Capstone 3 - Final Report

Final Project Report: Predicting Customer Lifetime Value (CLV) to Optimize Targeted Marketing Strategies in E-Commerce

Introduction

In the competitive e-commerce sector, accurately predicting Customer Lifetime Value (CLV) is crucial for identifying high-value customers and optimizing marketing strategies. CLV prediction helps businesses allocate resources efficiently, tailor loyalty programs, personalize marketing efforts, and maximize Return on Investment (ROI). This report discusses the importance of CLV in enhancing resource allocation, personalizing marketing, maximizing revenue, and facilitating strategic decision-making.

Problem Statement

In this capstone project, the goal was to improve targeted marketing strategies in a Brazilian e-commerce setting by accurately predicting Customer Lifetime Value (CLV) using a dataset of 100,000 anonymized orders from 2016 to 2018.

Dataset Description

The primary data, sourced from Kaggle, was composed of 5 datasets: https://www.kaggle.com/olistbr/brazilian-ecommerce. These were joined to create a single dataset that included order details, item-level data, product attributes, customer demographics, geographical information, and payment methods.

Data Wrangling

Data cleaning involved dropping missing values (less than 5%) and converting date features to datetime format. We then derived various features:

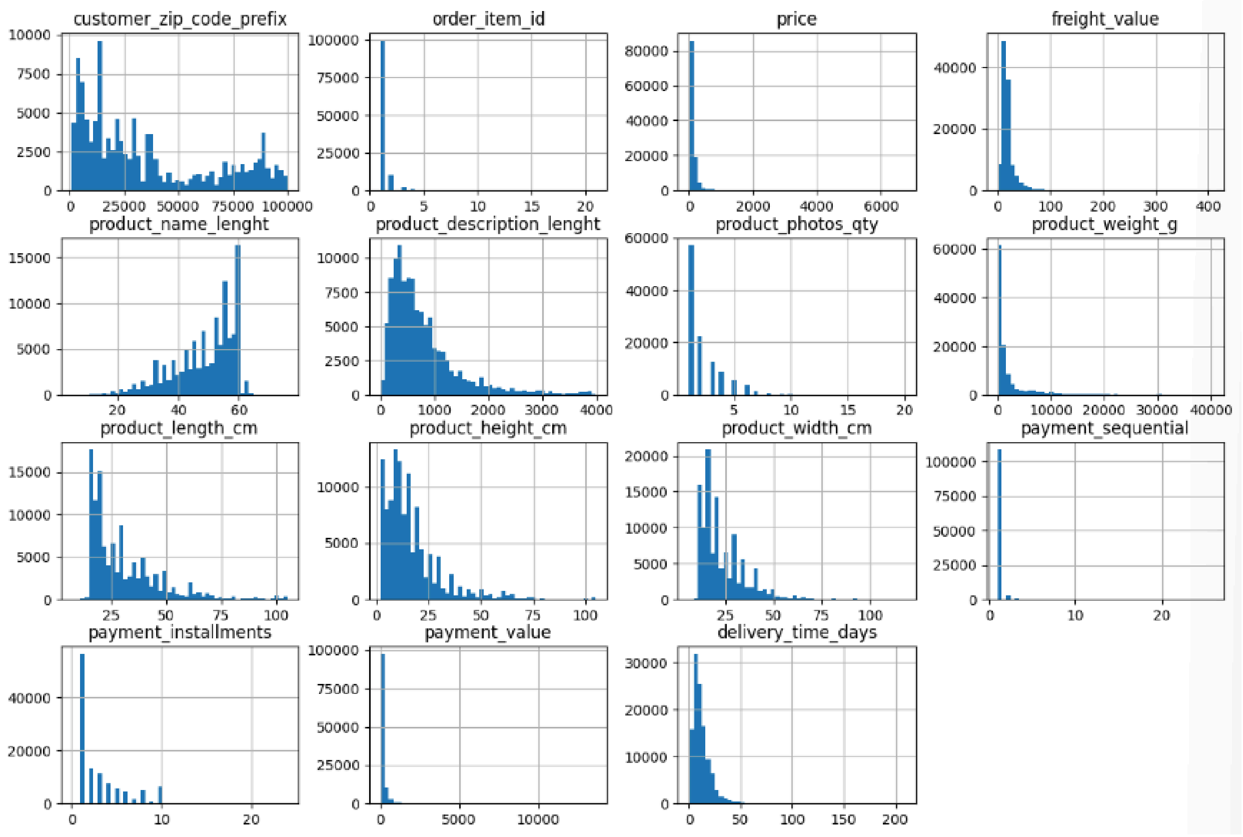
- **Customer-level aggregates** such as the total number of orders, total spend, average order value, order frequency, and customer tenure to summarize interactions with the e-commerce platform.
- **Recency, Frequency, Monetary (RFM)** to define the recency of the last purchase, the frequency of purchases made, and the monetary value of those purchases.
- **Product interaction** features including the number of unique products purchased, product category diversity, average product price, and total units purchased by each customer.

