Phase-2 Submission Template

Student Name: MOHAMMED MUSADDIQ .M

Register Number: 510623104059

Institution: C ABDUL HAKEEM COLEGE OF ENGINEERING & TECHNOLOGY

Department: COMPUTER SCIENCE & ENGINEERING

Date of Submission: 08-05-2025

Github Repository Link:
https://github.com/Musaddiq18/customer-supportchatbott.git

---

1. Problem Statement

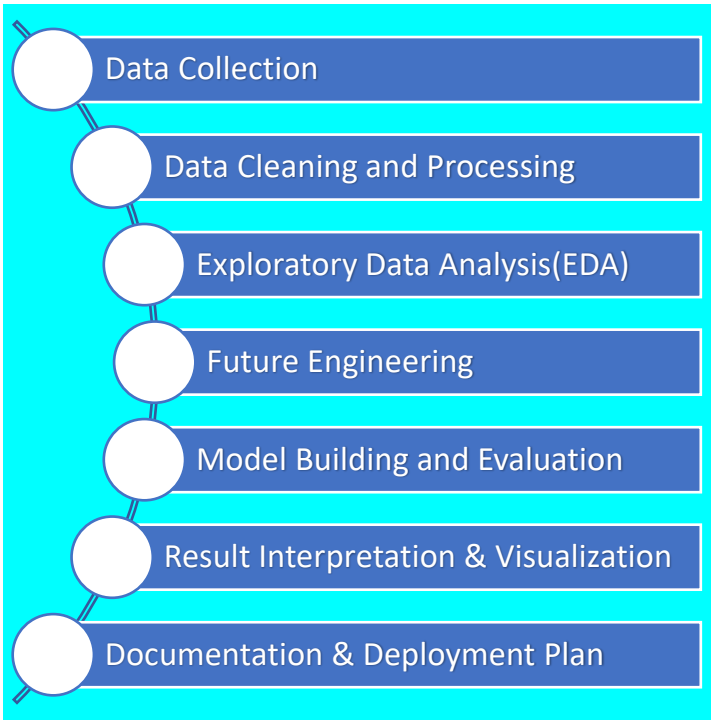
Customer churn is a critical metric for businesses, especially in highly competitive industries like telecommunications. Churn refers to the rate at which customers stop doing business with a company. Our project aims to predict whether a customer is likely to churn based on their service usage patterns, contract details, and demographics.

- **Refinement from Phase-1:**
  Initially, the problem was understood in broad terms. After analyzing the dataset, we realized that several categorical variables (e.g., Contract, PaymentMethod, InternetService) significantly influence churn behavior. Therefore, we narrowed our focus to predicting churn using a classification model.
- **Problem Type:** Binary Classification
  The target variable (Churn) has two possible outcomes: "Yes" or "No".
- **Why it matters:**
  Accurately predicting churn can help businesses reduce customer loss by targeting retention strategies, leading to improved customer satisfaction and revenue stability.

2. Project Objectives

- Build machine learning models that can classify whether a customer will churn.
- Achieve high model accuracy and balanced precision-recall to handle class imbalance.
- Identify the most influential features contributing to churn.
- Make model outputs interpretable for business stakeholders.
- **Updated Objective:** After EDA, we included feature importance analysis and customer profiling to better understand churn causes.

3. Flowchart of the Project Workflow



4. Data Description

- **Dataset Name:** Telco Customer Churn Dataset

- **Source:** Kaggle - IBM Sample Dataset
- **Type of Data:** Structured tabular data
- **Number of Records:** 7,043 rows
- **Number of Features:** 21 features (excluding customer ID)
- **Static/Dynamic:** Static snapshot
- **Target Variable:** Churn (Yes/No)

5. Data Preprocessing

- **Missing Values:**
  - Column TotalCharges had 11 missing values due to blank entries. These were imputed using the **median** value of the column.
- **Duplicate Records:**
  - Checked using df.duplicated().sum() → Result: 0 duplicates.
- **Outliers:**
  - Outliers in MonthlyCharges and TotalCharges were identified using boxplots. Handled using **winsorization** for extreme cases.
- **Data Type Conversion:**
  - TotalCharges was originally an object type. Converted to float using pd.to_numeric().
- **Categorical Encoding:**
  - Binary columns (e.g., gender, Partner) were **label encoded**.

