

Customer Lifetime Value Analysis

Pamela Gatica

Marketing Analytics
November 2025

What is Customer Lifetime Value?



Customer Lifetime Value (CLV) is the total revenue a business can expect from a customer over their entire relationship with the company.

It helps companies:

- ✓ Identify their most valuable customers.
- ✓ Improve marketing and retention strategies.
- ✓ Allocate resources more effectively.
- ✓ Predict future revenue.

What this project covers:

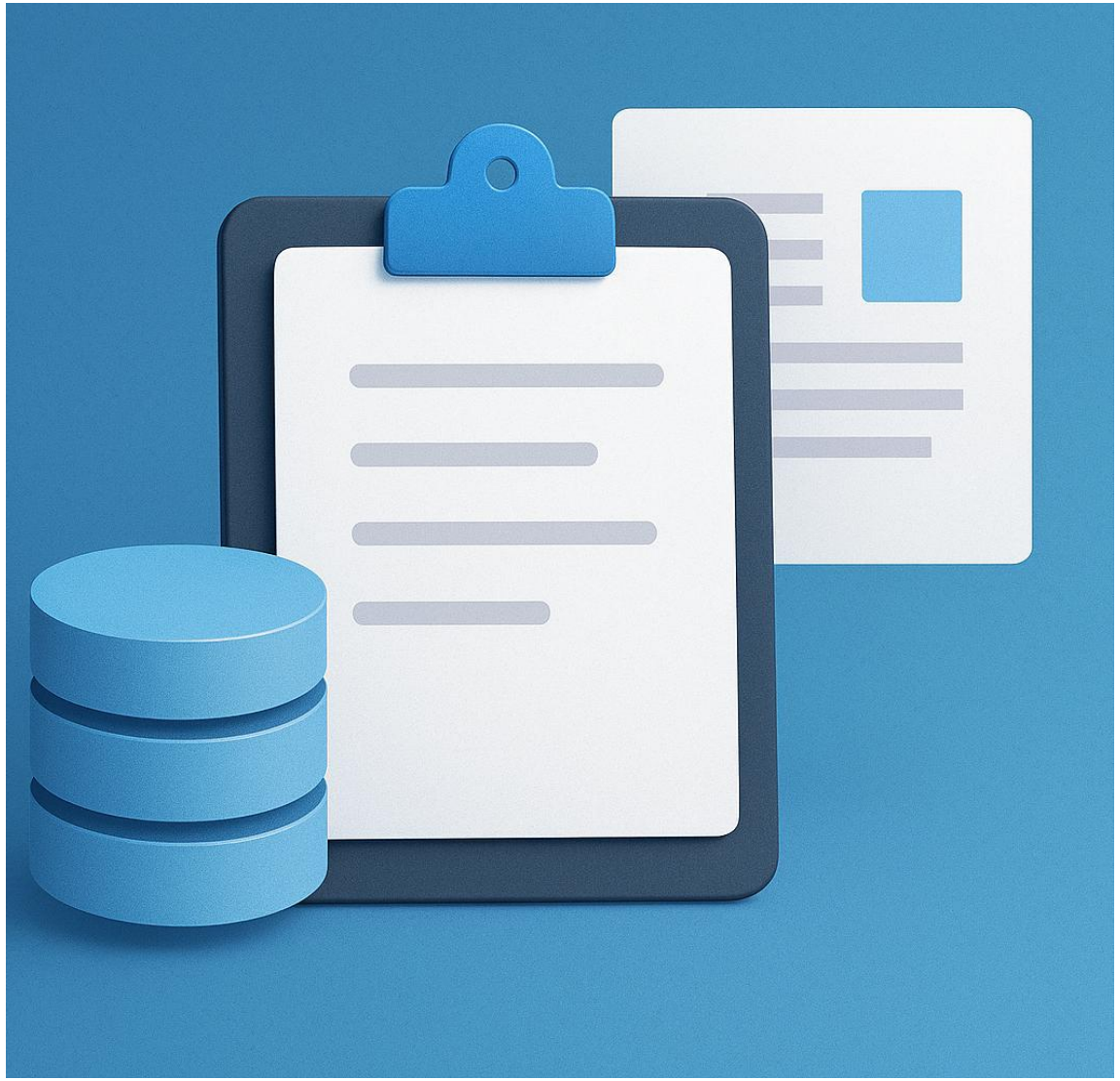
- ✓ Understanding and preparing the dataset / Exploratory Data Analysis (EDA).
- ✓ RFM (Recency, Frequency, Monetary) segmentation to find top customers.
- ✓ Customer behavior patterns (churn, loyalty, spending).
- ✓ Predicting CLV using machine learning models.
- ✓ Comparing model performance.