- **Shipping and delivery** were created such as average delivery time, average shipping cost, and on-time delivery rate, to gauge the impact of these factors on customer satisfaction and loyalty.
- **Payment behavior** features were crafted to identify the preferred payment method, the average number of installments chosen, and the total number of payment transactions made by each customer.

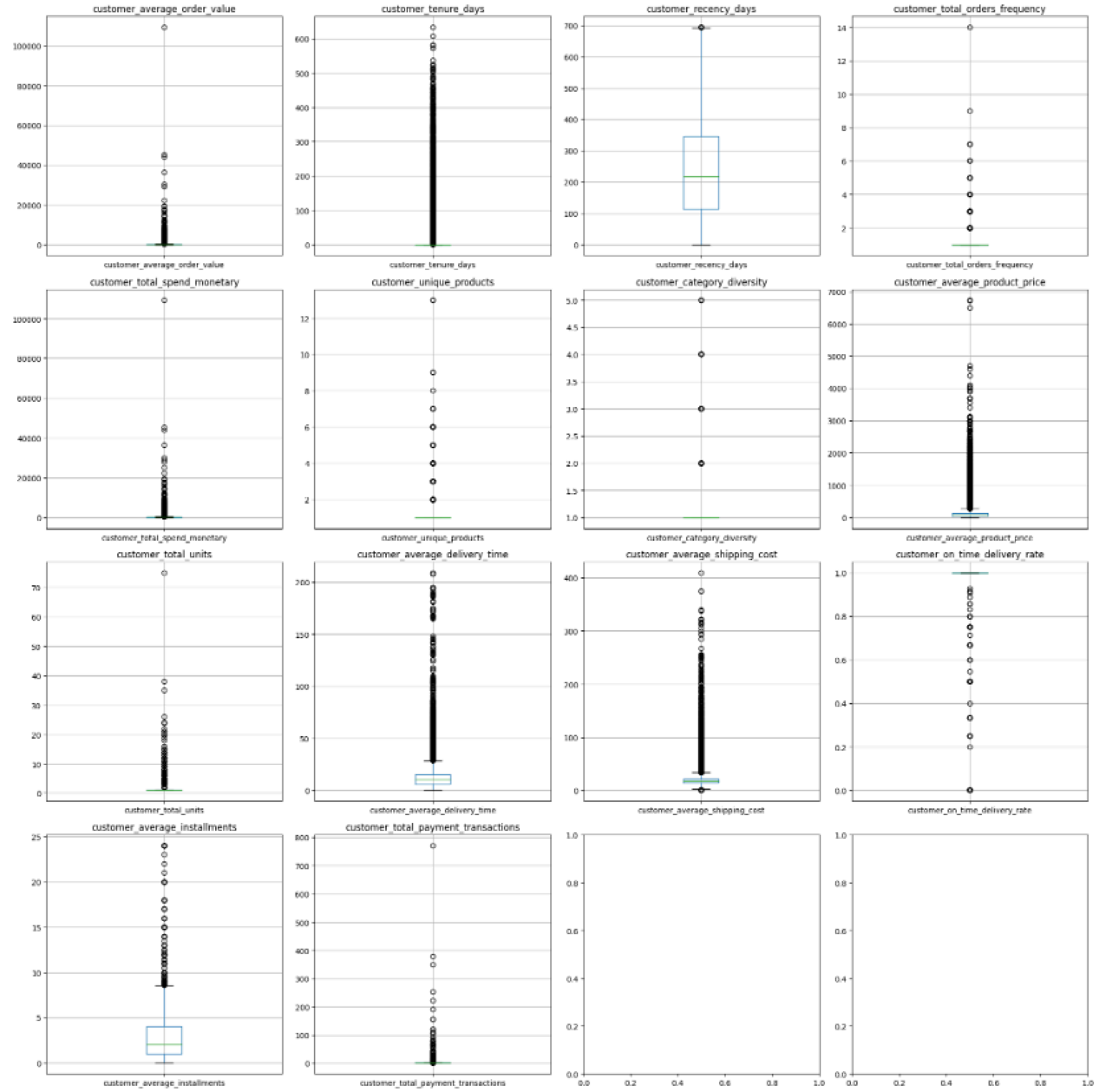
These derived features were then consolidated into an additional dataset to set the foundation for reliable CLV predictions.

