Customer LifeTime Value(CLV) Prediction

Final Presentation

CRN 21323

Introduction & Problem Statement

What is CLV - Customer Lifetime Value: total expected revenue from a customer over their lifetime.

Problem Statement:

- Traditional CLV models are often too simplistic.

- Need for flexible models that capture nonlinear, dynamic customer behavior.

Project Goal - Build a multimodal machine learning pipeline (Regression + Classification + Clustering) to predict and segment CLV.

Code Architecture & Preprocessing Pipeline

Data Cleaning - Removed nulls, duplicates, handled encodings

Feature Engineering - Computed Recency, Frequency, Monetary value, Derived Customer Age, Country, and Time features

Preprocessing - Label encoding for categorical variables, MinMax Scaling for model input

Tools Used - pandas, sklearn, matplotlib, seaborn

CustomerID	Recency	Frequency	MonetaryValue	TotalQuantity	AOV	Tenure	PurchaseFrequency	CLV
-1.0	1	13	5126.76	387562	1022.596475	372	28.615385	4.945277E+06
12346.0	326	1	104.00	100	104.000000	0	0	0.000000E+00
12747.0	2	11	4196.01	1275	381.455455	366	33.272727	1.535740E+06
12748.0	1	13	5126.76	23565	149.931388	372	28.615385	7.250682E+05
12749.0	4	5	4057.86	1471	811.572000	209	41.800000	8.480927E+05

Baseline Model - Linear Regression

Purpose: Establish a simple benchmark for CLV prediction

Model Details:

- Input Features: Recency, Frequency, Monetary, Age
- Target: Total monetary value spent per customer

Observations: Fast & interpretable, Poor performance on high-variance customers

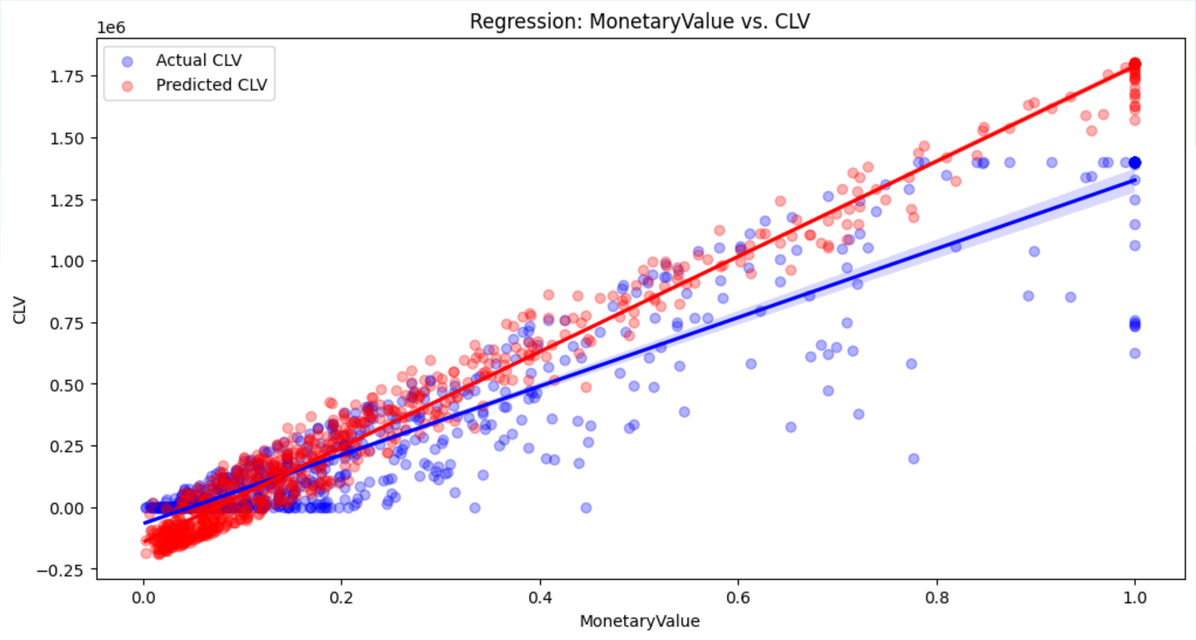
Advanced Regression - XGBoost

Why XGBoost - Captures non-linear relationships, Robust to outliers and skewed features

Techniques Used - Grid Search for hyperparameter tuning, Cross-validation for generalization

