Elective Project:

Enhancing AllLife Bank's Customer Experience Through Cluster Algorithm Segmentation

Table of Contents

Executive Summary

In anticipation of the upcoming financial year, AllLife Bank aims to fortify its position within the credit card market. Acknowledging insights from its marketing research team, the bank recognizes an opportunity to enhance market penetration. In response to this, the marketing team advocates for a strategic approach, proposing the implementation of personalized campaigns to attract new customers and maximize upselling opportunities among the existing customer base.

Additionally, market research reveals a prevailing negative perception among customers regarding the bank's support services. Recognizing the crucial role customer satisfaction plays in retaining and expanding the customer base, the operations team is committed to addressing this challenge. To achieve this, they propose an overhaul of the service delivery model, with a primary focus on expediting query resolution to elevate the overall customer experience.

To effectively execute these initiatives, the heads of marketing and operations have jointly engaged the Data Science team. Leveraging data-driven insights and advanced analytics, the collaborative effort aims to optimize campaign targeting, refine customer engagement strategies, and streamline support services. The integration of data science into these key areas is anticipated to yield measurable improvements in market reach, customer satisfaction, and overall business performance.

This comprehensive approach, driven by data-driven decision-making, positions AllLife Bank to not only strengthen its existing customer relationships but also to successfully capture a larger share of the credit card market. The collaborative efforts across marketing, operations, and data science underscore a commitment to innovation and customer-centricity, setting the stage for a transformative and successful financial year ahead.

Business Problem Overview and Solution

AllLife Bank faces a critical challenge in maximizing the potential of its credit card customer base. The existing market penetration, as identified by the marketing research team, falls short of its optimum level. Moreover, customer perceptions of the bank's support services are subpar, necessitating a strategic intervention to elevate customer satisfaction and retention.

To address these challenges, AllLife Bank is embarking on a data-driven initiative focused on refining customer segmentation. The objective is to identify distinct customer segments within the credit card base, considering individual spending patterns and historical interactions with the bank. This strategic move aims to tailor marketing efforts more precisely, ultimately boosting customer acquisition and upselling opportunities.

Approach

The proposed solution involves leveraging a sophisticated machine learning model developed in Python. This model will analyze vast datasets, incorporating transactional data, spending behaviors, and past interactions. By employing advanced algorithms, the model will autonomously identify meaningful customer segments, allowing for a nuanced understanding of diverse customer profiles.

Benefits:

Enhanced Marketing Precision: By identifying specific customer segments, the bank can craft personalized campaigns that resonate with the unique preferences and behaviors of each group. This targeted approach is expected to significantly improve the effectiveness of marketing initiatives.

Improved Upselling Opportunities: Understanding the nuanced needs of different customer segments enables the bank to tailor upselling strategies. This can lead to increased credit card utilization, contributing positively to the bank's revenue stream.

Optimized Service Delivery: Insights from the machine learning model will also inform improvements in the service delivery model. By anticipating customer needs based on their historical data, the bank can streamline support services, ensuring faster query resolution and consequently enhancing overall customer satisfaction.

Technology Integration: The collaboration with the Data Science team involves deploying a Python-based machine learning model. This model will not only uncover hidden patterns within the data but also adapt and evolve over time, ensuring ongoing optimization of customer segmentation strategies.

Expected Outcomes: This initiative positions AllLife Bank to not only overcome its current challenges but also to proactively cater to the dynamic expectations of its credit card customers. The integration of machine learning into customer segmentation is anticipated to result in a more agile, responsive, and customer-centric approach, laying the foundation for sustained growth and success in the competitive credit card market.

Data Overview

The data utilized for the Customer Segmentation Project at AllLife Bank comprises comprehensive information on individual customers, focusing on key variables that provide insights into their financial behavior and interaction patterns with the bank. The dataset encompasses the following attributes:

1. **Sl_no (Customer Serial Number)**: A unique serial number assigned to each customer, serving as a primary identifier for individual records.
2. **Customer Key (Customer Identification)**: A unique identifier associated with each customer, facilitating accurate tracking and reference throughout the analysis.
3. **Avg_Credit_Limit (Average Credit Limit)**: This attribute represents the average credit limit extended to each customer. While the currency is not specified, assumptions can be made for contextual interpretation.
4. **Total_Credit_Cards (Total Number of Credit Cards)**: Indicates the overall count of credit cards held by a customer, offering insights into their credit portfolio.
5. **Total_visits_bank (Total Bank Visits)**: Quantifies the number of times a customer has physically visited the bank for queries or transactions, providing a measure of in-person engagement.

6. **Total_visits_online (Total Online Visits)**: Reflects the frequency of customer interactions through online channels, offering a glimpse into the digital engagement preferences of the customer base.
7. **Total_calls_made (Total Calls Made)**: Captures the total number of calls made by customers to the bank's call center, highlighting their reliance on telephonic communication for query resolution.

The richness of this dataset enables a holistic analysis of customer behavior, encompassing both traditional and digital touchpoints. The inclusion of credit-related metrics, alongside interaction frequencies, forms the foundation for the machine learning model. This data-driven approach aims to unlock meaningful patterns, facilitating the identification of distinct customer segments crucial for the targeted marketing and service delivery enhancements outlined in the project objectives.

