Phase-2 Submission Template

Student Name : Vinu Utyhramoorthy B

Register Number : 720323106061

Institution : Akshaya College Of Engineering And Technology

Department : B.E.Electronics Communicatinos And Engineering

Date of Submission :15\05\2025

Github Repository Link: https://github.com/23ec061/black-.bs/upload

---

1. Problem Statement

Customer churn is a significant challenge for subscription-based businesses. Identifying customers who are likely to stop using a service allows companies to take proactive steps to retain them. This is a **classification problem**, where the goal is to predict whether a customer will churn (Yes/No) based on historical data. Solving this problem helps businesses increase customer retention, reduce revenue loss, and optimize marketing efforts.

2. Project Objectives

- Develop machine learning models to predict customer churn.
- Achieve high accuracy, precision, and recall in predictions.
- Uncover key factors contributing to churn through feature analysis.
- Create interpretable results that can be used by business stakeholders.

3. Flowchart of the Project Workflow

Data Collection

→ **Data Preprocessing**

→ **Exploratory Data Analysis**

→ **Feature Engineering**

→ **Model Building**

→ **Model Evaluation**

→ **Insights & Visualization**

→ **Deployment (optional)**

4. Data Description

- **Dataset Source**: [e.g., Kaggle: Telco Customer Churn Dataset]
- **Type**: Structured data
- **Records & Features**: ~7000+ records and ~20+ features
- **Static or Dynamic**: Static
- **Target Variable**: Churn (Yes/No)

---

5. Data Preprocessing

- Removed duplicates and handled missing values using imputation.
- Converted categorical variables using Label Encoding and One-Hot Encoding.
- Treated outliers with capping techniques.
- Normalized numerical features using StandardScaler.
- Ensured data type consistency throughout the dataset.

6. Exploratory Data Analysis (EDA)

- **Univariate**: Histograms and count plots revealed imbalances in target variable and distributions.
- **Bivariate/Multivariate**: Correlation heatmaps and pair plots helped identify relationships.
- **Insights**:
    - Tenure and monthly charges significantly influence churn.
    - enior citizens and customers without dependents show higher churn rates.

7. Feature Engineering

- Created binary flags for senior citizens, contract types, and online services.
- Binned tenure into groups (new, mid, long-term customers).
- Created a total service count feature.
- *Dropped redundant columns like customer ID*

8. Model Building

- **Models Used**: Logistic Regression, Random Forest Classifier
- **Why**: Logistic Regression offers interpretability; Random Forest offers performance and feature importance.
- **Split**: 80/20 train-test split with stratification
- **Evaluation Metrics**: Accuracy, Precision, Recall, F1-score
    - Random Forest performed best with accuracy ~85% and recall ~83%.

9. Visualization of Results & Model Insights

- Confusion Matrix showed good balance between FP and FN.
- ROC-AUC score was ~0.88 for Random Forest.
- Feature Importance showed “Contract Type”, “Monthly Charges”, and “Tenure” as top predictors.
- Visual comparisons confirmed Random Forest’s superior performance.

10. Tools and Technologies Used

- **Language**: Python
- **IDE**: Google Colab
- **Libraries**: pandas, numpy, matplotlib, seaborn, scikit-learn, XGBoost
- **Visualization**: matplotlib, seaborn, Plotly (optional)

.

11. Team Members and Contributions

Name	Contribution
B vinu Uthramoorthy	-Data Cleaning ,EDA
Hariharan	-Feature Engineering
Kishore priyadharshan	-Model Development
dharmesh	- Documentation and Reporting
Karthikeyan	-Research assistant