Dataset Description

- **Dataset Size:** 10,000 customers
  Columns: 19 fields (numeric, categorical, dates)

- **Numerical columns (10):** ['customerid', 'totalpurchases', 'totalrevenue', 'discountsused', 'feedbackscore', 'frequency', 'avgpurchasevalue', 'recency', 'tenure', 'avgtimebetweenpurchases']
- **Categorical columns (7):** ['referralsource', 'churnindicator', 'productcategory', 'responsetolastcampaign', 'preferredpaymentmethod', 'supportticketsraised', 'hasloyaltycard']
- **Date columns (2):** ['firstpurchasedate', 'lastpurchasedate']

* No missing values in the dataset.

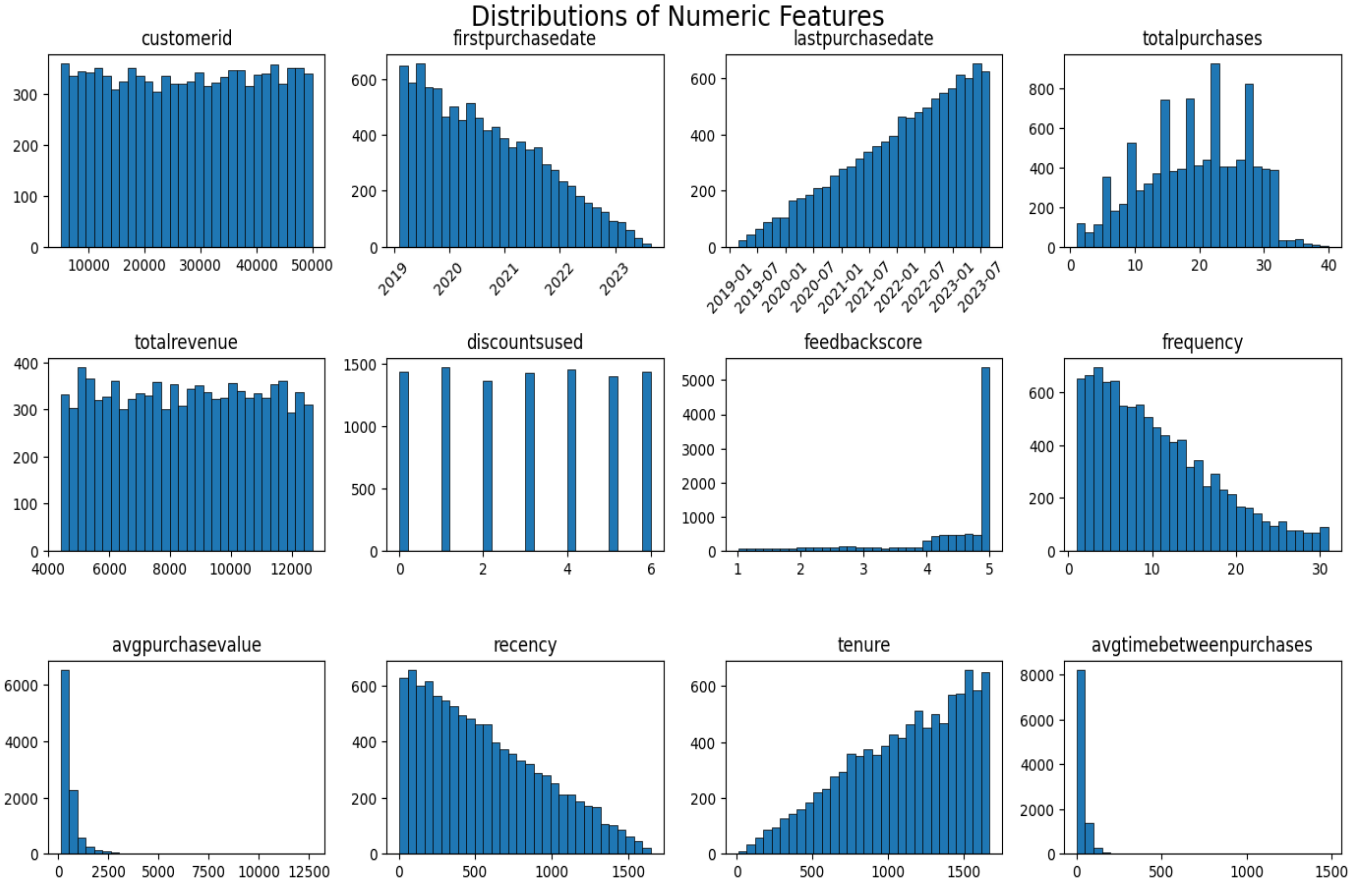


Numerical Distributions

•Most purchase behaviors (*totalpurchases, frequency, avgpurchasevalue*) are **right-skewed**, meaning only a small group of customers shows very high activity or spending.

•*Recency* is **generally low**, while *tenure* is **high**, indicating that many customers purchased recently and have been with the company for a long time.

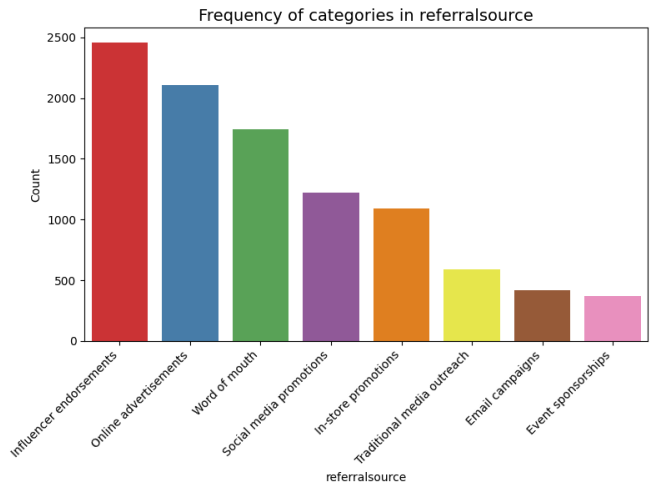
•*Feedback scores* are concentrated at the **higher end**, suggesting overall positive customer satisfaction.