Exploratory Data Analysis (EDA)

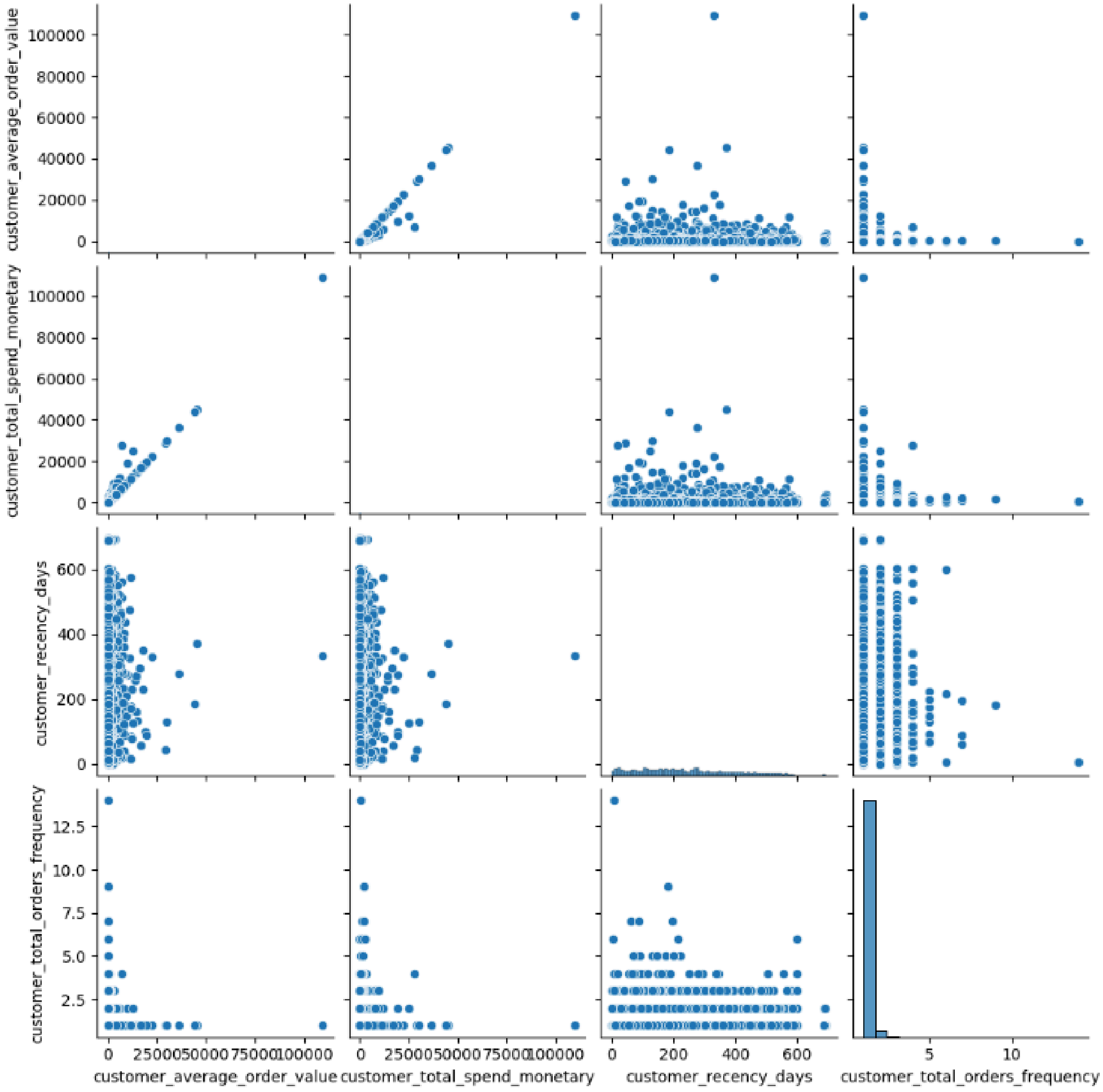
Next we used visualizations to begin exploring our data. Initial histogram analysis revealed that most customers are low-frequency buyers with limited product diversity and low average order values. The delivery times and shipping costs were generally favorable, with rare usage of financing options like installments.



