**Phase-1 Submission Template**

**Student Name:**JANANI.V

**Register Number:**712523121007

**Institution:** PPG Institute of Technology

**Department:**Bio medical Engineering

**Date of Submission:25.04.2025**

**1.Problem Statement**

Predicting customer churn using machine learning is a powerful approach to uncover hidden patterns in customer behavior, enabling businesses to proactively retain valuable clients. Here's a comprehensive guide to help you embark on this project

**2.Objectives of the Project**

1. Develop a predictive model to identify customers at high risk of churning using historical and behavioral data.

2. Uncover hidden patterns and insights from customer data that contribute to churn, using data mining and machine learning techniques.

3. Improve customer retention by enabling proactive intervention strategies based on model predictions.

4. Optimize business decision-making by identifying key factors influencing churn (e.g., customer service interactions, usage patterns, pricing issues).

5. Compare multiple machine learning algorithms (e.g., logistic regression, decision trees, random forests, XGBoost, etc.) to determine the most effective model for churn prediction.

**3.Scope of the Project**

1.Data Collection and preprocessing

2.Exploratory data analysis

3.Model development

4.feature importances and pattern recognition

5.Model evaluation and validation

**4.Data Sources**

1. Public Datasets (for academic or prototype purposes):

                - Telco Customer Churn Dataset (Available on Kaggle)

                - Contains customer demographics, services signed up for, account information, and churn label.

2. Enterprise Customer Databases (for real-world application)

              -CRM systems (e.g., Salesforce, HubSpot)

             - Billing systems and usage logs

             - Customer support logs (e.g., ticket history, chat logs)

3. Data Fields Typically Included:

               -Customer ID

              - Demographics (age, gender, region)

              -Service usage data (call minutes, internet usage, subscription type)

              - Billing details (payment method, monthly charges, total charges)

**5.High-Level Methodology**

**● Data Collection**

1. Gather relevant customer data, including:

- Demographic information

- Transaction history

- Usage patterns

- Feedback and complaints

2. Ensure data quality and integrity

**● Data Cleaning**

Handle missing values, remove duplicates, and correct inconsistencies.

● **Exploratory Data Analysis (EDA)**

**Visualizations:** Use tools like Seaborn and Matplotlib to create histograms, box plots, and correlation matrices.

**Statistical Analysis:** Identify significant differences between churned and retained customers using statistical tests.

**● Feature Engineering**

**Identify relevant features:**

- Extract meaningful features from data

 - Create new features (e.g., usage metrics, customer tenure)

**Select most informative features:**

- Use techniques like correlation analysis, mutual information, or recursive feature elimination

**● Model Building**

**Baseline Models:**

Start with Logistic Regression and Decision Trees to establish a performance baseline.

**Advanced Models:**

● Random Forest: Effective for handling large datasets and capturing non-linear relationships.

● XGBoost: Known for its performance and speed in classification tasks.

● Ensemble Methods: Combine multiple models to improve accuracy.

**● Model Evaluation:**

Use metrics like Accuracy, Precision, Recall, F1-Score, and Area Under the ROC Curve

● **Deployment**:

Integrate the model into a business application or dashboard for real-time predictions.

**6.Tools and Technologies**

**Machine Learning Frameworks**

● Scikit-learn: A popular Python library for machine learning.

● TensorFlow: An open-source machine learning framework developed by Google.

● PyTorch: An open-source machine learning framework known for its ease of use and flexibility.

**Data Preprocessing and Analysis**

● Pandas: A Python library for data manipulation and analysis.

● NumPy: A Python library for numerical computing. ● Matplotlib and Seaborn: Python libraries for data visualization.

**Data Storage and Management**

● Relational databases: MySQL, PostgreSQL, or Oracle for storing structured data.

● NoSQL databases: MongoDB, Cassandra, or Redis for handling large amounts of unstructured or

semi-structured data.

● Cloud storage: AWS S3, Google Cloud Storage, or Azure Blob Storage for storing and managing large datasets.

**Model Deployment and Serving**

● Cloud platforms: AWS SageMaker, Google Cloud AI Platform, or Azure Machine Learning for deploying and managing machine learning models.

● Containerization: Docker for packaging and

deploying models in containers.

● Model serving: TensorFlow Serving, AWS

SageMaker, or Azure Machine Learning for serving models in production environments.

**Other Tools**

● Jupyter Notebook: An interactive environment for data exploration, visualization, and model

development.

● Apache Spark: A big data processing engine for handling large-scale data processing and machine learning tasks.

● Feature engineering tools: Featuretools or tsfresh for automating feature engineering

**Team Members Roles**

Nidhidharshini.K :Problem Statement and Objectives of the project

Janani .V :Scope of the project

Charumathi .K: Data sources

Mohana.A:High-level methodology

Iniyavarshini .S :Tools and Technologies