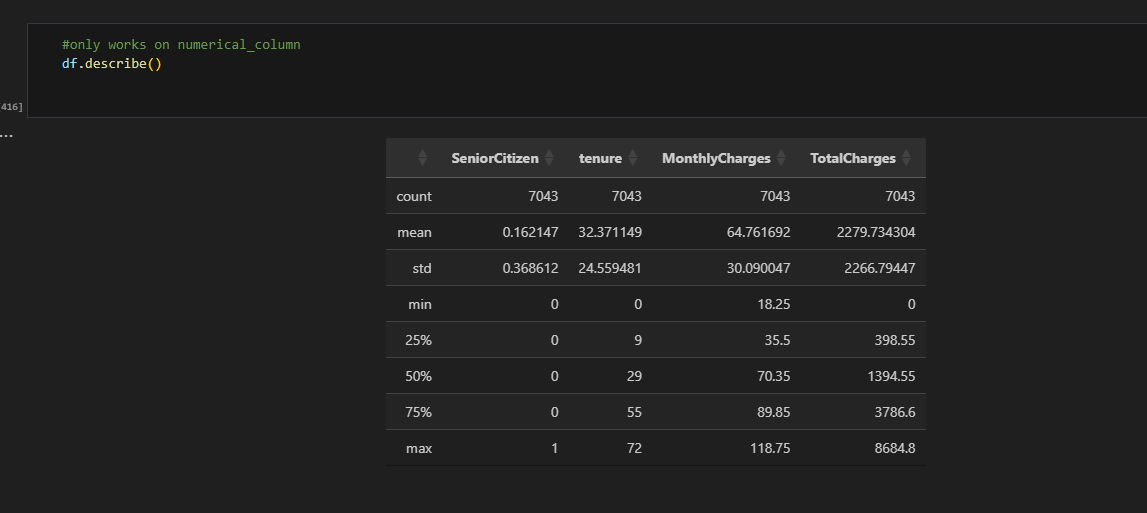
### **Key Features:**

* customerID: Unique identifier for each customer
* gender: Gender of the customer
* SeniorCitizen: Indicates if the customer is a senior citizen (1) or not (0)
* Partner: Whether the customer has a partner
* Dependents: Whether the customer has dependents
* tenure: Number of months the customer has stayed with the company
* PhoneService: Whether the customer has phone service
* MultipleLines: Whether the customer has multiple lines
* InternetService: Type of internet service (DSL, Fiber optic, No)
* OnlineSecurity: Whether the customer has online security
* OnlineBackup: Whether the customer has online backup
* DeviceProtection: Whether the customer has device protection
* TechSupport: Whether the customer has tech support
* StreamingTV: Whether the customer streams TV
* StreamingMovies: Whether the customer streams movies
* Contract: Contract term of the customer (Month-to-month, One year, Two year)
* PaperlessBilling: Whether the customer has paperless billing
* PaymentMethod: Payment method (Electronic check, Mailed check, Bank transfer, Credit card)
* MonthlyCharges: The amount charged to the customer monthly
* TotalCharges: The total amount charged to the customer
* Churn: Whether the customer churned (Yes or No)