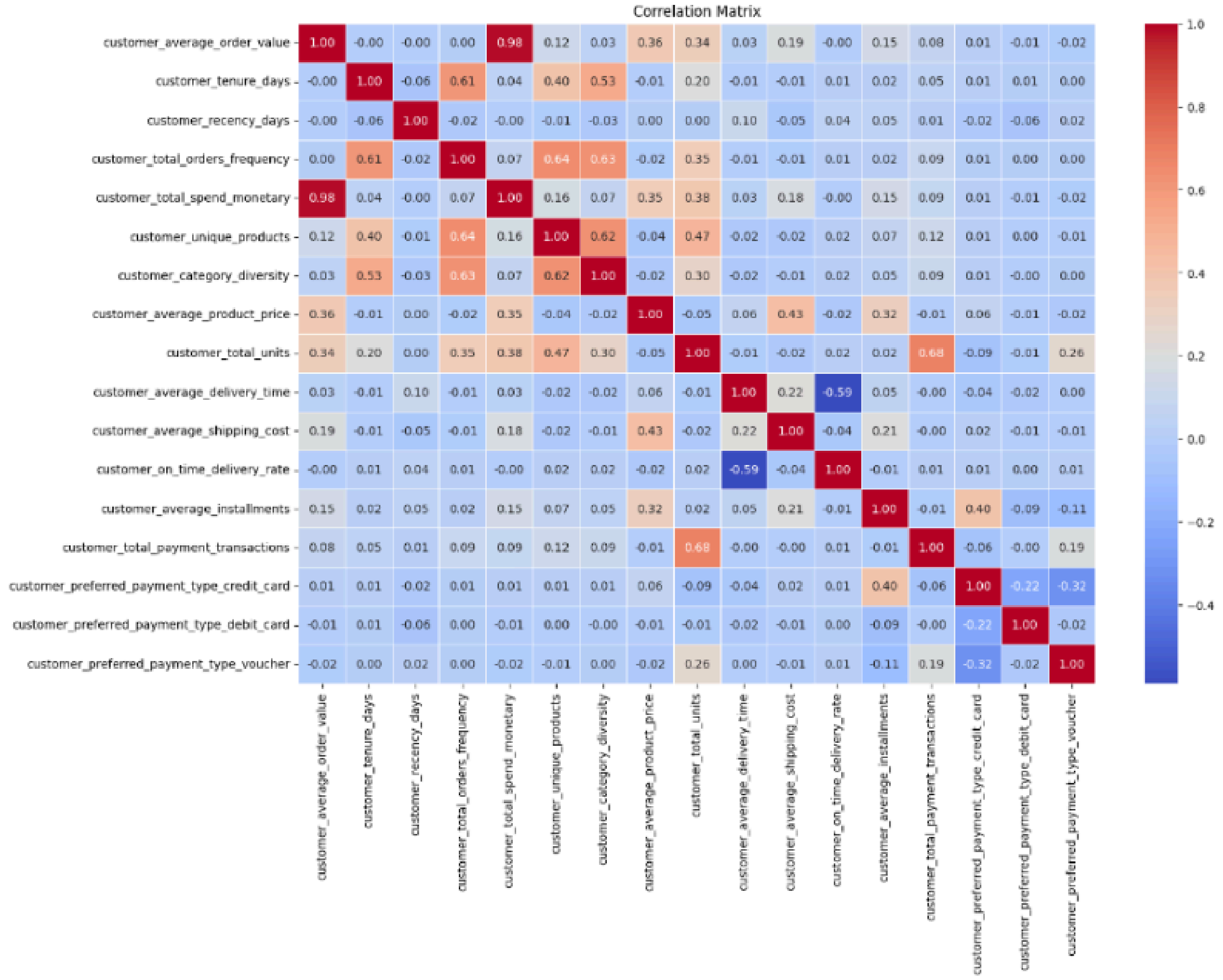
Boxplots showed significant outliers and skewness, indicating higher average order values and total spend among a select group of high-value customers.



