

# Segmenting Customers



# with Predictive CLV

Debbie Trinh

July 28, 2023

# Introduction

Companies spend a lot of money to acquire and retain customers. Data and customer lifetime value (CLV) can help to identify which customers are the most profitable to retain and upgrade.

This data-driven approach will aim to leverage the power of predictive CLV to identify and segment the highest valued customers. This analysis will also dive into notable trends of the behaviors and demographics of the top 10% of customers. Furthermore, segmenting on CLV will enable us to tailor marketing strategies / loyalty programs to different customer segments, improving targeting, personalization, and overall customer satisfaction and loyalty.



# Data Overview, Imputation, and Transformation

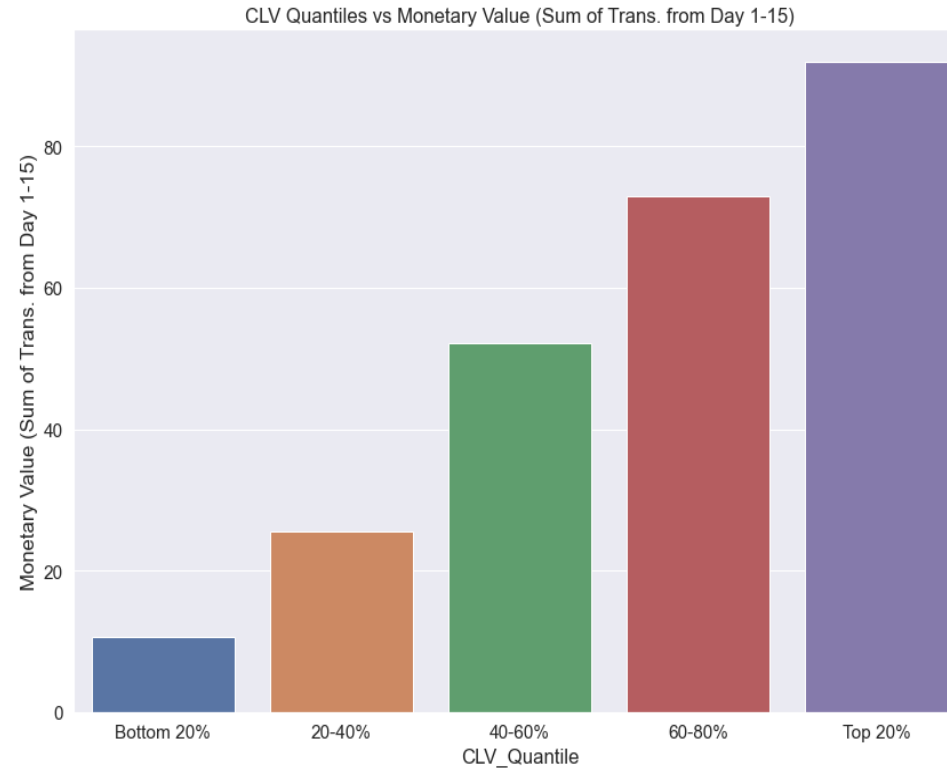
The Starbucks reward offers dataset from Kaggle simulates the Starbucks Reward Program, consisting of 76,277 marketing offers sent to 17,000 users over 30 days.

The dataset includes three tables: an event log tracking user actions such as receiving, viewing, completing offers, and transactions; customer profiles containing demographic information like age, gender, income, and membership year; and marketing offers sent through various channels.

To construct the Customer Lifetime Value (CLV) dataset, the original data was divided into two equal 15-day periods. The first period's demographic, behaviors, and spending data were used to predict spending in the latter period. CLV is calculated as the sum of transactions during the latter period and segmented into five quantiles. Quantiles are statistical measures that divide a dataset into equal segments, representing different customer groups based on value. The "CLV\_Quantile" column was added to analyze representative customer attributes and spending/engagement behaviors within each quantile. Null numerical variables were imputed with 0. Null numerical variables were imputed with 0. Null categorical variables were imputed under the label "Unknown", followed by one-hot encoding.

