Assignment 2 – Machine Learning

Data Preparation and Feature Engineering

The first step in the analysis involved loading and inspecting each dataset for completeness, data types, and duplicates, addressing missing values wherever present. Given that transaction and marketing data were not at the customer level, feature engineering was undertaken to create meaningful features that aggregated this data at the customer level. This process included categorizing ages, amalgamating engagement metrics into a numeric score, and creating new metrics such as the number of campaigns each customer was exposed to and responded to, the most common campaign each customer accepted, and various transaction-related metrics. By joining all tables at the customer level and conducting thorough checks for inconsistencies, the data was prepared for further analysis. Additional features such as customer life (last purchase date - join date) and a composite RFM score were created to better understand and predict CLV.

Exploratory Data Analysis (EDA)

The EDA revealed several key insights. Age did not appear to be a strong predictor of CLV, with similar distributions of CLV across different age groups. Gender also showed no significant relationship with CLV. However, a positive relationship was observed between the total number of transactions and CLV, as well as between frequency and monetary value of purchases and CLV. Interestingly, while the number of campaigns a customer was exposed to did not significantly affect CLV, participating in at least one successful campaign did show a positive impact. The stability of CLV across different categories suggested that successful campaigns helped, but were not the sole drivers of high CLV.

Categories and Promotions: A deeper dive into spending patterns across product categories revealed that electronics had the highest contribution to CLV (57.51%), followed by home goods (28.27%), and clothing (14.22%). This indicated that customers who spent more on electronics tended to have a higher lifetime value. Promotion types also played a significant role in influencing CLV. High-value promotions such as "Buy One Get One" and "Free Shipping" were found to attract higher CLV customers, while discount promotions, though effective in engaging customers, generally attracted those with lower lifetime values. Customers who were not influenced by specific promotions still maintained a reasonable baseline CLV, suggesting a core group of consistently valuable customers.

Combined Insights: The combined insights from the analysis highlighted several strategic opportunities for EcomX Retailers. First, the effectiveness of promotions was clear: "Buy One Get One" offered the most engagement and highest CLV, followed by "Free Shipping." Discounts, while engaging, yielded lower CLV. This suggested that EcomX should prioritize high-value promotions to maximize customer lifetime value. Secondly, there was a segment of non-responsive customers with moderate CLV, indicating potential for targeted strategies that did not rely on direct promotions. By focusing on high-value promotions and refining strategies for non-responsive segments, EcomX could enhance customer retention and value.

Correlation Matrix Findings: The correlation analysis provided further insights into the factors influencing CLV. Engagement metrics, both individually and collectively, showed no strong relationship with CLV. However, customer life showed a negative correlation with average CLV per month, suggesting that customers might spend more initially and then reduce spending over time. Recency had a negligible correlation with CLV, while frequency and monetary value showed strong negative correlations, indicating that lower scores (implying higher purchase frequency and spending) were associated with higher CLV. Among product categories, electronics spending had the highest contribution to CLV, followed by home goods and clothing.

Model Building

For model building, regression models were used to predict continuous values of CLV and average CLV per month, and classification models were used to predict categories. Separate models were built for continuous and classification prediction. For classification, models for both binary and multiclass classification were developed. For continuous predictions, separate models for CLV and average CLV per month were built, while classification models focused on categorizing CLV values. Linear regression was employed for CLV prediction and KNN regressor for average CLV per month, with model selection based on cross-validation scores between linear regression and KNN regressor. For classification, Naive Bayes was used for binary classification and KNN for multiclass classification, again selected based on cross-validation scores. Two versions were created for each model: one with an RFM composite score (sum of R, F, M scores) and another with separate R, F, M scores to determine the best approach for prediction.

Details of all models are mentioned in Table 1, and details of all input variables are in Table 2 in Appendix.

Model Evaluation and Results

For predicting Total Customer Lifetime Value (CLV), the best model is Model 2, which uses a linear regression approach with Recency, Frequency, and Monetary value treated as distinct features. This model outperforms others in terms of R-squared score and error metrics, providing nuanced and valuable insights that allow for a better understanding of each feature's impact on CLV. Leveraging these insights can enhance EcomX Retailers' marketing strategies, resource allocation, and customer retention efforts, ultimately driving higher profitability.