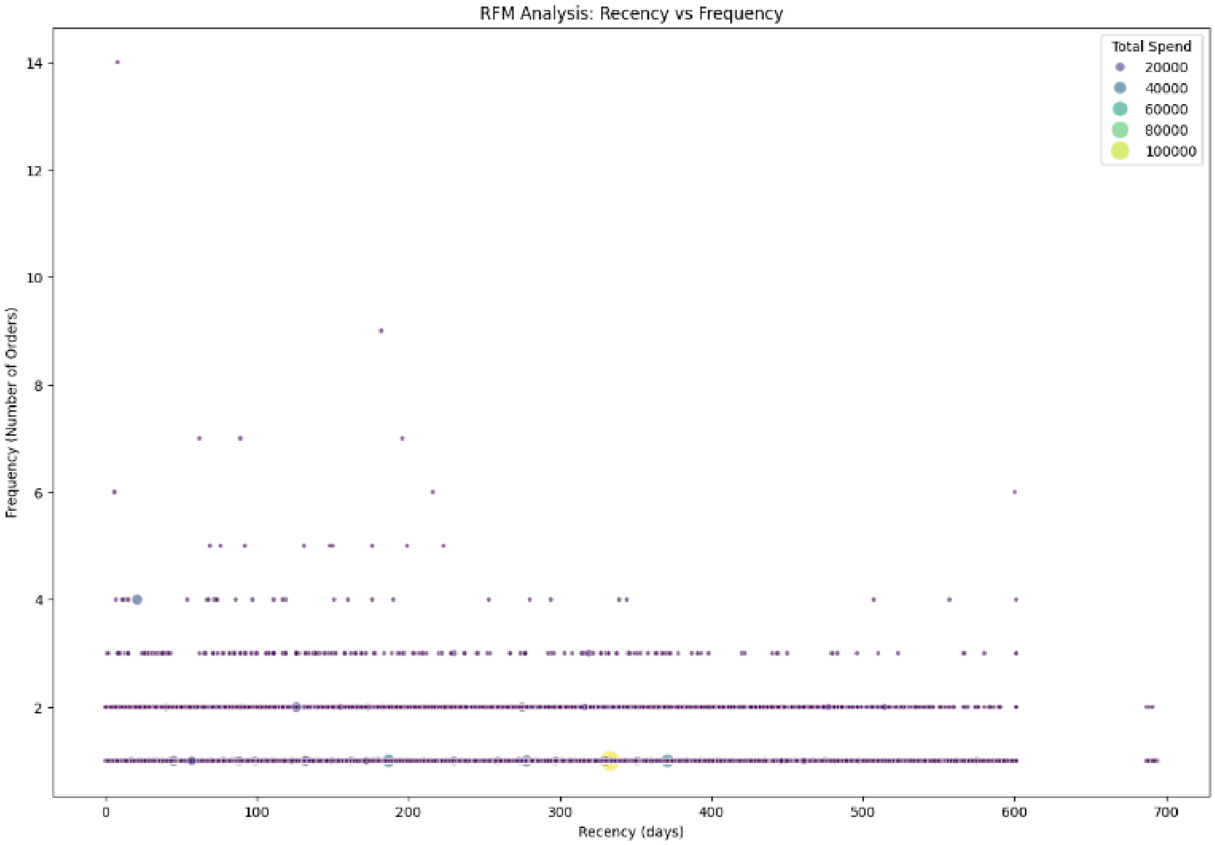
Pairplots highlighted a strong linear relationship between customer total spend and average order value.



