Alvina Aamir (221749742)

MBAN Data Science Summary Report Assignment 2

Analysis Approach

The analysis began with loading the data and conducting initial exploration; this involved importing relevant libraries (pandas, matplotlib, seaborn), uploading the dataset, and displaying the first rows. Next, data cleaning and preprocessing were completed where missing values were checked and handled, data types were displayed and converted if necessary (date columns to datetime), and new columns were created for customer tenure. Descriptive statistics for the numerical columns and frequency distribution for the categorical variables were completed to summarize the data and understand their distributions. Visualization tools such as histograms and heatmaps were used to visualize the data. Feature engineering was performed to create new features, such as early behavior indicators and normalized monetary value. Specifically, customer tenure in days was calculated, and monetary value was normalized by dividing the total transaction amount by customer tenure. Early behavior indicators such as early total spent, early frequency, and early recency were derived for the first 30 days of customer activity. For modeling, two models were selected: Linear Regression and K-Nearest Neighbors (KNN). Cross-validation was performed to evaluate the models, and grid search was used to optimize hyperparameters for the KNN model. Performance metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), were used to evaluate and compare the models.To obtain specific insights, customers were segmented into high and low early spenders based on spending more or less than $200 in the first 30 days, high and low frequency based on having more or less than 5 transactions in the first 30 days, and based on engagement metrics such as site visits and email opens. The average CLV for these segments was calculated by normalizing the monetary value and multiplying it by the average customer tenure days within each segment.

Key Metrics

The key metrics used in the analysis included Customer Tenure, Early Total Spent, Early Frequency, Early Recency, and Normalized Monetary Value. Customer Tenure, measured in days, helps in understanding customer engagement and the effectiveness of retention strategies. Early Total Spent is the total amount spent by the customer within the first 30 days, indicating early customer value and potential long-term value. Early Frequency, the number of transactions within the first 30 days, shows early engagement levels. Early Recency, the number of days since the last transaction within the first 30 days, indicates recent activity and engagement. Normalized Monetary Value, the total transaction amount divided by customer tenure, normalizes the monetary value based on how long they have been a customer, providing a fair comparison.

Findings and Insights

From the visualization and statistical tools, it was found that early spending and frequency are significant predictors of CLV. Specifically, customers who spent more than $200 in their first 30 days had an average CLV of $140,644.97, compared to an average CLV of $18,106.69 for those who spent less than $50. This shows a strong correlation between early spending and long-term value. Similarly, customers with more than 5 transactions in the first 30 days had an average CLV of $133,000.65, compared to $16,858.11 for those with fewer than 2 transactions. Engagement metrics like the number of site visits and emails opened also correlate with higher CLV. Customers who visited the site more than 10 times in the first month had an average CLV of $43,089.23, while those who visited fewer than 3 times had an average CLV of $168,847.89. Those who opened more than 5 promotional emails had an average CLV of $39,445.39, compared to those who did not engage with emails (data unavailable). Another key insight was the importance of normalizing monetary value by customer tenure. It was observed that customers with a high normalized monetary value (greater than $2 per day) consistently show higher engagement and loyalty metrics. This normalization allows for a fair comparison across customers who have been with EcomX for varying lengths of time, reaffirming the importance of sustained customer interaction and targeted marketing efforts.

Model Performance and Selection

KNN and Linear Regression were chosen for this data because while Linear Regression provides a straightforward baseline for regression tasks, KNN is effective in capturing non-linear relationships, offering a robust comparison for predicting CLV. Here is a detailed breakdown of their performance:

1. **Linear Regression:**
   - **Cross-Validation MAE:** -0.56, **Test MAE:** 0.54, **Test RMSE:** 0.72
   - Linear Regression provided a good baseline but had a higher RMSE compared to the other model, indicating that it might not capture the complexity of the data as well as other models.
2. **K-Nearest Neighbors (KNN):**
   - **Best Params:** {'n_neighbors': 5}, **Test MAE:** 0.47, **Test RMSE:** 0.65
   - KNN showed significant improvement over Linear Regression after hyperparameter tuning with Grid Search, having the lowest RMSE, making it the best performing model in terms of predictive accuracy.

Based on the performance metrics (MAE and RMSE), KNN was selected as the best model. The Grid Search cross-validation helped in optimizing the hyperparameters, ensuring that the best version of the model was used. KNN's ability to capture non-linear relationships in the data contributed to its superior performance.

Recommendations to the Business

Targeted retention strategies should be put in place for high-value customers based on their early spending and engagement behaviors. Implementing targeted marketing strategies for customers who spend more than $200 in their first 30 days can be highly effective, as these customers have an average CLV of $140,644.97. Personalized offers and loyalty programs tailored to these high-value customers can help retain them. Increasing customer engagement through personalized emails and recommendations can also be beneficial. Customers who open more than 5 promotional emails have an average CLV of $39,445.39, so focusing on email engagement strategies can help retain valuable customers. Additionally, customers who visit the site more than 10 times in the first month have an average CLV of $43,089.23, indicating that encouraging frequent site visits through incentives and personalized content can boost CLV. Monitoring early spending and transaction frequency as key indicators of customer value is crucial. Customers with more than 5 transactions in the first 30 days have an average CLV of $133,000.65. By identifying and targeting potential high-value customers early in their lifecycle, EcomX Retailers can implement strategies to retain these customers. Enhancing the overall customer experience is another vital recommendation. This includes improving the website and app usability, offering excellent customer service, and providing personalized content to reduce churn.

Business Impact Estimates

If the recommendations above are followed, quantitatively, the business can potentially increase revenue by targeting and retaining high-value customers. For instance, assuming a 5% improvement in retention rates for high-value customers identified through early behavior, the potential revenue increase could be significant. By improving retention of high-value customers, the revenue can increase by approximately $154,402, assuming a 5% improvement in retention. Qualitatively, the business will see improved customer satisfaction and loyalty, which can lead to a stronger brand reputation, thus improving ROI. By implementing these strategies, EcomX Retailers can effectively increase CLV, improve customer satisfaction, and drive business growth.