Phase-1

Student Name :SurendhaR K

Register Number :511323104105

Institution : Kingston Engineering College

Department : CSE

Date of Submission: 25-04-2025

TITLE : Predicting customer churn using machine learning to uncover hidden patterns

Problem Statement

In today's competitive and data-driven economy, customer retention has emerged as a pivotal concern for companies, particularly those operating in subscription-based or service-intensive industries such as telecommunications, financial services, and media streaming platforms. The term "customer churn" refers to the phenomenon where customers discontinue their relationship with a business over a period. High churn rates directly impact an organization's revenue and operational stability since acquiring new customers is more expensive than retaining existing ones.

The traditional approaches to predicting churn—mainly based on heuristic or rule-based systems—often lack the sophistication to accommodate dynamic and evolving customer behaviors. Such models are usually limited in their predictive power and fail to offer actionable insights.

In response to this challenge, our project presents a machine learning-driven approach to customer churn prediction. By utilizing historical usage data, demographic attributes, and service interaction patterns, our aim is to construct a robust, scalable model that can accurately identify customers who are at a higher risk of churning. This predictive capability enables businesses to proactively implement retention strategies, such as customized offers or loyalty programs, thereby improving both customer satisfaction and profitability.

Objectives

The primary aim of this project is to develop a machine learning model that can predict customer churn with high accuracy. The following are the specific objectives:

- **Model Development**: Create an efficient and scalable churn prediction model using supervised learning techniques.
- **Feature Identification**: Analyze the dataset to uncover significant factors contributing to customer churn.
- **Feature Engineering**: Apply data transformation and feature creation techniques to optimize model performance.
- **Insight Generation**: Visualize data patterns and model outcomes to derive actionable insights.
- **Deployment (Optional)**: Integrate the model within an interactive web application or dashboard to support real-time business decision-making.

Scope

Inclusions

- Thorough analysis of customer demographic and service usage patterns.
- Data preprocessing steps including cleaning, encoding, and transformation.
- Implementation of multiple classification algorithms such as Logistic Regression, Random Forest, and Gradient Boosting.
- Model evaluation using industry-standard metrics and visual aids.
- Presentation of analytical findings through dashboards and interactive visualizations.

Limitations

- The project is based on a single static dataset, specifically the Telco Customer Churn dataset from Kaggle.
- Real-time deployment and integration with live data sources are not within the scope.
- Model interpretability is limited to tools like SHAP; deep explainable AI platforms are not included.

Data Sources

The data utilized for this project is sourced from Kaggle's **Telco Customer Churn** dataset. It comprises:

- **Demographic Details**: Gender, Senior Citizen status, Partner/Dependents.
- **Account Information**: Tenure, Contract Type, Paperless Billing, Payment Method, Monthly/Total Charges.
- **Service Information**: Type of internet service, online security, tech support, streaming services, and phone-related services.
- **Target Variable**: Churn, which indicates whether the customer has left the company.

This dataset is ideal for binary classification problems and is well-suited for supervised machine learning approaches.

High-Level Methodology

The overall approach is divided into several stages:

Data Collection

- The dataset is downloaded from Kaggle in CSV format.
- It is imported into a Python environment using pandas for initial exploration.

Data Cleaning

- Null or missing values, especially in columns such as TotalCharges, are handled using imputation techniques.
- Categorical variables are transformed into numerical representations using label encoding and one-hot encoding.
- Irrelevant or redundant features (like customer IDs) are dropped.

Exploratory Data Analysis (EDA)

- Visual tools such as matplotlib, seaborn, and pandas_profiling are used.
- Churn trends are examined with respect to:
    - Contract Type (monthly, one-year, two-year
    - Monthly Charges
    - Customer Tenure
- Cross-tabulations and correlation heatmaps help to identify feature dependencies and redundancies.

Feature Engineering

- New derived metrics such as Average Charges per Tenure are created.
- Skewed numeric data is normalized using techniques like log transformation.
- Encoding methods such as one-hot encoding are used for non-numeric data fields.

Model Building

- A variety of models are trained to benchmark performance:
    - Logistic Regression
    - Decision Trees
    - Random Forest
    - XGBoost
- Hyperparameter tuning is conducted using GridSearchCV.

Model Evaluation

- Evaluation metrics include:
    - **Accuracy** – Overall correctness of the model.
    - **Precision and Recall** – Useful in imbalanced datasets.
    - **F1-Score** – Harmonic mean of precision and recall.
    - **ROC-AUC** – Measures ability to distinguish between churn and non-churn.
- K-Fold Cross-Validation is employed for robust performance assessment.

Visualization and Interpretation

- **Feature Importance** is plotted for tree-based models.

- **SHAP (SHapley Additive exPlanations)** is used for interpreting model decisions on an individual level.
- Results are summarized with graphical representations for easier business understanding.

Deployment

- A prototype interface is developed using **Streamlit** or **Flask**.
- It allows users to input new customer data and receive churn predictions in real-time.
- Enhances accessibility and business usability of the churn model.

Tools and Technologies

Category	Tools/Frameworks
Programming Language	Python 3.x
IDE	Google Colab, Jupyter Notebook
Data Handling	pandas, numpy
Visualization	matplotlib, seaborn, plotly
Machine Learning	scikit-learn, xgboost, shap
Deployment (optional)	streamlit, flask
Version Control	GitHub (for team collaboration and updates)

Team Members and Role Allocation

Name	Role	Responsibilities
Emayavan M	Data Scientist	Data wrangling, model building, tuning, and evaluation.

Sri vigneshwar R	Backend Developer	Development of deployment dashboard and API integration (Streamlit/Flask).
Niranjan R	Data Analyst	Conduct EDA, visualize trends, interpret results, and generate analytical insights.
Surendhar K	Documentation Lead	Compile reports, prepare presentations, manage GitHub updates and team workflow.