Phase-1

Student Name: Dinesh D

Register Number: 712523205701

Institution: PPG Institute of Technology

Department: B. tech Information Technology

Date of Submission:

1.Problem Statement

Customer churn is a significant issue for many businesses, as it leads to a loss of revenue and increases the cost of acquiring new customers. In industries like telecom, banking, and subscription-based services, retaining existing customers is more cost-effective than acquiring new ones. This project addresses the real-world problem of customer churn by leveraging machine learning to uncover hidden patterns in customer behavior and identify those at high risk of leaving. By predicting churn early, businesses can take proactive steps to improve customer retention and satisfaction.

2.Objectives of the Project

Customer churn is a significant issue for many businesses, as it leads to a loss of revenue and increases the cost of acquiring new customers. In industries like telecom, banking, and subscription-based services, retaining existing customers is more cost-effective than acquiring new ones. This project addresses the real-world *problem of customer churn by leveraging machine learning to uncover hidden patterns in customer behavior and identify those at high risk of leaving. By predicting churn early, businesses can take proactive steps to improve customer retention and satisfaction.*

3.Scope of the Project

This project will focus on analyzing features such as customer demographics, subscription plans, payment methods, service usage, and customer support interactions. It will explore different machine learning models suitable for binary classification tasks. The scope is limited to publicly available datasets, and the deployment, if done, will be in the form of a basic web application or dashboard. Advanced deep learning methods or real-time prediction systems are beyond the scope of this project due to time and resource constraints.

4.Data Sources

The dataset used in this project is the Telco Customer Churn dataset, which is publicly available on Kaggle. This dataset contains various customer attributes, including service details, contact information, payment methods, and churn status. The dataset is static, meaning it will be downloaded once and used throughout the project without real-time updates. It provides a well-balanced mix of categorical and numerical data suitable for machine learning applications.

Kaggle: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Telco customer chum: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

5.High-Level Methodology

● ***Data Collection***

The data will be obtained by downloading the Telco Customer Churn dataset from the Kaggle platform and then loaded into a Python environment using the panda's library.

● ***Data Cleaning***

Potential issues such as missing values in columns like Total Charges, duplicate entries, and inconsistent data formats will be identified and addressed through imputation, removal, or transformation.

● ***Exploratory Data Analysis (EDA)***

EDA will be conducted using various statistical summaries and visualizations such as bar plots, box plots, correlation heatmaps, and distribution plots to identify trends, patterns, and relationships within the data.

● ***Feature Engineering***

New features may be created by grouping existing columns (e.g., tenure buckets) or combining data to create interaction terms. Categorical variables will be encoded appropriately, and numerical values may be standardized for better model performance.

● ***Model Building***

Multiple classification models such as Logistic Regression, Random Forest, XGBoost, and Gradient Boosting Classifier will be trained and compared. These models are chosen due to their effectiveness in handling classification tasks and ability to provide interpretable results.

- ***Model Evaluation***

Model performance will be evaluated using metrics such as Accuracy, Precision, Recall, F1 Score, and AUC-ROC. Cross-validation techniques will be used to ensure the models generalize well to unseen data.

- ***Visualization & Interpretation***

Key findings and model predictions will be visualized using bar charts, pie charts, SHAP value plots, and feature importance graphs to interpret the contribution of each feature towards churn prediction.

- ***Deployment***

If time and resources permit, the project will be deployed as a web-based application using tools like Streamlit. The deployment will allow users to input customer data and receive churn predictions along with explanations.

6.Tools and Technologies

Programming Language: Python.
Notebook/IDE: Google Colab and Jupyter Notebook
Libraries: pandas and numpy. For data visualization, matplotlib, seaborn, and plotly , scikit-learn, xgboost, lightgbm.
Optional Tools for Deployment: Streamlit or Flask

7.Team Members and Roles

NAME	*ROLE*	
S.Thilshan	*Project Lead*	*Responsible for overall project planning,data processing and model handling*
P.Deeksha	*Data Analyst*	*Handles exploratory data analysis (EDA), feature engineering, and insights generation.*
D.Dinesh	*Visualization Specialist*	*Creates visualizations and interprets model results using tools like SHAP and Plotly.*
B.Srimathi	*Deployment Engineer*	*Develops and manages deployment using tools like Streamlit or Flask, and prepares final presentation.*
B.Tamilarasan	*Documentation & Reporting Lead*	*Prepares project documentation, creates presentations, and compiles reports for final submission.*