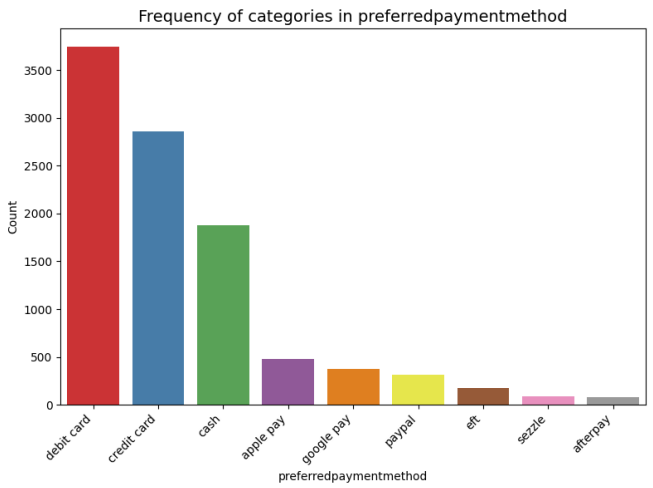


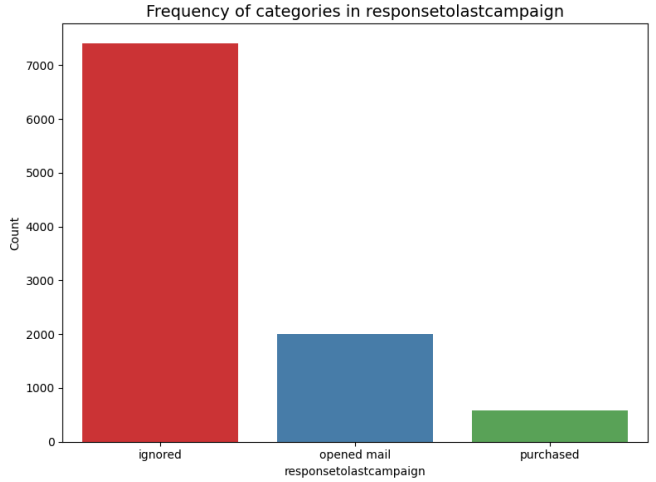
Overall, the data reveals a **small segment of high-value customers** and a **large base of lower-activity customers**.

