Phase-1SubmissionTemplate

Student Name: M Soorya Prakash

Register Number: 410723104048

Institution: Dhanalakshmi College of Engineering

Department: Computer Science

Date of Submission: 28.04.2025

PREDICTING CUSTOMER CHURN USING MACHINE LEARNING TO UNCOVER HIDDEN PATTERN

1. Problem Statement

Customer churn is a critical concern for businesses, directly impacting revenue and growth. With the increasing availability of customer interaction and transaction data, it has become feasible to predict churn before it happens. Traditional methods often fail to capture subtle patterns leading to customer attrition.

This project aims to develop a machine learning-based solution that accurately predicts customer churn by leveraging historical data and identifying previously undetected behavioral patterns. The model will help businesses:

- Detect high-risk customers before they churn.
- Understand the key factors contributing to churn.
- Design targeted retention strategies based on data-driven insights.

2. Objectives of the Project

- To predict customer churn by building a machine learning model that learns from historical customer data and accurately forecasts whether a customer is likely to leave.
- To analyze behavioral trends and interaction history to detect early warning signs of churn—such as reduced usage, frequent complaints, or late

payments.

- To identify and rank the key features (such as contract type, usage pattern, customer support interaction, etc.) that most significantly influence a customer's decision to stay or leave.
- To empower businesses with actionable insights so they can design personalized retention campaigns, improve customer experience, and proactively address churn triggers.

3. Scope of the Project

- This project will focus on creating a generalized machine learning framework that can be applied across various industries, including telecommunications, e-commerce, banking, insurance, and SaaS companies. It involves:
- Extracting and analyzing structured data from customer relationship management (CRM) systems, including demographic information, billing history, service usage, and support tickets.
- Detecting behavior patterns and segmenting customers based on their likelihood to churn.
- Applying statistical and machine learning techniques to model and interpret these patterns.
- Supporting decision-makers in understanding which interventions are likely to retain at-risk customers.
- Enabling real-time predictions through deployment on an interactive platform or dashboard.
- This research could significantly impact business sustainability by reducing customer attrition, increasing lifetime value (CLV), and improving the overall quality of customer service.

4. Data Sources

- The project will utilize various data sources to train and validate the churn prediction models:
- Public datasets such as the “Telco Customer Churn” dataset from Kaggle,

which includes features like tenure, monthly charges, and contract types.

- Company-specific CRM data, which may include customer demographics, transactional logs, service history, and complaint tickets.
- Behavioral data like login frequency, product usage patterns, and time between transactions.
- Support interaction logs, which can highlight customer frustration or dissatisfaction.
- Synthetic data may be generated using techniques like SMOTE to handle class imbalance between churned and non-churned customers.
- All data will be anonymized and used in compliance with privacy and data protection regulations (e.g., GDPR).

5. High-Level Methodology

1.Data Collection

Data will be acquired from internal company databases or open-source platforms like Kaggle. APIs may be used to pull live data from CRM tools, support systems, and web activity logs. Web scraping may be applied to collect review and feedback data where permitted.

2.Data Cleaning

Missing values will be addressed using appropriate imputation techniques. Outliers and noise will be handled through normalization and filtering. Categorical variables (e.g., contract type, payment method) will be encoded using One-Hot or Label Encoding. Feature scaling methods like Standardization or MinMax Scaling will be used for continuous variables.

3.Exploratory Data Analysis (EDA)

EDA will be used to identify trends and correlations in customer behavior. Techniques include:

- **Histograms and Boxplots** to understand distribution of numerical features.
- **Correlation matrices** to find feature interdependencies.
- **Churn ratio visualizations** across demographics, tenure, contract type, etc.
- **Time-series plots** to understand churn trends over time.

4.Feature Engineering

New features will be created to capture subtle behavioral signals:

- **Tenure buckets**, service upgrade patterns, average support call duration, and days since last interaction.
- **Customer segmentation features**, such as loyalty score and engagement frequency.
- Text data (e.g., complaints) may be vectorized using **TF-IDF or embeddings**.
- Dimensionality reduction techniques (like PCA) will be used to simplify the feature space.

5.Model Building

We will begin with simple, interpretable models like:

- **Logistic Regression**, **Decision Trees**, and **Random Forests**. Advanced models will include:
- **XGBoost**, **CatBoost**, and **LightGBM** for improved accuracy.
- **Neural networks**, including **deep learning architectures**, may be used on larger datasets or time-series inputs.
- **Ensemble models** to combine the strengths of multiple algorithms.

6.Model Evaluation

Performance will be measured using:

- **Accuracy, Precision, Recall, and F1-Score**.
- **ROC-AUC curve** to evaluate classification thresholds.
- **Confusion matrix** to assess false positives and false negatives.
- **Cross-validation** to ensure model generalizability.

7.Visualization and Interpretation

Business stakeholders need to understand model outputs clearly. To that end:

- **SHAP or LIME** will be used to explain individual predictions.
- Dashboards will show churn likelihoods by customer segment.
- Feature importance charts and churn driver summaries will support actionable insights.
- Interactive reports will highlight "high-risk" customers and reasons behind their churn likelihood.

8. Deployment

The model will be integrated into a real-time dashboard or web application using:

- **Streamlit**, **Flask**, or **Gradio**. This system will allow non-technical users to:
- Upload customer data,
- Get churn predictions, and

- Receive tailored recommendations for customer retention strategies.

6.Tools and Technologies

1. Programming Language

- Python will be the primary language due to its powerful libraries for data science and machine learning.

2. Notebook/IDE

- Google Colab for cloud-based experiments.
- Jupyter Notebook for local development and EDA.
- VS Code for writing modular Python scripts.

3. Libraries

- Pandas, NumPy: Data manipulation and transformation.
- scikit-learn: Model building, training, and evaluation.
- XGBoost, LightGBM, CatBoost: Advanced boosting models.
- TensorFlow/Keras, PyTorch: Neural networks.
- Matplotlib, Seaborn, Plotly: Data visualization.
- SHAP, LIME: Model interpretation.
- Streamlit, Flask, Gradio: For building user interfaces.

6. Optional Tools for Deployment –

- **Gradio**: For building lightweight and intuitive prediction interfaces.
- **Docker**: For containerizing the application and ensuring environment consistency across systems.
- **AWS/GCP/Azure**: For cloud hosting and scalability.

7. Team Members and Roles

NAMES	ROLE	RESPONSIBILITY
M Soorya Prakash	Leader	Data Collection and Cleaning
Murugesh M	Member	Data visualization and Interpretation
Logesh R	Member	Exploratory Data Analysis
Magesh V	Member	Model evaluation
Antony Sanjay P	Member	Model Building