Project Plan: Customer Lifetime Value (CLV) Analysis and Prediction

Goal:

Analyze customer behavior and predict Customer Lifetime Value (CLV) by leveraging simulated data based on real-world trends, enhanced by data scraped from Trustpilot and Amazon.

1. Data Collection

1.1 Scraping Data from Amazon:

Categories:

- grocery(chocolate candy),

- health (multivitamins)

- electronics(windows laptops)

- home(air conditioners)

Data to collect:

- Product title

- Product price

- Rating (1-5 stars)

- Amount of reviews

- Business name

- Category

- Subcategory

1.2 Scraping Data from Trustpilot:

For each category 15 samples:

- 5 high-rated businesses (4.5+ stars) with large review counts (>3,000 reviews)

- 5 mid-rated businesses (3.5–4.5 stars) with moderate review counts (1000-3000 reviews)

- 5 lower-rated businesses (<3.5 stars) with smaller review counts (<1000 reviews)

200–300 reviews per business

Data to collect:

- user\_name

- rating

- date

- location

- review\_text

1.3 Tools:

Scraping tools:

- BeautifulSoup.

- Selenium.

- Parsehub.

- Apify

Data storage:

- PostgreSQL (pgAdmin4).

1.4 Simulating Transactional Data:

- Generating transactional data for customers:

- Assigning customer IDs.

- Associating purchases with businesses and categories based on popularity (ratings and reviews).

- Generating transaction dates using uniform distribution over a set time period.

- Calculating Recency, Frequency, and Monetary Value (RFM) for each customer.

2. Data Cleaning and Preparation

2.1 Cleaning the Data:

- Handling missing values in reviews and ratings.

- Removing duplicates and resolving inconsistencies in business names and categories.

2.2 Data Transformation:

- Converting review dates into usable temporal features.

- Encoding categorical variables using one-hot or label encoding.

2.3 Feature Engineering:

- Calculating RFM metrics for each customer:

- Recency: Days since the last purchase.

- Frequency: Number of transactions.

- Monetary Value: Total revenue contributed by the customer.

- Additional features:

- Average transaction value.

- Engagement score based on review counts and ratings.

3. Data Analysis

3.1 Exploratory Data Analysis (EDA):

- Visualizing customer purchase behavior (purchase frequency, average spend).

- Analyzing correlations between RFM metrics and CLV.

- Analyzing high-value products and their contribution to revenue.

3.2 Customer Segmentation:

- Using K-Means Clustering on RFM metrics to segment customers:

- Loyal High-value customers, Loyal customers, at-risk customers (low frequency), New customers.

- Visualizing customer clusters.

4. Model Development

4.1 Target Variable:

- Predicting total revenue a customer will generate over 6 month.

4.2 Model Selection:

- Training regression models to predict CLV:

- Linear Regression.

- Random Forest Regressor.

- Evaluating models using RMSE, MAE, and R².

4.3 Feature Importance Analysis:

- Identifying key drivers of CLV using:

- Feature importance from tree-based model.

- SHAP values.

5. Dashboard Development

5.1 Integrating Data in PowerBI:

- Importing simulated transactional data, rfm metrics and model predictions into PowerBI.

5.2 Dashboard Visualizations:

- Customer Segments: Clusters of customers based on RFM metrics.

- CLV Predictions: Predicted CLV values for each customer using bar charts.

- Feature Importance: Highlighting key factors driving CLV predictions.

- Interactive Filters: Enabling filtering by region, category and customer segment.

6. Documentation and Insights

6.1 Final Report:

- Summarizing data collection and simulation methods.

- Describing modeling approaches and key results.

- Providing actionable recommendations to enhance customer retention and revenue.

6.2 Insights:

- Identifying high-value customers and revenue-driving products.

- Suggesting targeted marketing and promotions based on customer segments.

Deliverables:

1. Cleaned and structured datasets and simulated transactional data stored in PostgreSQL.

2. Machine learning models to predict CLV.

3. Interactive PowerBI dashboard with customer segmentation, CLV predictions, and insights.

4. Comprehensive report with methodology, analysis, and business recommendations.