JULY 17, 2024

Assignment 2: Building Machine Learning Models

MBAN 6110 S- DATA SCIENCE I

PRIYA CHAUDHURI
221060736

Data Preparation Approach

1) Data Joining and Merging
   - Joined `customer`, `engagement`, and `transaction` datasets using `customer_id` and matched the most recent marketing campaign to each transaction based on `transaction_date` and `campaign_date`.

2) Data Loading and Initial Cleaning
   - Converted `join_date` and `last_purchase_date` to datetime format and ensured `customer_id` is a string across datasets.
   - Filled missing values in `age` with the mean age and `gender` with 'Unknown'.
   - Identified and dropped duplicate rows based on `customer_id` to ensure data integrity.

3) Feature Engineering
   - Total Transaction Amount: Summed transaction amounts for each customer.
   - Frequency of Purchases: Counted the number of transactions for each customer.
   - Recency: Calculated the number of days since the last purchase.
   - Tenure: Calculated the number of days between joining and last purchase.
   - Campaign Details: Aggregated the number of campaigns and responses for each customer.
   - RFM Scores: Created Recency, Frequency, and Monetary (RFM) scores and combined them to form an RFM score.
   - Identified high-value customers based on their RFM scores (>=11).
   - Average Amount Spent per Day: Calculated the average transaction amount per day for each customer.

Key Insights from Trend Analysis

1) The number of site visits, emails opened, and clicks show a right-skewed distribution, indicating a small number of highly active customers.
2) Most customers have recent purchase activities and are involved in one or two campaigns, with varying responses.
3) Product Type Electronics show a higher median and a wider range of spending compared to Clothing and Home Goods
4) Transaction amount is relatively consistent across genders, with some outliers indicating higher spending among females and unknown gender categories. Similar views with Campaign Type.

Feature Selection

Age segments customers for tailored marketing, while site visits, emails opened, and clicks indicate engagement and purchase potential. The number of campaigns and responses reflect customer involvement and responsiveness. Product category reveals spending habits, gender aids in creating targeted campaigns, and promotion type highlights the effectiveness of different promotions. These features together help predict customer lifetime value and identify high-value customers, leading to better-targeted marketing, improved retention, and enhanced business outcomes.

Predicting Amount Spent Per Day

The trend analysis showed a significant right-skew in transaction amounts, site visits, emails opened, and clicks. Predicting the average amount spent per day by each customer allows EcomX to:

- Identify high-revenue customers.
- Understand spending patterns and customer behaviour over time.
- Optimize marketing spend by focusing on customers with higher predicted spending.

Model Evaluation: These models predict exact spending amounts, which is useful for precise financial forecasting and budget management.

- Linear Regression: Mean Squared Error: 0.0018, R-squared: 0.044
- Decision Tree Regressor: Mean Squared Error: 0.0016, R-squared: 0.1505
- Random Forest Regressor: Mean Squared Error: 0.0012, R-squared: 0.3541

Selected Model: Random Forest Regressor performed the best with the highest R-squared value (0.3541) and the lowest Mean Squared Error (0.0012).

Predicting High-Value Customers

The RFM analysis provided a clear segmentation of customers based on their recency, frequency, and monetary values. This approach helps EcomX to:

- Identify top-tier customers who are most valuable to the business.
- Develop targeted marketing strategies and implement loyalty programs and personalized offers to enhance customer experience and loyalty.

Model Evaluation: These models classify customers into high-value or low-value categories, helping tailor marketing strategies and resource allocation.

- KNN: Accuracy: 0.8735 | F1 Score: 0.0307 | Recall: 0.0172 | Precision: 0.1429
- Naive Bayes (NB): Accuracy: 0.8565 | F1 Score: 0.0401 | Recall: 0.0258 | Precision: 0.0909
- Decision Tree (DT): Accuracy: 0.773 | F1 Score: 0.1303 | Recall: 0.1459 | Precision: 0.1176

Selected Model: Decision Tree balances accuracy with much better F1, recall, and precision scores as compared to the other models.

It's capturing low-value customer cases better so the company would now concentrate on the low-value customers for marketing, retention and resource-related strategies.

Estimated Business Impact

Based on the feature importance results, age, number of site visits, and engagement metrics like email opens and clicks are critical in predicting both daily spending and identifying high-value customers. By leveraging these critical features, EcomX Retailers can implement more effective strategies to increase revenue, retain customers, while also identifying low-value customers for better resource allocations.

Predicting Amount Spent Per Day

- Enhanced Revenue Forecasting: Accurate daily spending predictions help create precise budgets, reduce the risk of under- or over-spending, and improve long-term planning for optimal resource allocation.
- Optimized Marketing Spend: Focus marketing on high-spending segments to increase ad efficiency and conversion rates, reducing waste and freeing funds for strategic initiatives.
- Improved Customer Insights: Understanding spending behaviours allows for tailored marketing, improving engagement and satisfaction. Insights guide product development and inventory management, ensuring popular items are stocked and new products meet preferences.

Predicting Low-Value Customers

The confusion matrix suggests that the model is more effective at identifying low-value customers.

- Increased Customer Retention: For low-value customers design cost-effective loyalty programs, personalized campaigns that encourage these customers to become more engaged and potentially increase their CLV over time.
- Resource Allocation: This is critical for managing resources efficiently and ensuring that marketing efforts are not wasted on customers with lower CLV. Improved allocation enhances operational efficiency and reduces acquisition and retention costs.

Assumptions:
Total annual revenue: $1 million
Marketing budget: $100,000
Average spending per customer: $100
Number of customers: 10,000

Predicting Amount Spent Per Day	Predicting Low-Value Customers
1) **Enhanced Revenue Forecasting:** 5% reduction in unnecessary expenditures => $1 million * 5% = $50,000 annually 2) **Optimized Marketing Spend:** 15% higher conversion rate => $100,000* 15% = $15,000 annually 3) **Improved Customer Insights:** 10% increase in repeat purchases => 10,000 customers * 10% * $100 = $100,000 annually	1) **Resource Allocation:** Reduce customer acquisition costs by 20%, saving on marketing costs => $100,000* 20% = $20,000 annually 2) **Increased Customer Retention:** Increase retention rate by 15% => 10,000 customers * 15% * $100 = $150,000 annually
Total Savings: $65,000 annually **Total Additional Revenue: $100,000 annually**	**Total Savings: $20,000 annually** **Total Additional Revenue: $150,000 annually**