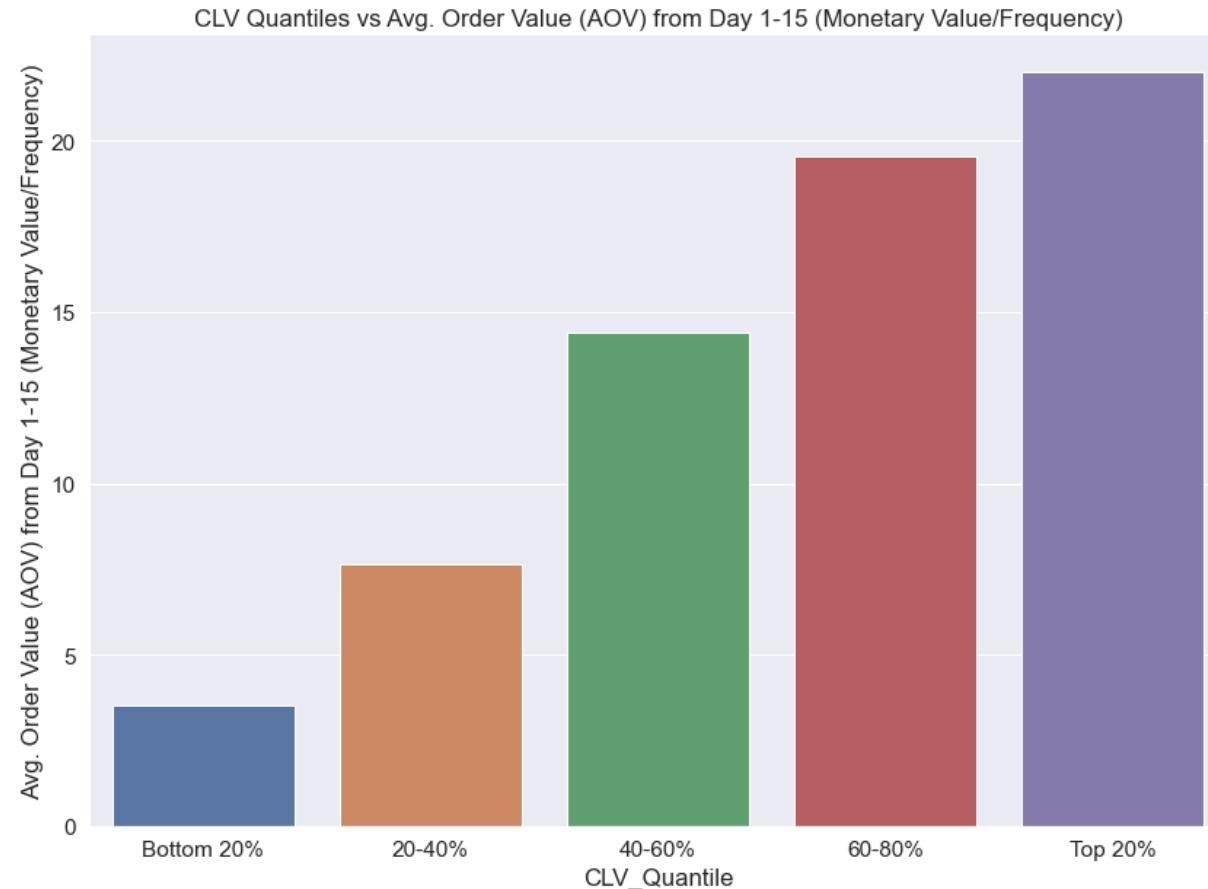
Features engineered for the dataset include Recency, Frequency, Monetary, Average Order Value (AOV), Response Rate, View Rate, and Completed Offers by Channel, all specific to Days 1-15. Customer attributes such as age, gender, household income, and membership year were also included. The target variable is defined as the sum of transactions for Days 16-30.

# Top 20% Most Valuable Customers spend \$6/day on average, equivalent to one specialty drink



Marketers use RFM (Recency, Frequency, and Monetary) to segment customers. In Starbucks' case, there is little difference in Recency or Frequency between the highest and lowest value segments. However, there is a large difference in transaction amount. The Top 80% of customers visit Starbucks four times in 15 days.

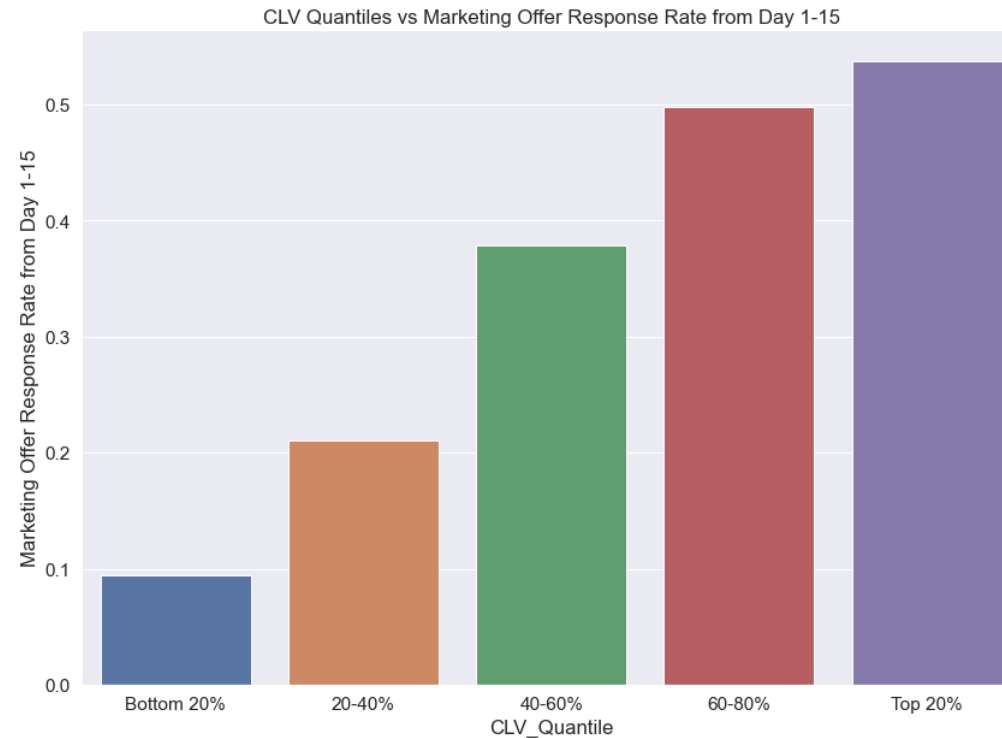
# Top 20% spend ~\$22/visit, Bottom 20% spend \$3.5/visit on average, an \$18.50 difference!



\$3.5 is worth 1 bakery item, \$22 is worth 3 specialty items (sandwich, protein box, and specialty drink).