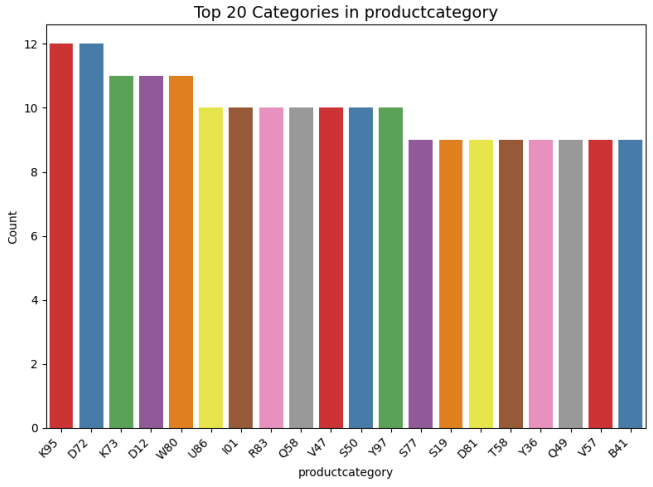
Categorical Distribution

Marketing & Customer Engagement





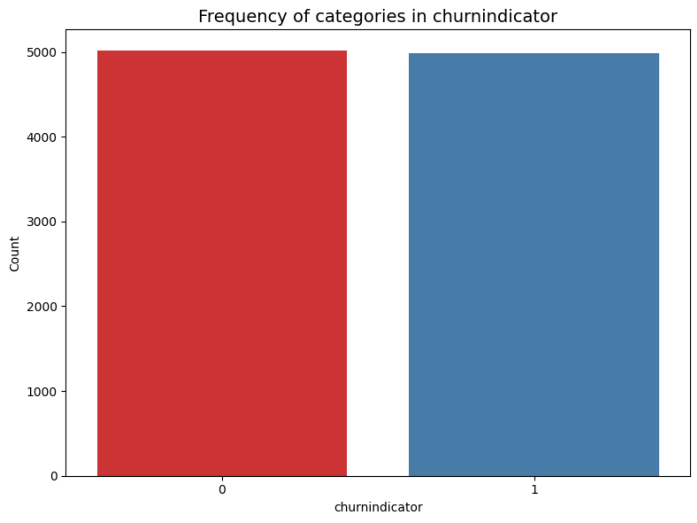


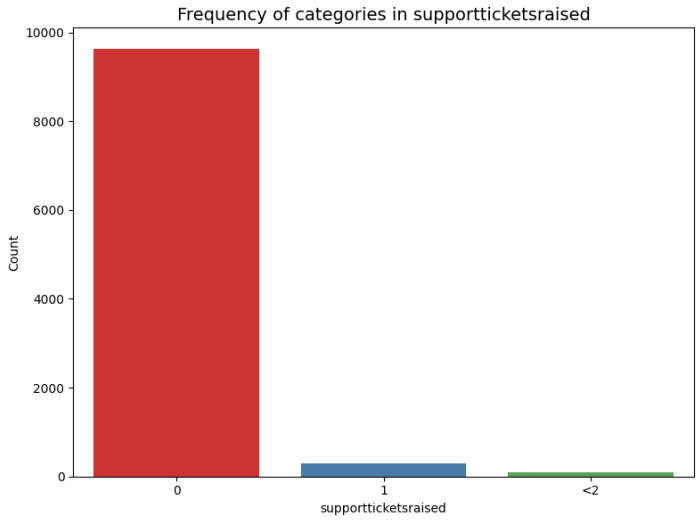


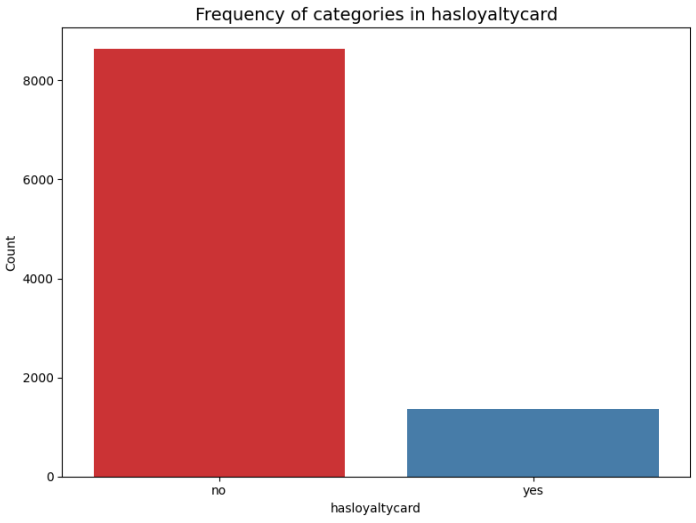
- Most customers come from influencers and online ads.
- Campaign engagement is very low; most messages are ignored.
- Debit card and credit card are the dominant payment methods.
- Product categories are evenly distributed, with no single category dominating.