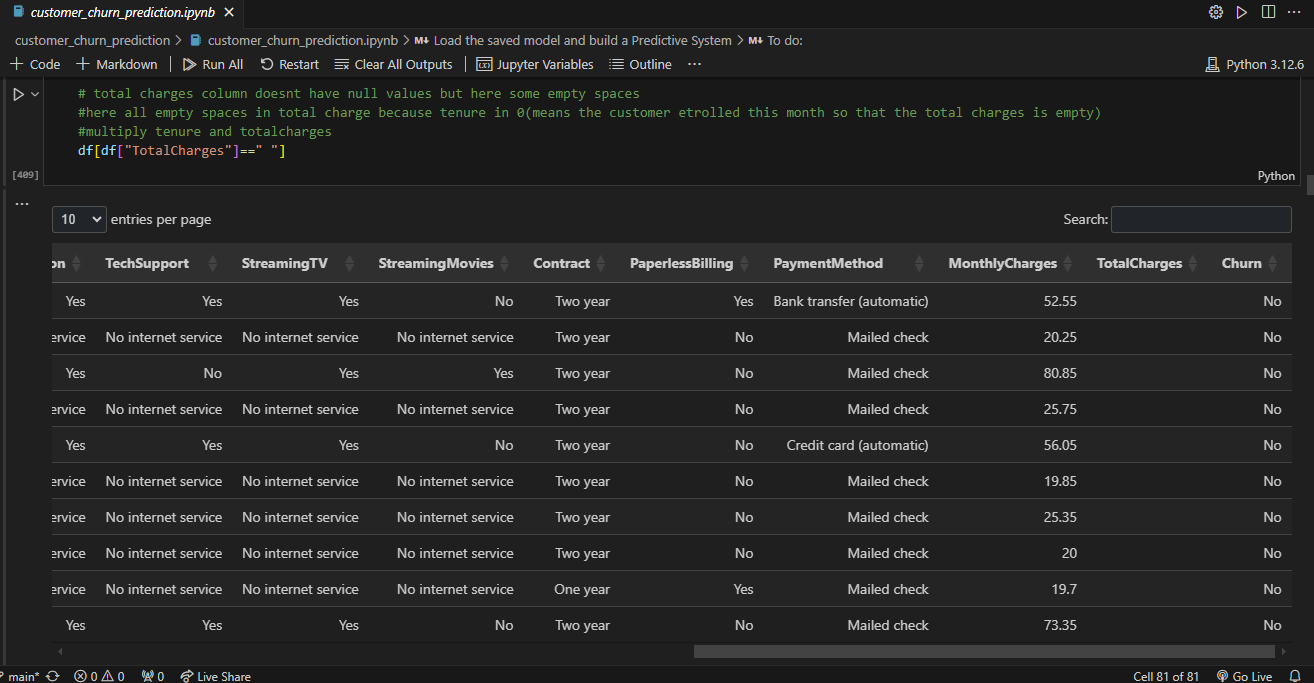


**"The image shows the dataset's structure, including the number of entries (7,043), column names, data types (integer, float, object), and non-null counts. This helps in understanding missing values and the overall dataset composition before preprocessing."**

### 

****

"The image shows a summary of the numerical columns in the dataset, with key statistics such as the count of non-null values, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile), 75th percentile (Q3), and the maximum value for each column."

****

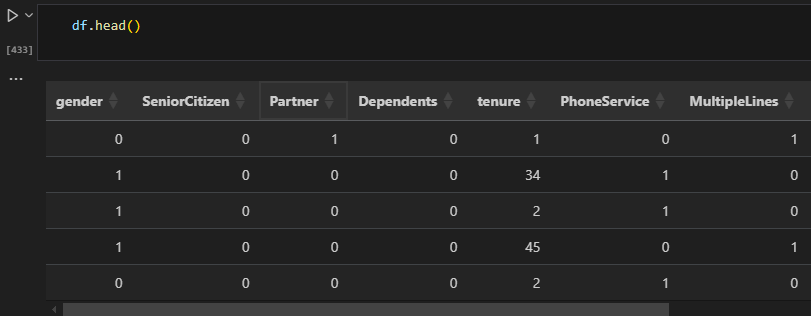
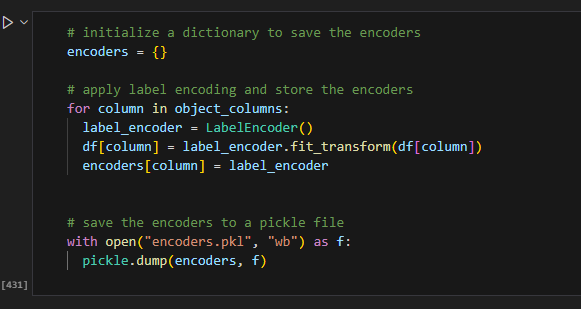
**“**The dataset contains some blank spaces (" ") instead of actual numerical values. These blank spaces act as **missing values** and need to be handled before model training. The code **filters rows where TotalCharges is an empty string (" ")** to check how many missing values exist.

## Data Preprocessing

### **Handling Missing Values:** The TotalCharges column had missing values, which were imputed using the median of the column.

### **Encoding Categorical Variables:** Categorical features were converted into numerical values using label encoding and one-hot encoding as appropriate.