When predicting Binary Classes of CLV, Model 6 is the most effective. It boasts high accuracy (93.05%), precision (97.55%), and a well-balanced F1-score (92.65%). This model surpasses its predecessor (Model 5) across all key metrics, making it the best choice for accurately identifying high-value customers. The model's reliable performance enables more precise customer segmentation, targeted marketing, and efficient resource allocation, significantly enhancing EcomX's customer engagement and retention strategies.

For Multiclass CLV prediction, Model 7, a multiclass logistic regression model, is the preferred choice despite its lower accuracy (75.05%) compared to binary models. It achieves a reasonable balance with precision (74.72%), recall (75.05%), and an F1-Score of 74.83%. Although it is less effective due to the complexity of predicting multiple classes, it provides sufficient accuracy and balance for EcomX Retailers to use in customer segmentation and marketing efforts. For predicting average CLV per month, none of the models provided demonstrated adequate predictive power, indicating a need for alternative modeling techniques or improved feature sets.

Key Business Impact

By leveraging Model 2 (Linear Regression with Recency, Frequency, and Monetary as distinct features), Model 6 (High-Accuracy Binary Classification), and Model 7 (Multiclass Logistic Regression), EcomX Retailers can significantly enhance their marketing strategies and customer retention efforts. The following assumptions and calculations underpin these results:

- ***Model 2 (Linear Regression):*** Assumes an average CLV increase of 15% for the top 10% of high-value customers. With a current average CLV of $1,000 (assumed), this translates to an increase of $150 per customer. For 1,000 high-value customers, this results in a potential revenue increase of $150,000 through targeted marketing campaigns.
- ***Model 6 (High-Accuracy Binary Classification):*** Assumes targeted marketing campaigns improve response rates by 10%. With a current response rate of 5% (assumed) for 10,000 marketing messages and an average purchase value of $50, the improved response rate of 5.5% leads to an additional revenue of $2,500. The model's high precision (97.55%) and recall (88.22%) ensure marketing efforts are well-targeted, minimizing wasted resources and maximizing ROI.
- ***Model 7 (Multiclass Logistic Regression):*** Assumes personalized retention strategies effectively engage customers. The model's insights support identifying and targeting 250 additional high-value customers with an average CLV of $1,000, resulting in an additional revenue of $250,000.

Overall, the combined impact of these models facilitates strategic decision-making for upselling and cross-selling, driving an additional $50,000 in revenue. This holistic approach, powered by predictive modeling, results in a total estimated incremental revenue of $452,500, allowing EcomX Retailers to optimize resource allocation, improve customer engagement, and achieve substantial business growth and profitability.

Conclusion and Recommendations

By integrating advanced predictive models, EcomX Retailers can significantly enhance marketing effectiveness, customer retention, and revenue. The combined use of Model 2 for CLV prediction, Model 6 for high-accuracy binary classification, and Model 7 for multiclass classification enables targeted marketing and personalized retention strategies, leading to an estimated additional revenue of $452,500.

For optimal results, following are advised:

- Continuous Improvement: Regularly update models with new data and refine parameters to maintain accuracy.
- Enhanced Data Collection: Collect more comprehensive customer data and integrate additional features to improve CLV predictions further.

Appendix

Number	Model Type	Prediction Target	Key Metrics
1	Linear Regression	Total Customer Lifetime Value	R-squared score: 0.87
2	Linear Regression	Total Customer Lifetime Value	Outperformed Model 1 in R-squared score and error metrics
3	KNN Regression	Average CLV per Month (Version 1)	R-squared score: 0.32
4	KNN Regression	Average CLV per Month (Version 2)	R-squared score: 0.33
5	Binary Classification	Binary CLV	Accuracy: 90.25%, Precision: 93.56%, Recall: 86.30%, F1-score: 89.79%
6	Binary Classification	Binary CLV	Accuracy: 93.05%, Precision: 97.55%, Recall: 88.22%, F1-score: 92.65%
7	Multiclass Logistic Regression	Multiclass CLV	Accuracy: 75.05%, Precision: 74.72%, Recall: 75.05%, F1-score: 74.83%
8	Multiclass Classification	Multiclass CLV	Accuracy: 26%, Precision: 12.82%, F1-score: 17.17%

Table 1 - Model Details

Version	Input Features
1	['total_transactions', 'num_campaigns', 'num_yes_campaigns', 'cust_life', 'Engagement_Score', 'RFM_Composite', 'most_common_yes_promotion']
2	['total_transactions', 'num_campaigns', 'num_yes_campaigns', 'cust_life', 'Engagement_Score', 'R', 'F', 'M', 'most_common_yes_promotion']

Table 2 - Input Variables