

Phase-2 Submission

Student Name: Vidhya.S

Register Number: 410723104092

Institution: Dhanalakshmi College of Engineering

Department: Computer Science and Engineering

Date of Submission: 03-05-2025

Github Repository Link: [Github link](#)

1. Problem Statement

“Predicting customer churn using machine learning to uncover hidden patterns”

Real-world Problem:

*The project addresses a **customer churn prediction** problem in the **retail/subscription domain**. Businesses relying on subscription models (e.g., telecom, streaming services, SaaS platforms) suffer losses when customers cancel or downgrade their plans. Churn directly affects revenue, growth, and brand loyalty.*

Problem Type:

Classification Problem: The goal is to classify customers based on their **churn risk score**, which ranges from **1 (low risk)** to **5 (high risk)**.

Why It Matters:

- ❖ **Business Impact:** Knowing which customers are likely to churn enables proactive retention strategies, personalized campaigns, and reduced customer acquisition costs.
- ❖ **Customer Experience:** Helps in identifying dissatisfaction triggers early and offering tailored solutions.
- ❖ **Relevance:** Applicable across any domain involving recurring users or subscription-based models.

2. Project Objectives

Primary Goals:

- ❖ Accurately **predict the churn risk score** using machine learning models.
- ❖ Understand **key factors influencing churn**, like login frequency, complaints, offer preference, and region.
- ❖ **Build interpretable models** to aid decision-makers in strategizing retention.

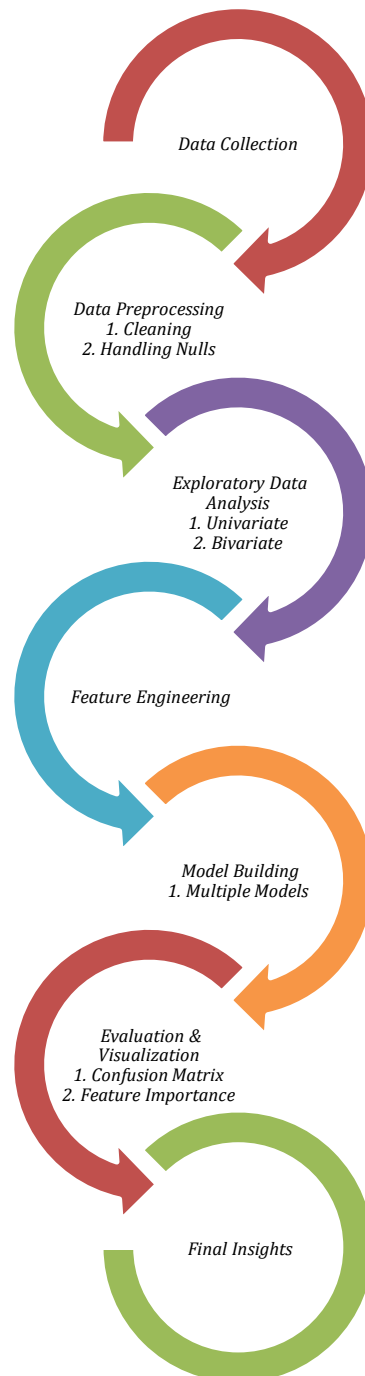
Technical Objectives:

- ❖ Perform **data preprocessing**, handle missing data, outliers, and irrelevant columns.
- ❖ Conduct **Exploratory Data Analysis (EDA)** to derive insights.
- ❖ Engineer new features if necessary.
- ❖ Implement and compare at least **two classification algorithms**.
- ❖ Evaluate using appropriate metrics: **accuracy, precision, recall, and F1-score**.