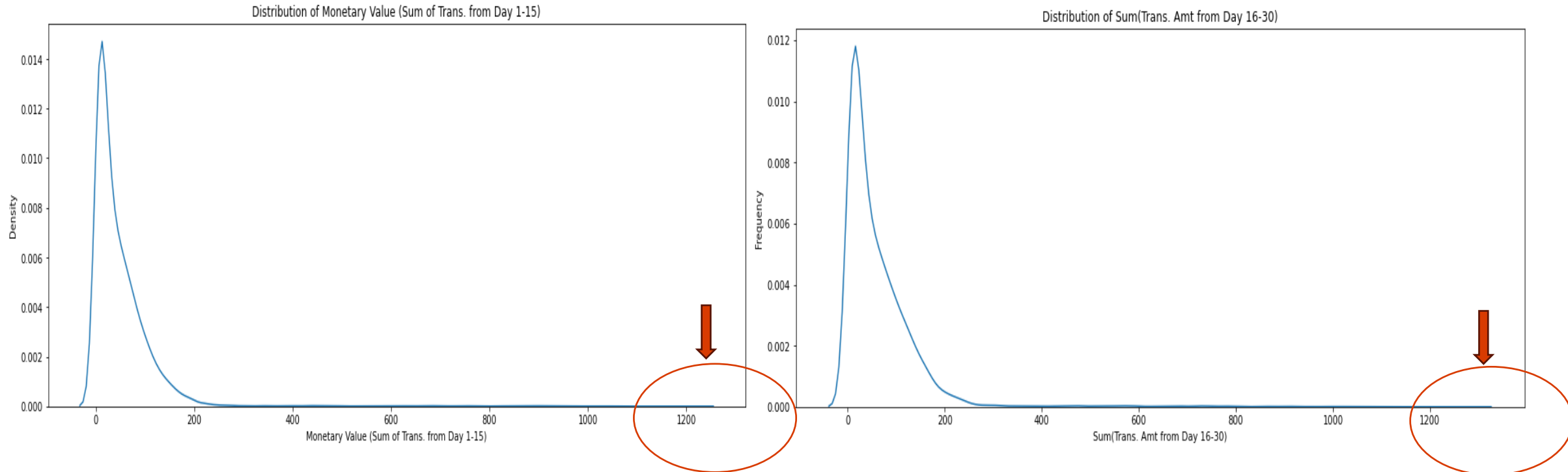
The difference in Average Order Value (AOV) is driving the difference in total spend.

# Response Rate is highly correlated with CLV, with a response rate difference of 44% between the Top 20% & Bottom 20%



As customer lifetime value (CLV) quantiles increase, there is a noticeable rise in the response rate, while the view rate remains relatively consistent across all quantiles. This finding indicates that customers, regardless of their lifetime value, have an equal likelihood of viewing the offer on average. It emphasizes the significance of response rate as a metric for measuring customer engagement.

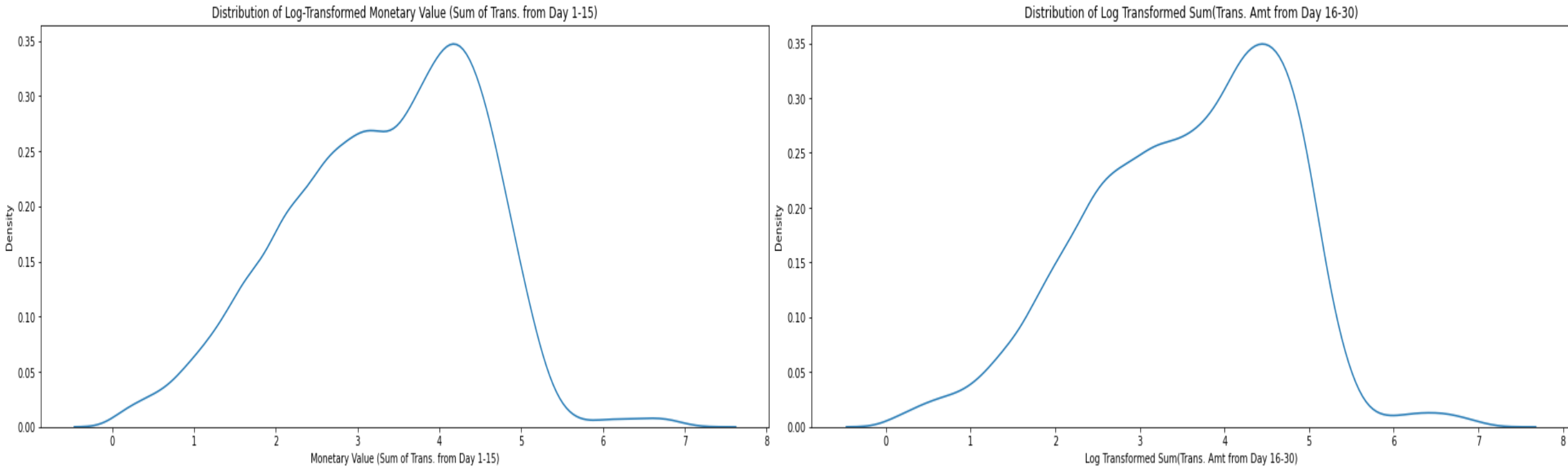
# CLV is highly right-skewed, some customers spend \$1.2K-\$1.3K in 15 days, hinting at corporate card usage for group orders



The median expenditure of a Starbucks customer was \$30 from Day 1-15 and \$40 from Day 16-30. However, the max spend from Day 1-15 is \$1.2K from Day 16-30 is \$1.3K.

The max spender from Day 1-15 spent ~\$600 per visit in 2 visits; the max spender from Day 16-30 spent ~\$500-\$600 per visit in 2 visits. This suggests these customers are probably using a corporate card to place a group order.

# Applying Log Transformation to Monetary Value Feature & Target Variable reduces skewness

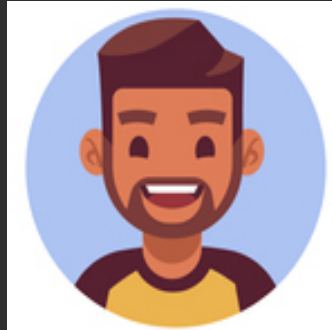


Log transformation can potentially improve model fit and reduces prediction errors (measured by MAE) by reducing skewness and normalizing the distribution.



