**University of Texas at Dallas**

**Report**

**Analytics Practicum – BUAN 6390.501**

**Milestone 3 --** **Draft Results & Model Evaluation.**

**Professor: Sunela Thomas**

**Group: 01**

**Melody Tavallaee**

**Harshita Yakkantii**

**Bhavith Akula**

# **Introduction:**

In previous stages of the project, we completed comprehensive data preparation, exploratory analysis, and initial model development aimed at understanding customer behavior and segmenting customers effectively. In this milestone, we focus on critically evaluating our predictive models and segmentation strategies to ensure their accuracy, reliability, and business relevance.

We assess model performance using key evaluation metrics, identify the most influential features driving customer spending, and conduct detailed error analysis to uncover model strengths and weaknesses. Additionally, we examine potential biases and limitations, ensuring that our findings are both statistically valid and practically applicable.

Beyond technical evaluation, we translate our results into actionable and meaningful business insights, offering clear recommendations to enhance customer engagement, optimize marketing strategies, and support revenue growth. This milestone plays a vital role in validating the robustness of our approach and setting the stage for final business-driven solutions.

# **Model Performance Evaluation**

## **Predictive Models Overview**

* **Models Used:** Linear Regression and Ridge Regression
* **Target Variable:** Total Spending (calculated as Quantity × Price)
* **Input Features:** Age, Quantity, and Price
* **Data Split:** 80% Training set, 20% Testing set

## **Evaluation Metrics**

* **Linear Regression**
  + R-squared (Training): 0.945
  + R-squared (Testing): 0.942
  + RMSE (Testing): ~480
* **Ridge Regression**
  + R-squared (Training): 0.943
  + R-squared (Testing): 0.941
  + RMSE (Testing): ~470

## **Interpretation**

Both models performed well in predicting total customer spending, with Ridge Regression slightly outperforming Linear Regression in terms of generalization. Ultimately, the Ridge's L2 regularization helped reduce potential overfitting, making it a more reliable model for future predictions.

# **Feature Importance and Error Analysis**

## **Feature Importance**

**Using coefficient analysis from Ridge Regression:**

* **Price**: Most important predictor of total spending (strong positive correlation)
* **Quantity**: Secondary strong contributor (positive relationship)
* **Age**: Relatively minor impact compared to Price and Quantity

A graph with blue squares

AI-generated content may be incorrect.

The Feature Importance Bar Chart highlights the relative impact of each feature on predicting customer spending. Price emerged as the most influential predictor, followed by Quantity, while Age had a minimal contribution. This confirms that purchasing behavior is primarily driven by transaction size and quantity rather than customer demographics.

## **Error Analysis**

* **Low-spending prediction errors**: Observed underestimation in predicting very low transaction amounts (e.g., < $100).
* **High-spending outliers**: Some high-value purchases (> $3,000) were not perfectly predicted, contributing to slightly larger residuals.
* **General trend**: Majority of predictions fell within acceptable error margins.

A graph with blue dots and a red line

AI-generated content may be incorrect.

Residual plots showed no significant heteroscedasticity, indicating the models' assumptions largely hold.

# **Assessment of Model Limitations and Potential Biases**

## **Limitations**

* **Feature Set Simplicity**: Only a limited number of features (Age, Quantity, Price) were used. Including additional variables like "payment method" or "shopping mall" might improve our ability to predict customer behaviors and audience segments.
* **Temporal Effects Ignored**: Customer spending habits may vary based on the time of year that they are shopping. Spending patterns over time (e.g., seasonality) were not incorporated into the models.
* **Model Linearity Assumption**: Both Linear and Ridge Regression assume linear relationships; real-world spending behavior may involve non-linear factors.

## **Potential Biases**

* **Sampling Bias**: The dataset only includes customers who made purchases; non-purchasing customer behavior is unknown. It could be beneficial to have data about the general population around the malls in order to use marketing to target other potential consumers and expand the customer base to include more people.
* **Age Bias**: Certain age groups (very young or very old) are underrepresented, potentially skewing generalizations.
* **Product Category Effects**: Models do not account for differences across product categories, which could influence spending behavior.

## **Insights for Improving Customer Engagement and Revenue.**

## **Insights**

### **1. High-Spenders**

Customers who purchase in larger quantities and at higher price points consistently contribute the most to overall revenue. These individuals typically exhibit strong purchasing power and lower price sensitivity, making them ideal targets for premium product offerings, loyalty programs, and exclusive promotions. Identifying and nurturing these high-value customers can significantly enhance profitability and brand loyalty.

### **2. Customer Segments**

Through clustering analysis, we identified distinct groups such as "High-Value Customers" and "Loyal Customers," both of which demonstrate significant revenue potential. High-Value Customers generate the highest total spending, while Loyal Customers exhibit consistent purchase behavior over time. Tailored marketing strategies—such as personalized discounts or membership rewards—can further deepen engagement with these segments and maximize customer lifetime value.

### **3. Spending Behavior**

Our analysis revealed that a majority of customers tend to concentrate their spending below the $1,000 threshold. This pattern suggests an inherent price sensitivity among the broader customer base. Businesses can leverage this insight by designing price-optimized product bundles, promotional discounts, and targeted campaigns that align with common spending behaviors, ultimately increasing conversion rates and average transaction values.

A graph with a line

AI-generated content may be incorrect.

The Elbow Method plot helps identify the optimal number of clusters by plotting the inertia score against different values of k. The plot shows a clear bend ("elbow") around k=4, suggesting that four clusters provide a good balance between model simplicity and segmentation accuracy. Based on this, we selected four distinct customer segments for further analysis.

## **Recommendations**

* **Targeted Promotions**: Focus marketing efforts on "High-Value Customers" with loyalty programs and exclusive deals. We also know that clothing is the most popular category amongst the customers, so we can leverage that knowledge to promote items in that category.
* **Personalized Engagement**: Use purchase history to suggest bundles or upsell opportunities to "Loyal Customers." Catering to customers’ personalized needs may encourage them to continue purchasing more items.
* **Retention Strategies**: Design re-engagement campaigns for "One-Time Buyers" to encourage repeat purchases. We see a potential to expand our customer base, and we can use remarketing strategies to convert “One-Time Buyers” to more loyal customers.
* **Price Optimization**: Explore strategic discounting for price-sensitive segments identified in the spending distribution analysis.

# **Conclusion:**

Our analysis demonstrates that the Ridge Regression model provides a highly accurate and generalized prediction of customer spending behavior. By leveraging RFM analysis combined with K-Means clustering, we successfully segmented customers into meaningful groups, offering actionable insights for targeted marketing strategies and customer engagement initiatives.

These models validate the underlying business hypotheses and provide a practical framework for optimizing customer relationships, identifying high-value segments, and personalizing outreach efforts. Ridge Regression helped mitigate overfitting issues, while clustering revealed distinct behavioral patterns critical for strategic decision-making.

However, while our current models show strong performance, there are clear opportunities for future improvement. Incorporating additional behavioral variables, capturing seasonality and purchasing trends, and experimenting with non-linear and ensemble modeling approaches could enhance predictive accuracy and deepen customer understanding.

Overall, this milestone establishes a robust analytical foundation for data-driven decision-making, bridging technical insights with business strategy to support sustainable growth and competitive advantage.