With correlation analysis we observed that customers with longer tenures often display higher category diversity but tend to order less frequently (average time between orders is high). We also further confirmed high positive correlation of average order value and total units purchased with total spend ('customer_total_spend_monetary'), which will be our target feature. Because 'customer_average_order_value' and 'customer_total_units' are not highly correlated with each other, it seems we do not need to be concerned with multicollinearity. These seem to be providing valuable information toward predicting total spend!



To visualize RFM, we generated a scatter plot depicting recency (the number of days since the last purchase) vs frequency (the number of purchases made). Color intensity and bubble size are used to indicate total spend (monetary). We see that most customers have a relatively low frequency of purchases, with a majority making between 1 and 3 purchases. Customers with higher frequency (more than 3 purchases) are considerably fewer in number, though there are some notable outliers with up to 14 purchases.



Other analyses showed that the top 20% of customers account for about 80% of total revenue. Temporal spikes in spending and order volume were most likely linked to marketing campaigns and/or seasonal trends. Payment analysis indicated a strong preference for credit cards, while geographical analysis pointed to high customer concentrations in São Paulo and Rio de Janeiro.

Preprocessing, Modeling, and Evaluation

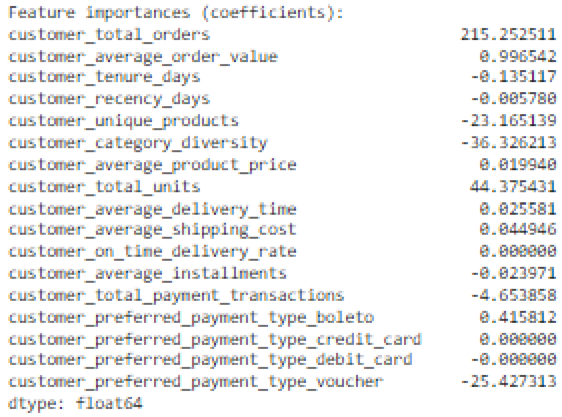
To ensure a realistic prediction scenario and avoid data leakage, the original data was split into training and test subsets and then the derived customer features were created again on each of these. We then applied one-hot encoding to categorical features on the training subset only. Finally we split each of our train and test subsets into X and y sets.

We wanted to evaluate whether there would be a difference in model results if our data was standardized so we standardized the X_train data and then transformed both the train and test data. We then used Linear Regression as our baseline model and evaluated standardized versus non-standardized. There was no difference between the two outcomes so we concluded that standardized data was not needed for further model evaluations.

We proceeded by evaluating several models, including Random Forest Regressor (a tree-based model using bagging), Gradient Boosting Regressor (another tree-based model using boosting), and Lasso Regression (a linear model with L1 regularization to manage any overlooked multicollinearity and perform feature selection). Ensemble and stacked modeling techniques were subsequently used to further investigate possibilities using aggregate predictions and incorporating a meta-model for combining base model outputs.

Among the models, Lasso Regression emerged as the most effective, presenting a mean squared error (MSE) of 2352.66 and an R² score of 0.9821. This model provided the best balance between predictive accuracy and preventing overfitting.