Performance - RMSE ↓, R^2 Score ↑

Best model for regression



CLV Classification - High vs Low

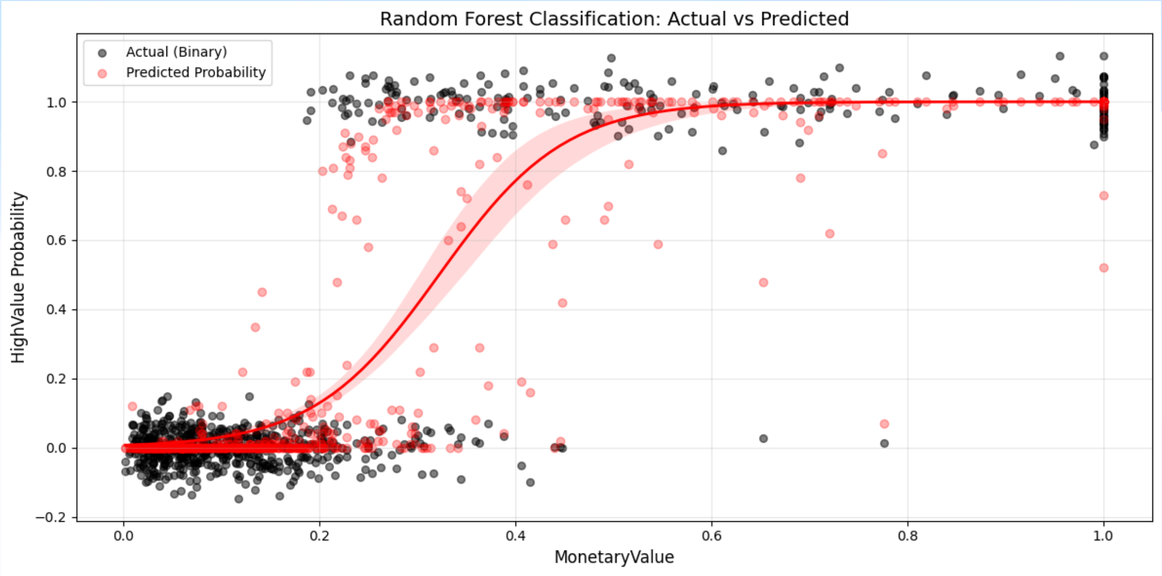
Why Classification - Businesses often need a binary decision: High-value or Low-value customers

Model - Random Forest Classifier

• Features: Recency, Frequency, Monetary, Age, Country

• Labels: 1 = High CLV, 0 = Low CLV (based on median split)

Outcome - Effective threshold-based segmentation, High precision in identifying top customers



Unsupervised Clustering – KMeans Segmentation

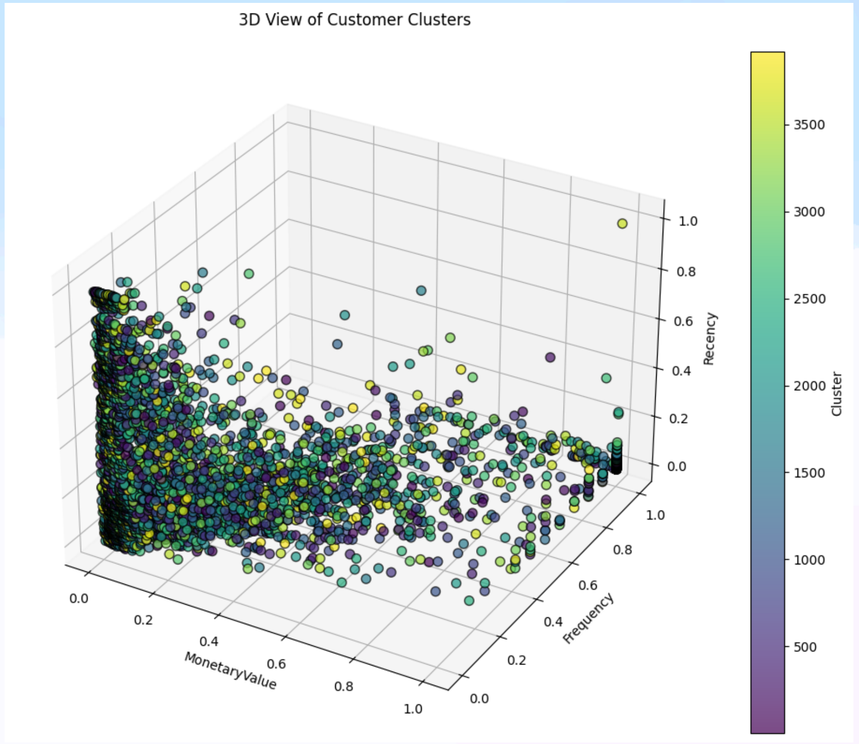
Why Clustering? - Identify natural groupings in customer behavior, Enable personalized marketing strategies

Model - Means

- Input: Scaled RFM + Customer Age features
- Clusters: 4 (based on Elbow Method)

Key Segments:

- Cluster 0: Loyal High-Spenders
- Cluster 1: Infrequent, Low-Spenders
- Cluster 2: New or Seasonal Customers



Conclusion & Future Work

Summary- Built full-stack CLV pipeline using Regression, Classification, and Clustering, Achieved strong accuracy and segment insights.

Business Value - Enhanced targeting, improved ROI, better customer retention

Future Scope:

- Deploy as API for real-time predictions
- Add NLP features from reviews
- Experiment with Deep Learning (e.g., LSTMs)

Harshith Reddy Gundra. 700780724

Sainath Konda. 700757121

Nikhila Potla. 700754837

Sai Bhavani Prasad. 700754838