### **Feature Scaling:** Numerical features were scaled using standardization to ensure uniformity.

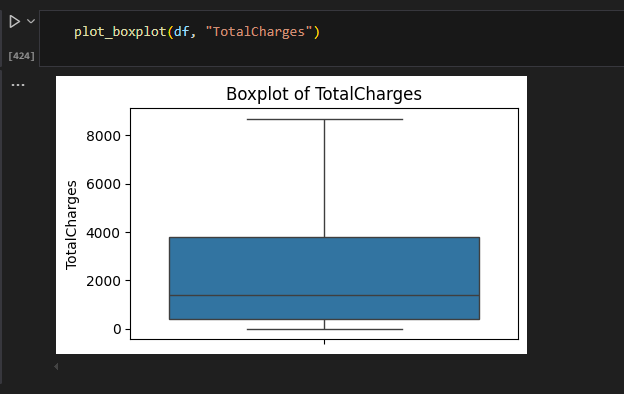
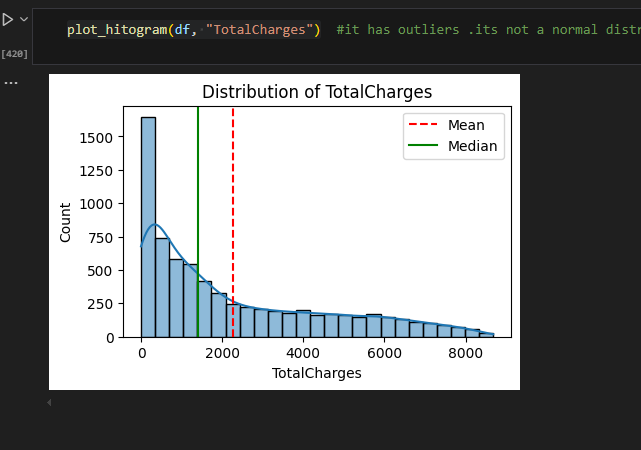


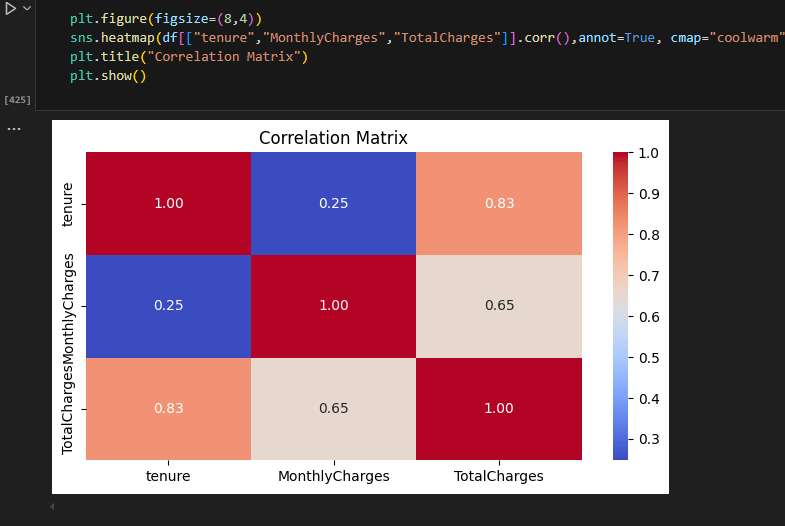
## Exploratory Data Analysis (EDA)

### EDA was conducted to understand the distribution and relationships of features:

### **Descriptive Statistics:** Calculated mean, median, mode, and standard deviation for numerical features.

* **Correlation Analysis:** Examined correlations between features to identify multicollinearity.
* **Visualization:**

1. **Histograms:** To observe the distribution of numerical features.
2. **Heatmap:** To understand the relationship between numerical features in the dataset, a correlation heatmap was created.
3. **Box Plots:** To detect outliers in numerical features.



## Model Development

### Several machine learning algorithms were employed to predict customer churn:

1. **Logistic Regression:** A baseline model to assess linear relationships between features and churn.
2. **Decision Tree Classifier:** To capture non-linear relationships and interactions between features.
3. **Random Forest Classifier:** An ensemble method to improve prediction accuracy and control overfitting.
4. **LightGBM:** A gradient boosting framework optimized for speed and efficiency.

Since the dataset is **imbalanced** (with more "No Churn" cases than "Yes Churn" cases), we applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the data before training the models.

## ****Module Description****

## Data Collection Module

* **Purpose:** Load and explore the dataset.
* **Description:**
* Reads the customer churn dataset using **pandas**.
* Displays basic information using **df.info()**, **df.describe()**, and **df.head()**.
* Identifies missing values and data types for preprocessing.

## Data Preprocessing Module

### **Purpose:** Clean and prepare the dataset for modeling.

### **Description:**

#### Handles **missing values** (e.g., converting TotalCharges to numeric).

#### Uses **SMOTE** to balance the dataset (fixes class imbalance).

#### Encodes categorical variables using **Label Encoding** and **One-Hot Encoding**.

#### Scales numerical data (if needed) to ensure consistency.

## 5.3 Exploratory Data Analysis (EDA) Module

### **Purpose:** Understand the relationships between features and visualize data patterns.

* Description:

#### Generates **histograms and bar graphs** for data insights.

* + - * Creates a **correlation heatmap** to analyze feature relationships.
      * Identifies outliers using **boxplots**.
      * Examines how different customer behaviors impact churn.