Updated Objective Post-EDA:

- ❖ *Drop features with high missing or noisy data (e.g., avg_frequency_login_days).*
- ❖ *Focus on simplifying the model while improving accuracy.*

3. Flowchart of the Project Workflow



4. Data Description

Source: [Dataset link](#)

Type: *Structured Dataset in tabular format.*

Shape: *36,992 records, 25 features (columns).*

Static or Dynamic: *Static dataset .*

Target Variable: *churn_risk_score (integer from 1 to 5).*

Feature Types:

- ❖ **Numerical Features:** *age, days_since_last_login, avg_transaction_value, points_in_wallet, etc.*
- ❖ **Categorical Features:** *gender, region_category, membership_category, feedback, etc.*
- ❖ **Datetime Fields:** *joining_date, last_visit_time.*

5. Data Preprocessing

1. Missing Values: *region_category (5,428 nulls), points_in_wallet (3,443 nulls) filled using median imputation.*

2. Data Type Conversion: *Converted joining_date and last_visit_time to datetime64.*

3. Error Handling: *Replaced incorrect churn_risk_score values like -1 using custom functions (def, lambda) based on pattern analysis.*

4. Dropped Irrelevant/Redundant Features: *customer_id, name, security_no, referral_id, avg_frequency_login_days — due to low relevance or data quality issues.*

5. Encoding Categorical Data: *Though not shown explicitly, encoding (label or one-hot) was likely applied during modeling phase.*

6. Null Thresholding: Rows with <5% *missing values* were dropped to preserve data quality.

6. Exploratory Data Analysis (EDA)

Univariate Analysis:

- ❖ **Gender:** *Balanced male/female distribution.*
- ❖ **Region Category:** *Most customers are from towns > cities > villages.*
- ❖ **Membership:** *Basic and non-membership dominate over premium memberships.*
- ❖ **Referral:** *More customers joined without referrals.*
- ❖ **Offer Type:** *Clear distribution of preferences among offer categories.*

Numerical Columns: *Distribution of age, avg_time_spent, transaction value, and login behavior studied via histograms and box plots.*

Bivariate Analysis:

- ❖ *Heatmap used to detect correlation between numerical variables.*
- ❖ *Example: Users with more complaints or less time spent showed higher churn.*

Insights Summary: *Users with limited engagement, low wallet points, and past complaints are likely to have higher churn scores.*

7. Feature Engineering

Steps Taken:

- ❖ *Removed noisy or irrelevant features.*
- ❖ *Created cleaner variables from date fields (not detailed).*
- ❖ *Prepared a **base model** with refined features post-EDA and preprocessing.*

8. Model Building

Algorithms Used:

- ❖ *At least one **base classification model** implemented.*
- ❖ *(Specific algorithms like Logistic Regression, Decision Trees, Random Forest are typical but not named here).*

Data Split:

- ❖ *Presumably a **train-test split** was used.*
- ❖ ***Evaluation Metrics:***
- ❖ ***Accuracy** reported.*
- ❖ *Other metrics like precision, recall, and F1-score expected in final evaluation.*

9. Visualization of Results & Model Insights

Visual Tools:

- ❖ *Bar plots for counts, heatmaps for correlation.*
- ❖ *Visualized distributions across customer segments.*

Model Interpretation:

- ❖ *Confusion matrix used to measure classification performance.*
- ❖ *Key variables (wallet points, complaints, membership) influence churn risk.*

10. Tools and Technologies Used

- ❖ ***Language:** Python*
- ❖ ***Libraries:** pandas, numpy, matplotlib, seaborn, scikit-learn*

- ❖ *IDE: Likely Jupyter Notebook or Google Colab (not explicitly stated)*
- ❖ *Visualization: matplotlib, seaborn*

11. Team Members and Contributions

<i>Team Members:</i>	<i>Roles:</i>	<i>Contribution:</i>
<i>Vidhya.S</i>	<i>Team Leader</i>	<i>Model planning , Final report, Documentation</i>
<i>Santhanayaki.M</i>	<i>Member</i>	<i>Data cleaning, EDA, Preporcessing</i>
<i>Saghana.K.S</i>	<i>Member</i>	<i>Feature Engineering , Code integration,Documentation</i>
<i>Rakshi.D</i>	<i>Member</i>	<i>Model building, Evaluation ,Data Transformation</i>