Phase-1 Submission

---

1. Problem Statement

Develop a machine learning model to predict customer churn by uncovering hidden patterns in customer behavior and usage data, enabling proactive retention strategies.

2. Objectives of the Project

- Identify key factors influencing customer churn.
- Build a machine learning model to predict churn.
- Uncover hidden patterns in customer behavior.
- Improve customer retention through insights.
- Evaluate model performance with relevant metrics.

3. Scope of the Project

This project focuses on using machine learning techniques to analyze customer data, identify patterns related to churn, and develop a predictive model. It includes data preprocessing, feature selection, model training and evaluation, and generating actionable insights for improving customer retention. The project is limited to available historical data and does not include real-time deployment or integration with live systems.

4. Data Sources

- **Customer Demographics** – Age, gender, location, income level.
- **Account Information** – Subscription type, tenure, contract details, billing method.
- **Usage Data** – Frequency of service use, session durations, feature usage.
- **Customer Support Logs** – Number and type of support interactions, resolution times.
- **Transaction History** – Payment patterns, late payments, service upgrades/downgrades.
- **Surveys/Feedback** – Customer satisfaction scores, Net Promoter Score (NPS).
- **Churn Labels** – Historical data indicating whether a customer churned or not.

5. High-Level Methodology

Data Collection – Gather historical customer data from various sources (e.g., CRM, billing systems).

Data Preprocessing – Clean, format, and handle missing values; encode categorical variables.

Exploratory Data Analysis (EDA) – Identify trends, patterns, and correlations related to churn.

Feature Selection/Engineering – Select or create the most relevant features for modeling.

Model Development – Train machine learning models (e.g., Logistic Regression, Random Forest, XGBoost).

Model Evaluation – Assess performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Insights & Interpretation – Analyze model output to extract business-relevant insights.

Recommendations – Provide strategic actions to reduce churn based on findings.

6. Tools and Technologies

Programming Language:

- Python (primary language for data analysis and modeling)

Data Handling & Analysis:

- **Pandas**, **NumPy** – Data manipulation and analysis
- **SQL** – Data extraction from databases (if needed)

Data Visualization:

- **Matplotlib**, **Seaborn**, **Plotly** – Visualizing data trends and patterns

Machine Learning:

- **Scikit-learn** – Traditional ML algorithms (e.g., Logistic Regression, Random Forest)
- **XGBoost**, **LightGBM** – Advanced ensemble methods for improved accuracy

Model Evaluation:

- **Scikit-learn metrics** – Precision, Recall, F1-score, ROC-AUC

Jupyter Notebook / Google Colab:

- For development, experimentation, and documentation

Version Control (optional):

- **Git**, **GitHub** – To track changes and collaborate

7.Team Members and Roles

1. **Project Manager**
   - Oversees project planning, timelines, and coordination among team members.
2. **Data Analyst**
   - Performs exploratory data analysis (EDA), visualizations, and derives initial insights.
3. **Data Engineer**
   - Manages data collection, cleaning, integration, and ensures data pipeline reliability.
4. **Machine Learning Engineer**
   - Builds, tunes, and evaluates ML models for churn prediction.
5. **Domain Expert (optional but valuable)**
   - Provides business context, helps interpret results, and ensures model aligns with real-world needs.
6. **DevOps Engineer (optional for deployment phase)**
   - Deploys the model into production and manages cloud or API infrastructure.