# Unveiling Customer Profiles using Descriptive Statistics: Demographics, Buying Patterns, & Engagement Levels By Quantile

Occasional Buyers (\$0.6M/yr)   Casual Shoppers (\$1.6M/yr)   Regular Patrons (\$3.3M/yr)   Engaged Members (\$5.3M/yr)   Top Spenders (\$9.6M/yr)



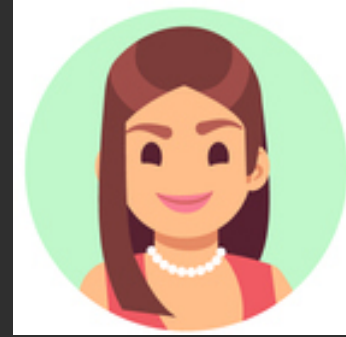
Customers	Gender (M/F Ratio)	Income Distribution (Top 2-3)	Average Frequency	Average Monetary Value (Day 1-15)	Average View Rate	Average Response Rate	Spent More of Less in Day 16-30 than Day 1-15	Average Sum of Transactions (Day 16-30)
Occasional Buyers (Bottom 20%) [\$0-\$14.65]	79% / 21%	30k-50k (56%), 50-75k (43%)	3	\$12	58%	11%	Less	\$6
Casual Shoppers (Bottom 20%-40%) [\$14.65-\$34.30]	69% / 31%	50-75k (47%), 30-50K (39%), 75-100k (11%)	4	\$29	63%	24%	Less	\$19
Regular Patrons (40%-60%) [\$34.30-\$64.49]	59% / 41%	50-75k (45%), 30-50k (29%), 75-100k (20%)	4	\$53	68%	38%	Less	\$40
Engaged Members (60-80%) [\$64.49-\$107.86]	49% / 51%	50-75k (42%), 75-100k (29%)	4	\$73	71%	50%	More	\$77
Top Spenders (Top 20%) [\$107.86-\$1287.25]	46% / 54%	50-75k (41%), 75-100k (37%)	4	\$92	72%	54%	More	\$179

# Super Spenders vs Rest of Population: Demographics, Buying Patterns, & Engagement Levels

Rest of Population (\$14.4M/year)



Super Spenders (\$6M/year)



Customers	Gender (M/F Ratio)	Income Distribution (Top 3)	Average Frequency	Average Monetary Value (Day 1-15)	Average View Rate	Average Response Rate	Spent More of Less in Day 16-30 than Day 1-15	Average Sum of Transactions (Day 16-30)
Rest of Population (0-90%) [\$0-\$144.35)	60% / 40%	50-75k (44%), 30k-50k (30%), 75k-100k (20%)	4	\$52	67%	36%	Less	\$50
Super Spenders (Top 10%) [\$144.35-\$1,287.25)	44% / 56%	50k-75k (39%), 75k-100k (39%), 100k-120k (14%)	4	\$97	72%	54%	More	\$252

# Insights from Super Spenders (Top 10%) Analysis

**Spending Amount:** The most significant finding is the difference in spending amounts between the Super Spenders and the Rest of the Population. In both halves of the month, Super Spenders spend considerably more than the Rest of the Population. Super Spenders have a significantly higher average sum of transactions in the second half of the month compared to the Rest of the Population (\$252 vs \$50).

**Response Rate:** Super Spenders have a higher response rate than the Rest of the Population (54% vs 36%). This is a big difference and may indicate higher engagement or satisfaction among Super Spenders. **Action:** Are Super Spenders more responsive to certain types of marketing or specific products? Use this information to enhance customer engagement strategies.

**Gender Distribution:** Super Spenders have a higher proportion of females (56%) compared to the Rest of the Population (40%). This could indicate a gender bias in spending patterns. **Action:** Consider whether gender-specific marketing or product strategies could be beneficial.

**Income Distribution:** There's a difference in income distribution between the two groups, where there are considerably more individuals that make 75k and above in the Super Spender category. Spending patterns of Super Spenders are influenced by their disposable income.

**Frequency of Transactions:** Both groups have the same average frequency of transactions, suggesting that the total spend is more a function of amount per transaction than frequency. **Action:** Focus on strategies that increase the amount spent per transaction rather than the number of transactions.

# Benefits of Predictive CLV

**Strategic Resource Allocation:** Predictive CLV helps allocate resources strategically by identifying high-value customers and focusing marketing efforts, retention strategies, and resources on them.