Categorical Distribution

Customer Support & Loyalty



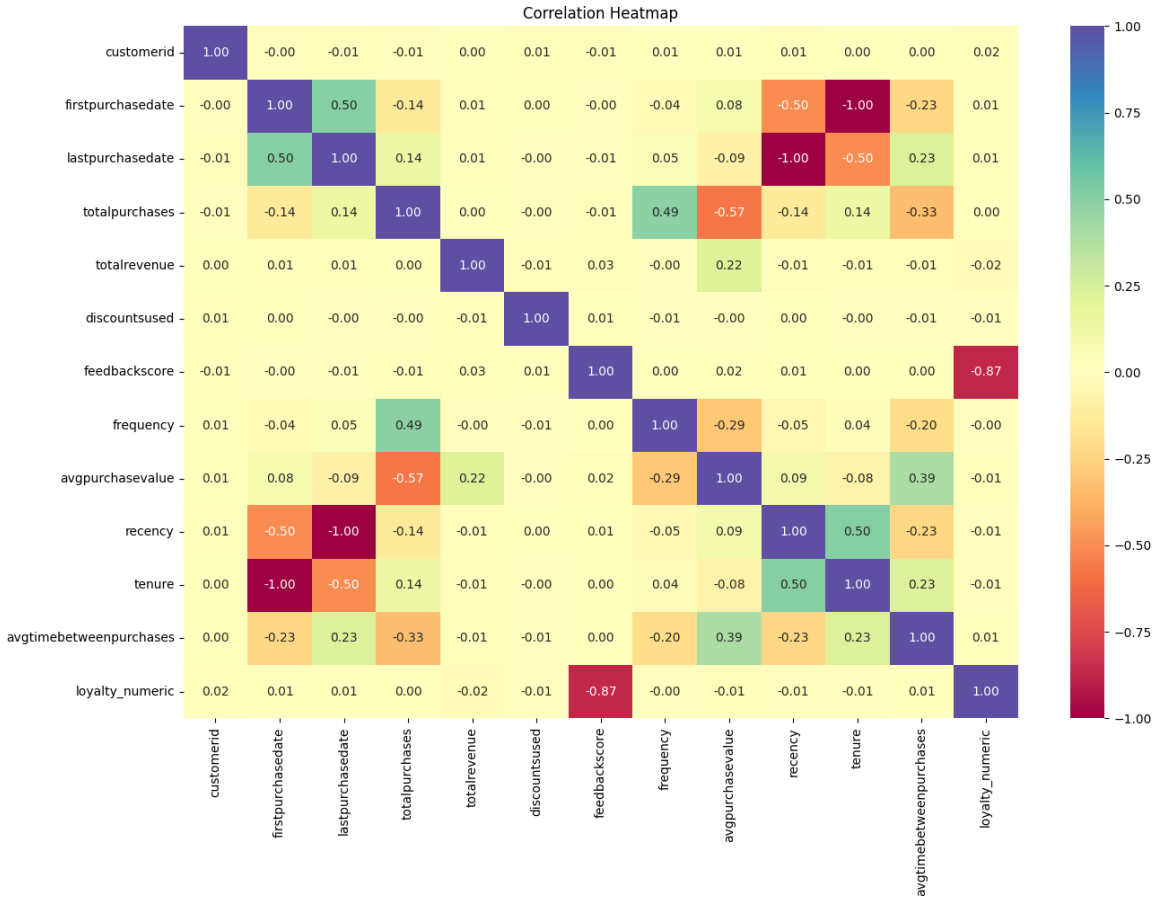


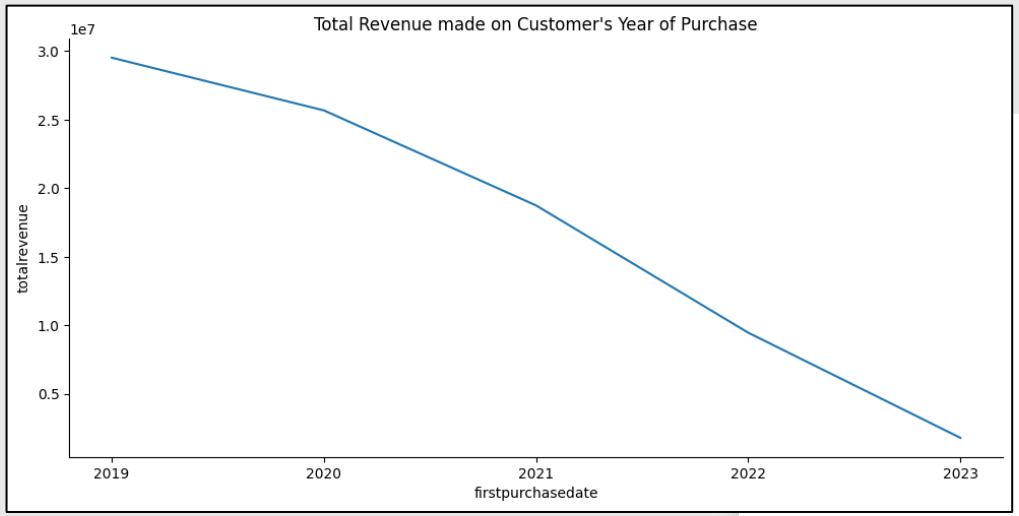


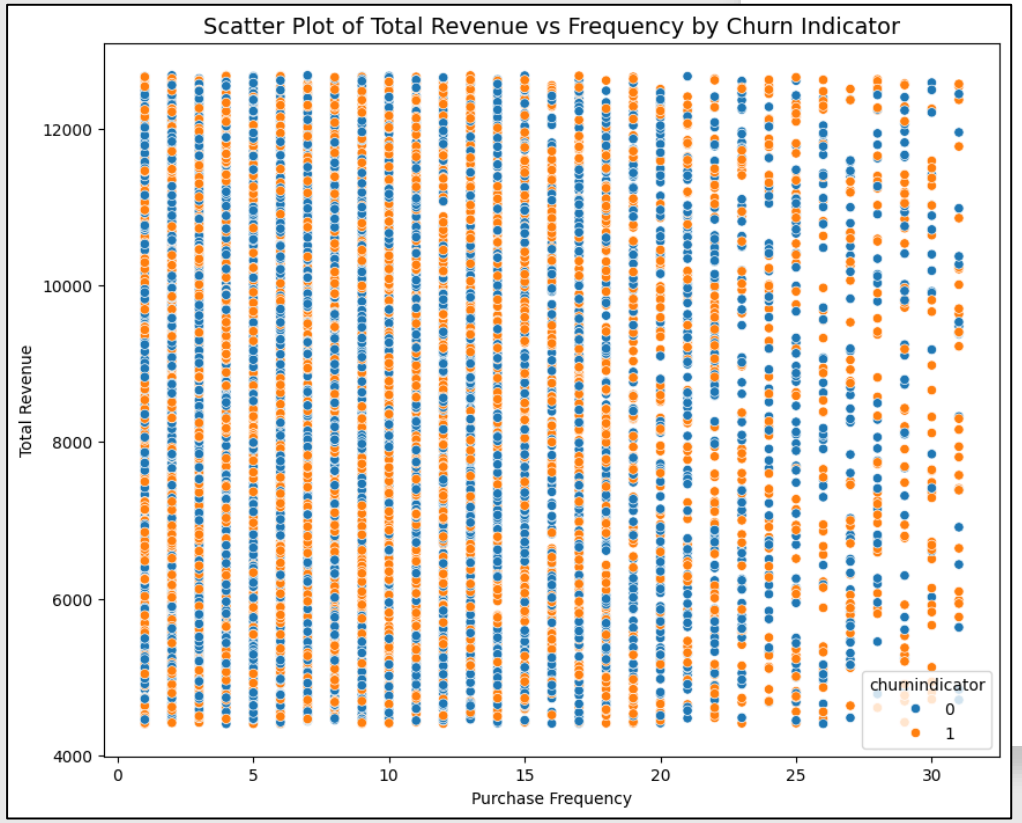
- Churn is **evenly balanced** between customers who stay and those who leave.
- Most users have **zero support tickets**, showing very low support interaction.
- Most customers **do not** have a loyalty card, indicating **low program adoption**.

Correlation Heatmap

- **Most features have very weak correlations**, meaning the variables are relatively independent from each other.
- **Time-based features** (recency, tenure, avg time between purchases) show the **only notable relationships**, which is expected because they measure customer lifecycle stages.
- **No feature shows a strong correlation with total revenue**, suggesting that revenue is influenced by multiple factors rather than a single dominant variable.







Customer Revenue Patterns and Churn Behavior

- Revenue declines significantly for newer customer groups, especially after 2021, suggesting that **recent customers either purchase less or churn earlier**.
- Churned customers appear across all frequency levels, indicating that **churn is not limited to low-frequency buyers**.
- Using the RFM model, the top 5% of customers (500 users) are identified as those with the higher RFM scores → buy more frequently, spend more and have purchased more recently.

Baseline CLV Models

Ridge, Random Forest and XGBoost

Ridge Regression	Shows high prediction error (MSE ≈ 5.5M), indicating limited ability to model CLV.
Random Forest	Performs significantly better, reducing error to ≈ 3.5M.
XGBoost	Achieves similar performance (≈ 3.56M) and trains much faster.

Overall, tree-based non-linear models fit the CLV patterns more effectively than linear approaches.

Advanced CLV Models



Decision Tree

- RMSE ≈ 99, R^2 ≈ 0.998
- Delivers highly accurate predictions, capturing non-linear customer behavior.



