Project Report

Customer Lifetime Value (LTV) Prediction

Objective

The goal of this project is to predict the Customer Lifetime Value (LTV) using historical transaction data. By estimating the future value of customers, businesses can make informed decisions for targeted marketing, customer retention strategies, and revenue forecasting.

Tools and Technologies Used

- **Programming Language**: Python
- **Libraries**: pandas, seaborn, matplotlib, scikit-learn, XGBoost
- **Environment**: Jupyter Notebook / Local Python setup
- **File Formats**: CSV (for data), PKL (for model)

Dataset Description

Source File:

- `customer_segmentation.csv`

Key Columns:

- `CustomerID`: Unique identifier for each customer
- `InvoiceDate`: Date and time of the transaction
- `Quantity`: Number of items purchased
- `UnitPrice`: Price per unit

Data Preprocessing and Feature Engineering

1. **Data Cleaning**:
    a. Removed rows with missing values
    b. Ensured all `Quantity` and `UnitPrice` values were positive
2. **Feature Engineering**:
    a. **Recency**: Days since last purchase (relative to the dataset's latest date)
    b. **Frequency**: Number of unique purchase events per customer
    c. **AOV (Average Order Value)**: Total Spend / Number of Orders
    d. These features were combined into a new dataframe (`customer_df`)

Model Building

Model Used:

- `XGBRegressor` from the XGBoost library

Features:

- `Recency`
- `Frequency`
- `AOV`

Target:

- `LTV`: Customer Lifetime Value calculated as total revenue over the observed period

Training and Validation:

- Data was split into training and testing sets (80/20)
- The model was trained on the training set and evaluated on the test set using:
    - **Mean Absolute Error (MAE)**
    - **Root Mean Squared Error (RMSE)**

Customer Segmentation

- Customers were segmented based on their predicted LTV into four categories:
  - `Very High`
  - `High`
  - `Medium`
  - `Low`

This segmentation allows the business to prioritize resources and tailor marketing campaigns.

Visualizations and Insights

Data Exploration:

- `.describe()` and `.info()` provided a statistical and structural overview of the dataset
- `.value_counts()` helped analyze categorical distributions

Visual Plots:

- **Histograms**: Showed the distribution of Recency, Frequency, and AOV
- **Boxplots**: Compared predicted LTV across customer segments
- **Scatterplots**: Revealed positive correlations between Frequency/AOV and LTV
- **Heatmap**: Displayed feature correlations using a color-coded matrix

Exported Files

File Name	Description
`final_ltv_predictions.csv`	Full customer data with features, predicted LTV, and segment

`predicted_ltv.csv`	Simplified version with only CustomerID and predicted LTV
`ltv_model_xgboost.pkl`	Trained XGBoost model for reuse on new data

Summary

- A robust model was built to predict LTV based on Recency, Frequency, and AOV.
- Visualization and segmentation provide actionable insights for business teams.
- The trained model and data files can now be used in production or shared with stakeholders.

Optional Enhancements #if used for real time data

- Implement automated pipelines for weekly updates
- Integrate clustering for deeper segmentation
- Deploy the model with a frontend for real-time LTV scoring