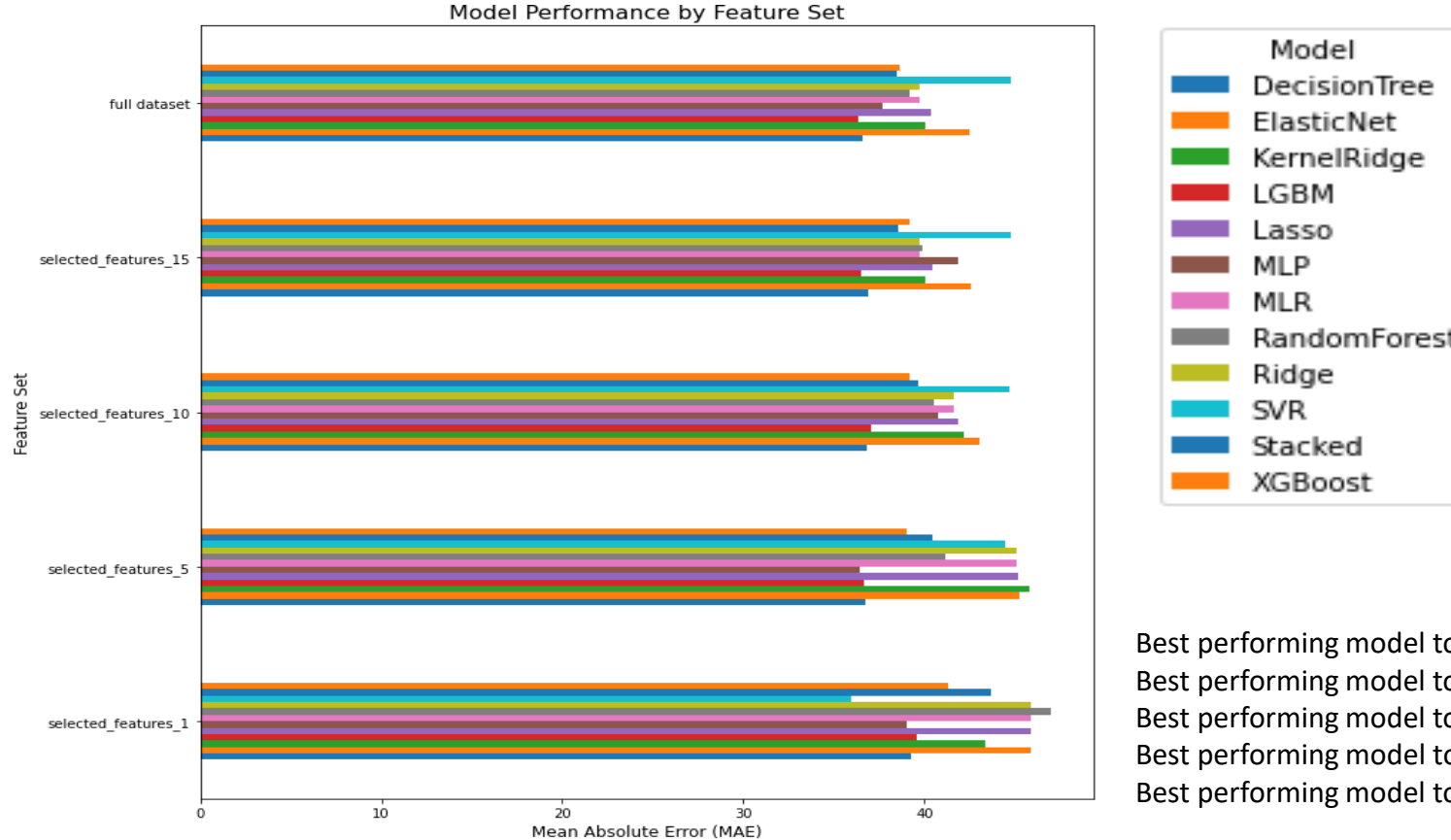
**Customer Segmentation:** Predictive CLV enables effective customer segmentation based on value and behavior, facilitating personalized experiences and tailored retention efforts, increasing customer satisfaction and loyalty.


**Customer Acquisition Optimization:** Predictive CLV optimizes customer acquisition costs by estimating the potential value of customers, allowing businesses to focus marketing efforts on acquiring high-value customers to achieve better ROI.

**Revenue Forecasting:** Predictive CLV provides insights into future revenue generation by predicting customer spending patterns, enabling businesses to forecast revenue, plan for growth, and make informed decisions.

**Improved Customer Relationship Management (CRM):** Predictive CLV enhances customer relationship management by understanding preferences, behaviors, and lifetime value, leading to personalized interactions and relevant offers.

# Selected LGBM as Winner Model utilizing Full Feature Set



Best performing model to minimize MAE for full dataset: LGBM (MAE: 36.32) 

Best performing model to minimize MAE for selected\_features\_15: LGBM (MAE: 36.50)

Best performing model to minimize MAE for selected\_features\_10: DecisionTree (MAE: 36.84)

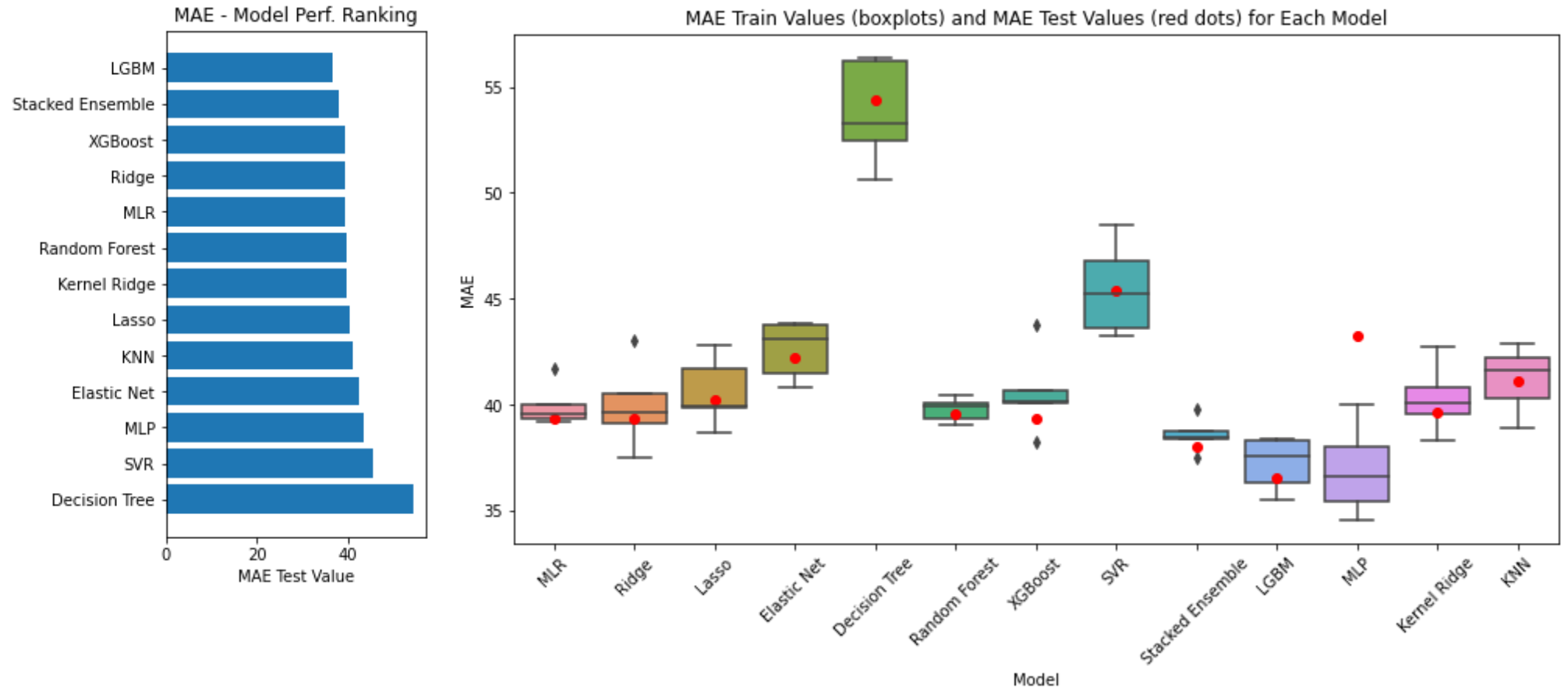
Best performing model to minimize MAE for selected\_features\_5: MLP (MAE: 36.42)

