🧠 Customer Lifetime Value Prediction — Project Documentation

---

1. Abstract

Customer Lifetime Value (CLV) prediction is a critical metric for understanding long-term customer profitability. This project applies data-driven approaches and machine learning models to predict the CLV of customers based on behavioral and transactional data.
By leveraging clustering, feature engineering, and regression models, the project delivers actionable insights for marketing strategies, customer segmentation, and retention campaigns.

---

2. Introduction

Businesses increasingly rely on predictive analytics to identify their most valuable customers. CLV estimation enables organizations to prioritize customers who contribute the most to revenue over time.
This project uses a machine learning workflow to analyze customer purchase patterns and forecast their lifetime value, helping decision-makers allocate marketing resources effectively.

---

3. Problem Statement

The challenge is to build a model that accurately predicts the **Customer Lifetime Value (CLV)** using historical customer data.
The objective is to identify which customers are likely to bring the most value and what behavioral factors contribute to this value.

---

4. Objectives

- To perform data cleaning, preprocessing, and exploratory analysis.
- To segment customers using **RFM (Recency, Frequency, Monetary)** analysis.
- To identify behavioral clusters using **K-Means clustering**.
- To train predictive models (e.g., **XGBoost**, **Linear Regression**) for CLV estimation.
- To visualize results and interpret key drivers of CLV.

---

5. Dataset Overview

- **Source:** Company's historical customer database (CSV format).
- **Attributes:**
    - CustomerID – Unique customer identifier

- Recency – Days since the last purchase
- Frequency – Number of purchases
- Monetary – Total spending by customer
- Tenure, Region, Gender, Age, etc.

The dataset is stored in the dataset/ folder.

---

6. Methodology

6.1 Data Preprocessing

- Handled missing values and duplicates.
- Standardized column names and encoded categorical variables.
- Scaled numerical features using **MinMaxScaler** for better model performance.

6.2 Exploratory Data Analysis (EDA)

- Analyzed spending patterns and frequency distributions.
- Visualized customer segmentation using **Seaborn** and **Matplotlib**.
- Identified outliers and behavioral clusters through boxplots and histograms.

6.3 Feature Engineering

- Derived RFM scores:
  - **Recency:** Days since last purchase.
  - **Frequency:** Total number of purchases.
  - **Monetary:** Average purchase value.
- Combined RFM scores to categorize customers into groups (e.g., Loyal, At Risk).

6.4 Modeling

- Applied **K-Means clustering** to group similar customers.
- Built regression models:
  - **Linear Regression**
  - **Random Forest Regressor**
  - **XGBoost Regressor**
- Compared model performances using:
  - RMSE (Root Mean Square Error)
  - MAE (Mean Absolute Error)
  - R^2 Score

6.5 Model Evaluation

- **XGBoost** achieved the best predictive accuracy with the lowest RMSE.
- Feature importance revealed that **Monetary** and **Frequency** are the strongest predictors of CLV.

---

7. Results and Discussion

- Identified top 20% of customers contributing ~80% of total revenue (Pareto principle).
- Discovered customer clusters based on purchasing patterns:
    - **Cluster 1:** Loyal high-value customers
    - **Cluster 2:** Potential churners
    - **Cluster 3:** New low-frequency buyers
- The prediction model allows marketing teams to personalize offers and improve retention.

---

8. Visualizations

All visuals are stored in the visuals/ folder and include:

- RFM distribution charts
- Cluster visualization (K-Means)
- Feature importance graphs
- Actual vs. Predicted CLV plots

---

9. Conclusion

The model effectively predicts customer lifetime value and segments customers based on behavior. The insights derived from the analysis can help businesses:

- Optimize marketing campaigns.
- Improve customer retention strategies.
- Focus on high-value segments for profitability.

---

10. Future Work

- Integrate **real-time CLV prediction** using streaming data.
- Enhance model performance with deep learning (e.g., LSTM networks).
- Automate pipeline deployment using Flask or FastAPI.

---

12. Author

Sanket Kumar

📧 sanketsv11@gmail.com

💼 Data Science | Machine Learning | AI Enthusiast