Key Findings



```
Feature importances (coefficients):
customer_total_orders                        215.252511
customer_average_order_value                   0.996542
customer_tenure_days                          -0.135117
customer_recency_days                         -0.005780
customer_unique_products                     -23.165139
customer_category_diversity                  -36.326213
customer_average_product_price                 0.019940
customer_total_units                          44.375431
customer_average_delivery_time                 0.025581
customer_average_shipping_cost                 0.044946
customer_on_time_delivery_rate                 0.000000
customer_average_installments                 -0.023971
customer_total_payment_transactions           -4.653858
customer_preferred_payment_type_boleto         0.415812
customer_preferred_payment_type_credit_card    0.000000
customer_preferred_payment_type_debit_card    -0.000000
customer_preferred_payment_type_voucher      -25.427313
dtype: float64
```

1. **Most Significant Features**: The Lasso Regression model identified critical features impacting CLV primarily as 'customer_total_orders', 'customer_total_units', 'customer_average_order_value', 'customer_tenure_days', and 'customer_category_diversity'. In particular, the quantity of past orders and total units purchased are strong predictors of future customer value.

2. **Customer Segmentation**: Most customers are low-frequency buyers with lower average order values and limited product diversity. However, a small segment of high-value customers (top 20%) accounts for about 80% of the revenue. These high-value customers display higher longevity and greater category diversity despite lower order frequencies.

3. **Payment Preferences**: There is a strong preference for using credit cards for transactions. Although other payment methods like 'boleto' (a type of bank slip that can be used by individuals

and businesses to pay for goods and services) and 'voucher' have some influence, they played a lesser role compared to credit card payments.

4. **Geographical Insights**: São Paulo and Rio de Janeiro are key geographical areas with high customer concentrations, making these regions strategic for targeted marketing efforts.

5. **Delivery and Experience**: Average delivery time and shipping cost positively influenced CLV, implying that efficient logistics and delivery services contribute to higher customer value.

Recommendations

1. Focused Marketing Campaigns:

- **High-Value Segment**: Develop tailored marketing strategies to retain and further engage the top 20% of customers, potentially through loyalty programs, exclusive offers, and personalized communications.
- **Low-Value Segment**: Implement targeted promotions to convert low-frequency buyers into regular customers. Consider offering incentives like discounts on bulk purchases or special promotions on diverse product categories.

2. Geographical Targeting: Allocate more resources for marketing in São Paulo and Rio de Janeiro, as these cities show the highest customer concentrations. Regional promotions, localized advertising, and partnerships with local influencers could enhance engagement.

3. Optimize Payment Options: While the majority of transactions are through credit cards, promoting alternative payment methods (like 'boleto' which shows a positive coefficient) can diversify the customer base. Ensure seamless and secure payment processes for better customer experience.

4. Enhance Customer Experience:

- Improve delivery times and reduce shipping costs where possible, as these factors positively affect CLV.
- Maintain high on-time delivery rates to foster customer satisfaction and loyalty.

5. Product Diversification: Encourage customers to explore a wider range of products. This can be achieved through personalized recommendations based on past purchases, cross-selling, and up-selling techniques.

6. Engagement and Retention Strategies:

- Focus on engaging customers with longer tenures by involving them in community-driven initiatives, feedback loops, and loyalty rewards.
- Reduce recency by re-engaging customers who haven't purchased in a while through reactivation campaigns, timely reminders, and attractive offers.

Ideas for Further Research:

1. **Behavioral Segmentation**: Explore clustering techniques to segment customers based on behavioral patterns beyond RFM (e.g., browsing habits, product preferences) to refine marketing tactics.

2. **Time Series Analysis**: Apply time series forecasting to predict seasonal trends and campaign impacts more precisely, enhancing strategic planning.

3. **A/B Testing for Campaigns**: Implement A/B testing to determine the effectiveness of different marketing strategies and promotions, allowing for data-driven decisions on optimizing marketing efforts.

4. **Customer Feedback Integration**: Incorporate qualitative data, such as customer reviews and feedback, to enhance the predictors used in the CLV model. Sentiment analysis and natural language processing techniques can provide insights into customer satisfaction and uncover potential areas for improvement.

Conclusion

This project underscores the critical role of predicting Customer Lifetime Value (CLV) in driving strategic decision-making within the e-commerce sector. A robust Lasso regression model achieved precise CLV predictions that offer actionable insights for targeted marketing, optimized product offerings, and improved customer experiences. These strategies can drive business growth and maximize ROI. Implementation of these recommendations and pursuing further research will enable even deeper customer insights and enhance long-term business strategies.