## 5.4 Feature Selection Module

### **Purpose:** Identify the most important features affecting churn.

* **Description:**
* Uses **correlation analysis** to remove redundant features.
* Extracts key features using **feature importance from Random Forest**.
* Ensures that only relevant information is used for model training.

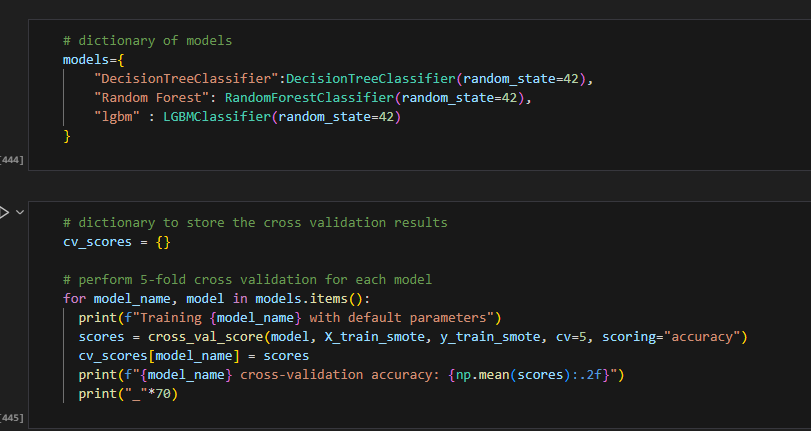
## 5.5 Model Training Module

### **Purpose:** Train multiple machine learning models for churn prediction.

* Description:
* Implements **four machine learning models:**

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. LightGBM

* Splits data into **training and testing sets**.
* Trains each model using the processed dataset.
* Applies **cross-validation** to improve generalization.



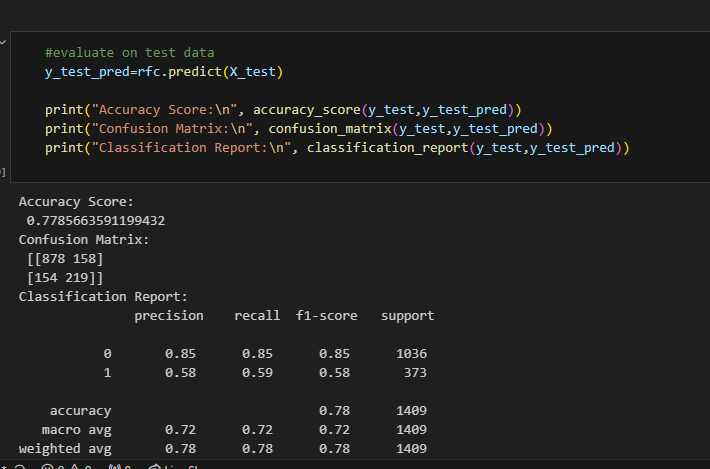
## 5.6 Model Evaluation Module

### **Purpose:** Compare model performance using evaluation metrics.

* Description:
* Evaluates models based on:

1. **Accuracy** – Measures overall correctness.
2. **Precision & Recall** – Checks churn detection effectiveness.
3. **F1-score** – Balances precision and recall.
4. **ROC-AUC Score** – Measures classification performance.

* Selects the **best-performing model** based on these metrics.

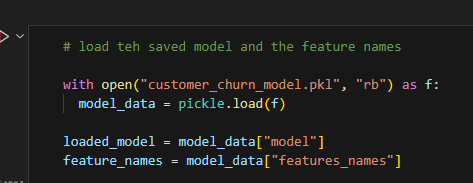


## 

## 5.7 Model Deployment & Prediction Module

### **Purpose:** Use the trained model to predict customer churn.

* Description:
* Takes new customer data as input.
* Uses the best model (Random Forest or LightGBM) to predict churn.
* Returns **"Churn" or "No Churn"** as the output.



# **6. Conclusion**

This project successfully demonstrates how machine learning can be used to predict customer churn. The implementation of multiple models allows for performance comparison, and the best model can be used to make accurate predictions.

## ****Future Enhancement****

* + 1. Integration with real-time customer data streams.
    2. Deployment as a web application for easy business use.
    3. Implementation of deep learning techniques for improved accuracy.
    4. Use of additional customer behavioral metrics for better insights.

## References

### **Dataset:** [**Telco Customer Churn Dataset**](https://github.com/anuselva1905/handle_missing_values/blob/d0a75bd7eff3cc522d9982ac8f6531b9eadf39a7/customer_churn_prediction/WA_Fn-UseC_-Telco-Customer-Churn.csv)**.**

1. **Github Repository:** [**Customer\_Churn\_Prediction**](https://github.com/anuselva1905/handle_missing_values/blob/main/customer_churn_prediction/customer_churn_prediction.ipynb)