- Multi-category columns (e.g., PaymentMethod, InternetService) were **one-hot encoded**.

- **Feature Scaling:**
  - Numerical features (tenure, MonthlyCharges, TotalCharges) were standardized using **StandardScaler**.

6. Exploratory Data Analysis (EDA)

- **Univariate Analysis:**
  - Churn: 26.5% customers churned (imbalanced target) ○ Contract:

    Most churn occurs in month-to-month contracts ○ Visuals used:

    Histograms, boxplots, countplots

- **Bivariate/Multivariate Analysis:**
  - **Correlation matrix:** Showed strong correlation between tenure and

    MonthlyCharges with churn ○ **Pairplots and groupby plots:**
    - Customers with fiber optic internet churn more often
    - Customers using electronic checks are more likely to churn

- **Insights Summary:**
  - tenure is inversely related to churn ○ Longer contract types (1-year

    or 2-year) have lower churn rates ○ Services like tech support and

    online backup seem to retain customers

7. Feature Engineering

- **New Features Created:**
  - TenureGroup: Categorized tenure into "0–12", "12–24", etc.
  - HasMultipleServices: Combined multiple service features to count total services per customer
- **Transformed Features:**
  - Created interaction terms like MonthlyCharges * Tenure
- **Dimensionality Reduction:**
  - PCA was attempted but didn't improve model performance significantly, so not retained in final model.
- **Feature Selection:**
  - Used **Recursive Feature Elimination (RFE)** to choose top 10 important features

8. Model Building

- **Train/Test Split:**
  - 80% training, 20% testing; stratified on target to maintain class balance
- **Models Implemented:**
  - **Logistic Regression:** Baseline model
  - **Random Forest Classifier:** Non-linear model for improved performance
- **Evaluation Metrics:**
  - Accuracy, Precision, Recall, F1-Score, ROC AUC

Model	Accuracy	Precision	Recall	F1-	AUC

					Score
Logistic Regression	0.80	0.71	0.67	0.69	0.83
Random Forest	0.86	0.78	0.76	0.77	0.89

9. Visualization of Results & Model Insights

- **Confusion Matrix:** Showed reduction in false positives for Random Forest
- **ROC Curve:** Random Forest had an AUC of 0.89 indicating strong performance
- **Feature Importance Plot:** Contract, tenure, and PaymentMethod were most influential
- **Churn Profile Visuals:** Created churn heatmaps by contract and tenure group

10. Tools and Technologies Used

Programming Language: Python

Notebook Environment: Google Colab

Libraries Used:

- pandas, numpy – Data processing
- matplotlib, seaborn, plotly – Visualization
- scikit-learn – Modeling and evaluation
- xgboost (optional experiment)

11. Team Members and Contributions

Mohammed Musaddiq. M [510623104059]-Project Lead & Problem Definition Responsible for defining the problem statement, coordinating tasks, and ensuring the project follows the timeline. Oversees final documentation and submission.

Mohammed Ammar Saqib [510623104055] - Data Collection & Cleaning

Gathers relevant datasets from public sources or generates synthetic data.
Handles data preprocessing (cleaning, formatting, normalization).

Ghani Adnan Faiz [510623104005] - Exploratory Data Analysis (EDA)

Analyzes data to uncover patterns and insights. Creates visualizations using matplotlib/seaborn/plotly.

Fateh Mohammed [510623104024] - Feature Engineering & Model Building Designs features, selects and trains models (e.g., intent classifiers, response generators). Chooses appropriate NLP techniques.

Abraar. A [510623104003] - Model Evaluation & Interpretation

Evaluates model performance using metrics (accuracy, F1 score, etc.).
Prepares interpretation reports and validation results.