Random Forest

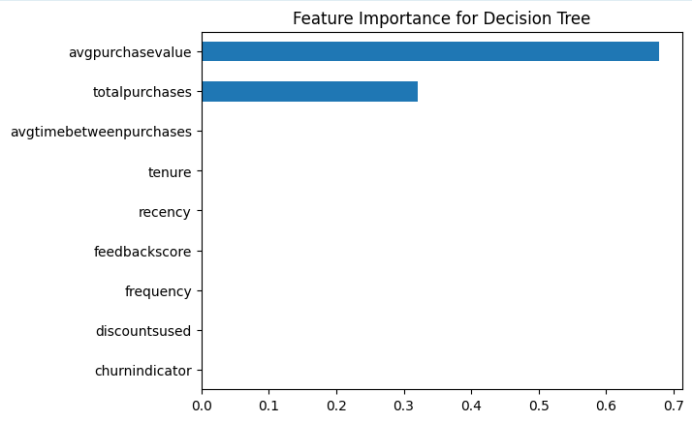
- RMSE ≈ 65, R^2 ≈ 0.999
- Provides the strongest overall performance and excellent generalization.

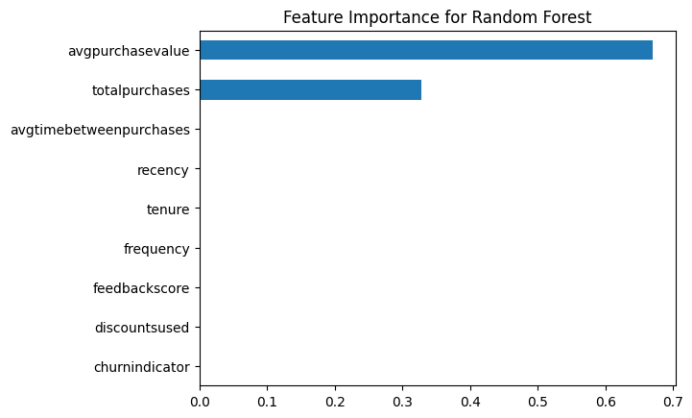


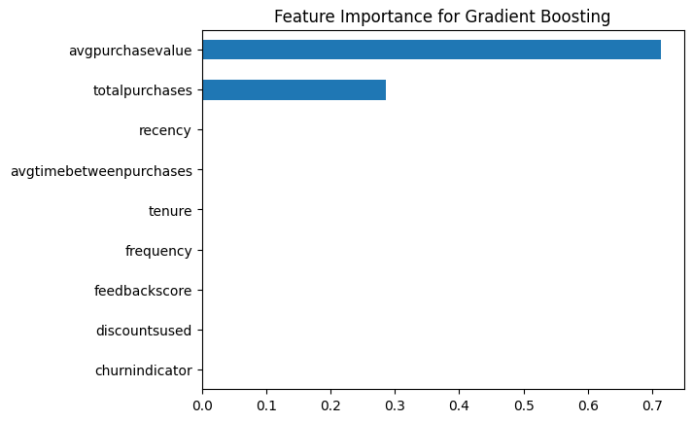
Gradient Boosting

- RMSE ≈ 333, R^2 ≈ 0.98
- Performs well but slightly less stable than Random Forest.

Overall, tree-based non-linear models consistently outperform linear approaches, both in accuracy and stability.







Feature Importance

- Across all tree-based models, **only two features show meaningful importance**:
  - ✓ *avgpurchasevalue* and *totalpurchases*, together explaining nearly all predictive power.

- The remaining features contribute close to zero, indicating that CLV in this dataset is driven almost entirely by how much customers spend per purchase and how many purchases they make.

Business Recommendations

Strengthen	Influencer and online ad investments, as these channels bring most new customers.
Improve	Campaign engagement through better targeting and more personalized messages.
Promote	The loyalty program, since adoption is low despite its positive impact on CLV.
Encourage	Repeat purchases with bundles, reminders, and targeted offers.
Increase	Visibility of customer support, ensuring users know how to access assistance.
Optimize	Payment options, highlighting the most preferred methods to reduce checkout friction.
Prioritize	High-value product categories, even though overall category distribution is balanced.


THANK YOU