Best performing model to minimize MAE for selected\_features\_1: SVR (MAE: 35.96)

Note: AOV and offers completed by channel features are dropped upfront to reduce multicollinearity because they are derived from existing columns in the dataset.

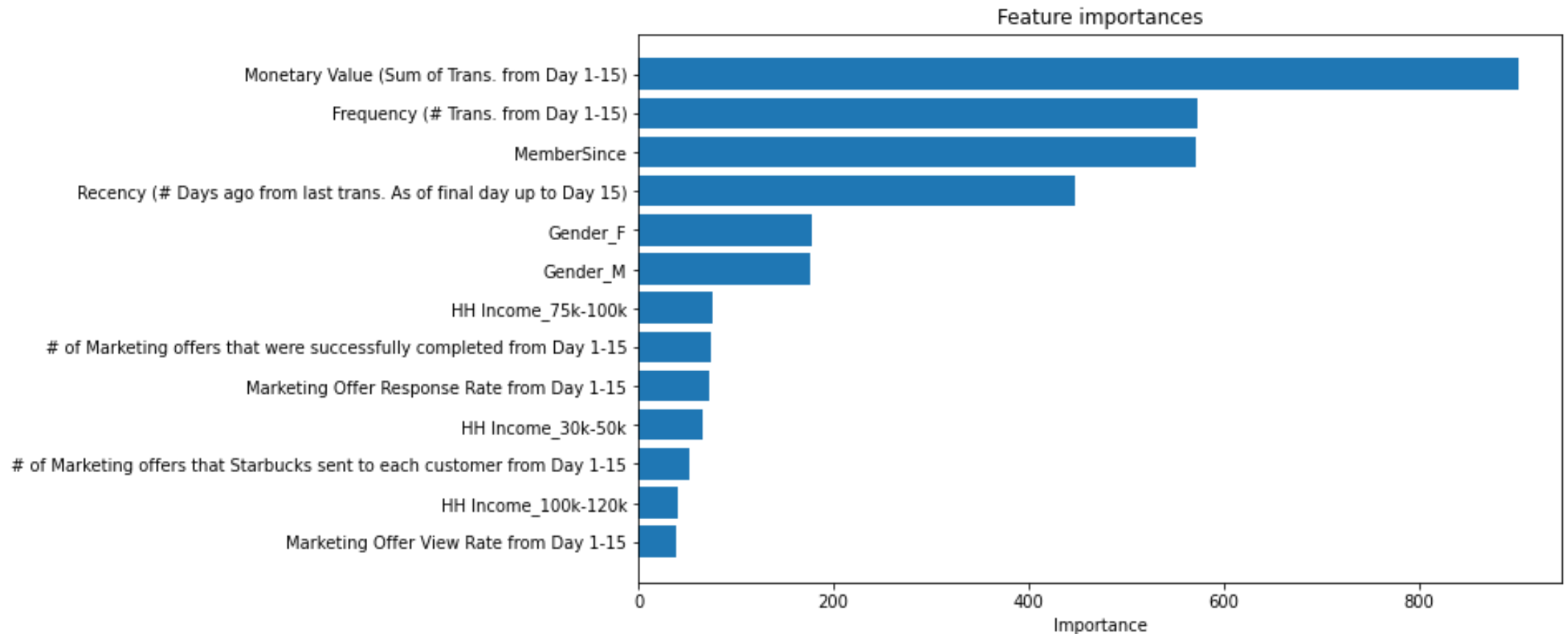
Selecting LGBM with the full dataset is a good decision because it performed relatively well, with an MAE of 36.32. Using the complete dataset helps LGBM capture complex patterns and provide more accurate predictions.

# LGBM emerged as Winner Model during model comparisons by minimizing Test MAE and maintaining a lower MAE range in K-fold cross-validation

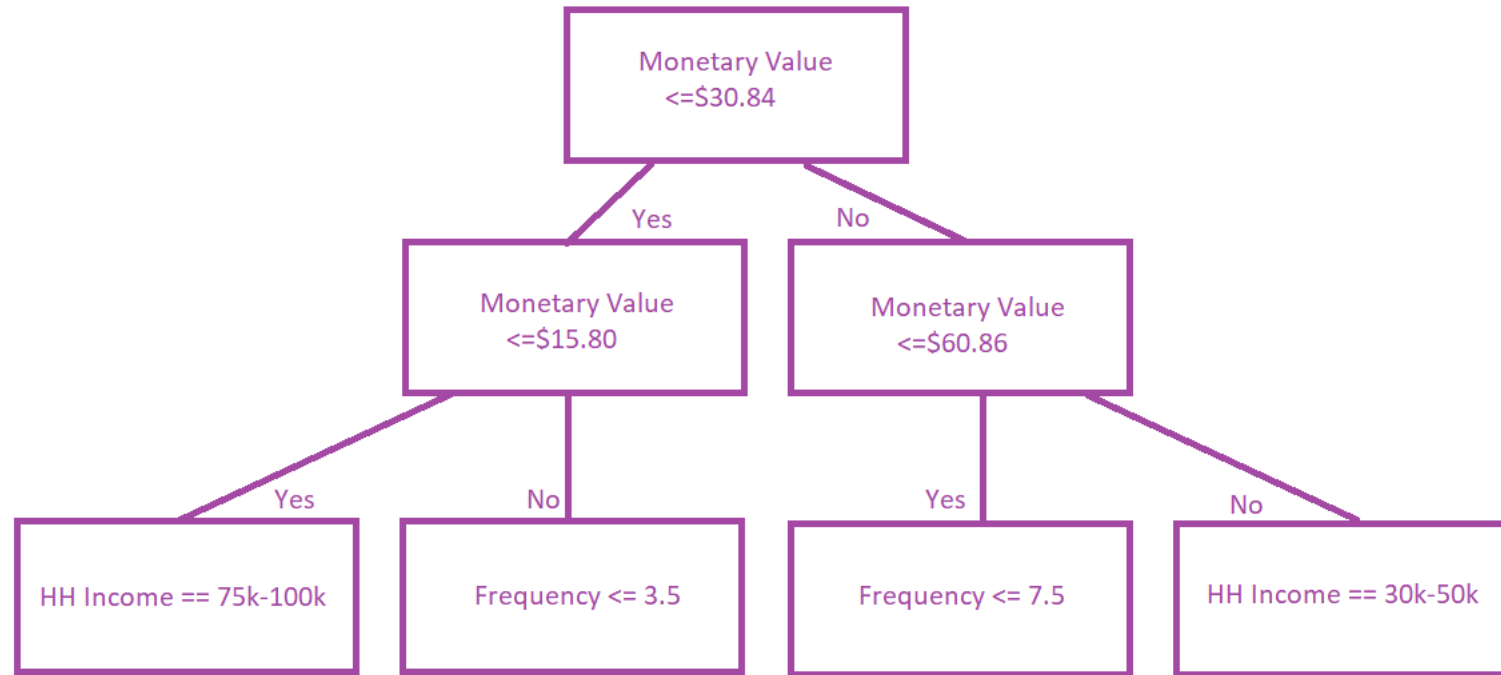


LGBM model is chosen because it minimizes the MAE. For Kfold CV=5, Shuffle=True, the MAE Train value range is lower and MAE Test value is the lowest vs other models. MAE measures how close the predicted values are to the actual values on average.

# LGBM Feature Importance shows that Monetary Value and Frequency are most important features



**Example of one LGBM Tree Split. It analyzes Spending, HH Income, Frequency in its initial branches, identifying influential predictors. This is consistent with feature importance.**



Initial splits capture important predictors. By considering these key features early on, the algorithm can identify the most relevant variables and create a strong foundation for subsequent splits. This approach allows for more effective and efficient decision-making, potentially leading to improved predictions and model performance.



# Winner Model: LightGBM without log transformation predicts Regular Patrons (60-80% Quantile) the best, predictions impacted by skew

Quantile	Avg. Actual CLV	Avg. Predicted CLV	MAE	MAPE
Bottom 20%	\$10.81	\$18.01	22.83	129.8%
20-40%	\$26.0	\$40.38	33.33	87.4%
40-60%	\$54.95	\$82.1	26.57	37.2%
60-80%	\$99.23	\$107.49	24.76	21.2%
80-90%	\$135.07	\$117.79	46.28	30.5%
90-95%	\$169.55	\$126.67	109.94	47.7%
95-99%	\$435.8	\$144.45	631.54	85.7%

The model shows a skew in CLV predictions, overestimating for the Bottom 80% and underestimating for the Top 20%. This is due to an uneven distribution of customers, with a higher concentration of lower-spending customers and relatively fewer high-spending customers, resulting in a weaker signal from the super spenders. Consequently, the tree model struggles to identify meaningful patterns or splits specific to the relatively fewer high-value customers, impacting its ability to accurately capture their behaviors. As a result, the model has difficulty predicting the lowest or highest value customers.

**The model consistently under predicts CLV for Super Spenders (Top 10% Actual CLV is \$144+) because of right-skew.**

User	Predicted CLV	Actual CLV	Difference
User 1	\$114.33	\$170.93	-56.6
User 2	\$122.75	\$201.84	-79.09
User 3	\$127.28	\$195.37	-68.09
User 4	\$107.98	\$219.99	-112.01
User 5	\$78.39	\$350.43	-272.04
User 6	\$104.73	\$171.51	-66.78
User 7	\$107.25	\$436.78	-329.53
User 8	\$120.03	\$183.9	-63.87
User 9	\$115.78	\$173.68	-57.9
User 10	\$143.81	\$172.54	-28.73

# Applying 95% Cap on Monetary Value & Target Variable improved MAE & MAPE for Bottom 60%

Quantile	Avg. Actual CLV	Avg. Predicted CLV	MAE	MAPE
Bottom 20%	\$10.81	\$15.51	18.71	105.8%
20-40%	\$26.0	\$36.77	27.14	71.7%
40-60%	\$54.95	\$72.06	19.88	27.7%
60-80%	\$99.23	\$93.89	30.80	26.2%
80-90%	\$135.07	\$102.38	59.30	39.1%
90-95%	\$169.55	\$109.17	78.48	45.8%
95-99%	\$171.24	\$118.5	78.67	45.9%

Applying a 95% Cap on the monetary feature and target variable improves predictions for Bottom 60% segment.

# Applying a Log Transformation on the Monetary Value & Target Variable to this dataset do not improve predictions for High-Value Segments

Quantile	Avg. Actual CLV	Avg. Predicted CLV	MAE	MAPE
Bottom 20%	\$10.81	\$11.21	16.94	88.6%
20-40%	\$26.00	\$30.79	22.33	55.5%
40-60%	\$54.95	\$62.64	20.89	26.8%
60-80%	\$99.23	\$85.54	41.41	34.0%
80-90%	\$135.07	\$94.60	74.27	47.1%
90-95%	\$169.55	\$99.94	151.27	62.0%
95-99%	\$435.80	\$110.89	657.88	88.9%

Log transformation reduces the variability in error metrics across quantiles and improves predictions for the Bottom 60% of customers. However, its performance is less effective in predicting the Top 40% or high-value customers. Therefore, log transformation does not produce a better model fit.

# Applying a 95% Cap AND Log Transformation on the Monetary Value & Target Variable to this dataset do not improve predictions for High-Value Segments

Quantile	Avg. Actual CLV	Avg. Predicted CLV	MAE	MAPE
Bottom 20%	\$10.81	\$8.23	12.36	70.5%
20-40%	\$26.00	\$20.31	21.59	56.2%
40-60%	\$54.95	\$40.91	27.04	36.0%
60-80%	\$99.23	\$72.13	53.28	45.7%
80-90%	\$135.07	\$84.74	83.63	55.4%
90-95%	\$169.55	\$91.82	102.40	59.8%
95-99%	\$171.24	\$101.16	102.13	59.6%

Consistent with the prior slide, log transformation reduces the variability in error metrics across quantiles and improves predictions for the Bottom 40% of customers. However, it does not perform as well in predicting the Top 40% or high-value customers. Therefore, log transformation does not produce a better model fit.

**Is modeling necessary to improve predictions? While past spend appears to be a good signal for future spend to segment users on CLV, we can get a slightly better result by including the full feature set.**

Quantile	Avg. Actual CLV	Avg. Predicted CLV	MAE	MAPE
Bottom 20%	\$10.81	\$8.65	17.67	99.9%
20-40%	\$26.0	\$21.07	33.33	87.0%
40-60%	\$54.95	\$44.57	41.75	55.4%
60-80%	\$99.23	\$78.63	60.64	51.9%
80-90%	\$135.07	\$110.27	81.53	54.0%
90-95%	\$169.55	\$141.66	131.01	58.2%
95-99%	\$435.8	\$245.66	653.16	89.2%

This approach only uses the monetary value from the first 15 days feature to forecast the sum of transactions from the 16th to 30th day without any predictive modeling.

# Insights and Actionable Recommendation

The best method to improve predictions was using a hyper-parameter tuned LGBM model without log transformation.

To address the right skew in the data, a log transformation was tested on the Monetary Value and Target Variable. While this transformation successfully reduced variability in error metrics and improved predictions for the Bottom 60% of customers, it surprisingly struggled in predicting the Top 40%. Therefore, the log transformation did not assist in predicting high-value segments with this dataset. It's worth noting that the dataset has limitations, as CLV is usually measured over a longer time horizon or with a richer set of features.

While past spend appears to be a good signal for future spend to segment users on CLV, we can get a slightly better result by including the full feature set. Utilizing predictive modeling is the preferred approach for CLV segmentation.

Since the average frequency of purchases is similar between the Top 10% and the rest of the population, it is recommended to focus on strategies that increase the amount spent per transaction (Average Order Value) rather than the number of transactions.

# Future Work

## **Introduce Starbucks Diamond Loyalty Program**

Implement a tailored loyalty program, "Starbucks Diamond," specifically designed for "super spenders" in the top 10% of CLV. As a short-term strategy, include customers spending \$200 or more per month. For the long-term, identify potential upgraders by approaching it as a binary classification problem, focusing on users displaying behaviors indicating their likelihood to upgrade. Identifying and retaining high-value customers is often more cost-effective than acquiring new ones.

## **Prioritize Regular Patrons (60-80% Quantile) as a Retention Strategy**

As the LGBM model without preprocessing (without 95% Cap or Log Transformation) predicts the Regular Patrons (60-80% Quantile) segment the best with the lowest errors, we can focus on encouraging these customers to transition into higher-value segments. Note that the Engaged Members and Top Spenders segments exhibit strong response rates and spending habits, so they would require less attention in terms of retention efforts.

## **Use K-Means Clustering to Create Models for Each Customer Segment**

Utilize K-Means Clustering to create customer segment models, to uncover other relevant patterns about customers, enabling personalized experiences and targeted communications. This approach allows businesses to focus on valuable segments and allocate resources effectively.



# THANK YOU!

**CONTACT:**



# Debbie Trinh



[www.linkedin.com/in/debbietrinh](https://www.linkedin.com/in/debbietrinh)