In addition to the substantive attributes of the dataset outlined above, it is noteworthy to highlight key structural characteristics. The dataset comprises 660 observations, each corresponding to a unique customer, distributed across 7 columns. A crucial aspect contributing to the dataset's reliability is the absence of missing values; all 660 entries in each column contain non-null values. There are duplicated values, however, that will need to be investigated further and treated.

The homogeneity of data types across all columns, specifically the use of integers, ensures consistency and facilitates streamlined analytical processes. These structural features reinforce the dataset's integrity, setting a robust foundation for the ensuing data analysis and machine learning endeavors in the Customer Segmentation Project.

EDA and Data Preprocessing

We made a few observations of the summary statistics based on each variable:

1. **Avg_Credit_Limit:** The average credit limit varies significantly, ranging from 3,000 to 200,000. The mean credit limit of 34,574 indicates a moderately distributed dataset. However, the high standard deviation of 37,625 underscores substantial dispersion, suggesting the presence of outliers.

2. **Total_Credit_Cards:** The total number of credit cards per customer ranges from 1 to 10, with an average of 4.71. The distribution is slightly right-skewed, with a median of 5, indicating a relatively balanced distribution.
3. **Total_visits_bank:** Customer visits to the bank vary, ranging from 0 to 5, with an average of 2.40. The distribution is moderately right-skewed, reflecting a tendency for customers to make fewer physical visits to the bank.
4. **Total_visits_online:** Online interactions exhibit a wider range, from 0 to 15, with an average of 2.61. The distribution appears right-skewed, indicating that while most customers have a moderate number of online visits, a subset engages more frequently.
5. **Total_calls_made:** The total number of calls made by customers ranges from 0 to 10, with an average of 3.58. The distribution appears right-skewed, suggesting that a significant portion of customers makes a moderate number of calls, while a smaller subset engages more extensively.

These summary statistics provide a comprehensive overview of the dataset, highlighting key trends, central tendencies, and variations in customer attributes. Subsequent exploratory data analysis will delve deeper into these patterns, informing strategic decisions for customer segmentation and targeted interventions.

Univariate Analysis

1. **Avg_Credit_Limit**: The distribution of average credit limits is positively skewed (skew = 2.2), indicating a tail towards higher values. Substantial outliers, exceeding $100,000, are observed, suggesting a minority of customers with exceptionally high credit limits. The majority of customers, however, fall within the range of $0 to $25,000, with even upper quartile credit limits remaining below $50,000.
2. **Total_Credit_Cards**: Total Credit Cards per customer exhibit a relatively balanced distribution, with an average of 5 on a scale of 0 to 10. The absence of significant skewness or substantive outliers suggests a consistent and moderate credit card ownership pattern across the customer base.
3. **Total_visits_bank**: The distribution of customer visits to the bank is well-distributed, with a mean between 2 and 3 visits. This balanced distribution

indicates a relatively consistent engagement pattern, with customers making a moderate number of in-person visits.

4. **Total_visits_online**: Total online visits display positive skewness (skew = 2.23). While the mean is 2 visits, suggesting most customers engage online moderately, the presence of a wide range exceeding 14 indicates a subset of customers with considerably higher online engagement. This suggests a notable disparity in online interactions among the customer base.
5. **Total_calls_made**: The distribution of total calls made by customers indicates a low-touch pattern, with a mean of 3 calls on a scale of 0 to 10. Importantly, there are no outliers, affirming that the majority of customers engage in a fairly restrained number of telephonic interactions with the bank.

These observations provide valuable insights into the variation and tendencies within each variable, guiding further analysis and strategic decisions in the context of customer segmentation and service enhancement initiatives.

Bivariate Analysis

Now we explore the relationship between relevant variables through the use of a heatmap.

1. **Avg_Credit_Limit and Total_Credit_Cards**: There is a positive correlation between Avg_Credit_Limit and Total_Credit_Cards. This suggests that customers with a higher average credit limit tend to possess more credit cards. This correlation aligns with common financial practices where individuals with higher creditworthiness may be offered a greater number of credit cards.
2. **Avg_Credit_Limit and Total_visits_online**: A positive correlation is observed between Avg_Credit_Limit and Total_visits_online. This indicates that customers with higher credit limits are more inclined to engage with the bank's online services. This correlation is logical, as customers with greater financial capacity might leverage online platforms for convenience and flexibility.
3. **Avg_Credit_Limit and Total_calls_made**: A negative correlation is identified between Avg_Credit_Limit and Total_calls_made. This suggests that as the average credit limit increases, customers tend to make fewer calls to the bank. This inverse relationship could be attributed to a higher level of financial

self-sufficiency among customers with elevated credit limits, reducing the need for telephonic assistance.

4. **Avg_Credit_Limit and Total_visits_bank**: There is a negative correlation between Avg_Credit_Limit and Total_visits_bank. This implies that customers with higher credit limits are less likely to make in-person visits to the bank. Again, this aligns with the idea that customers with greater financial capacity may prefer more convenient and efficient online interactions.
5. **Total_visits_bank and Total_visits_online**: A negative correlation is evident between Total_visits_bank and Total_visits_online, suggesting that the majority of customers prefer using either in-person visits or online channels but not both. This underscores a channel preference among customers, indicating distinct segments with varying interaction preferences.
6. **Total_visits_bank and Total_calls_made, Total_visits_online and Total_calls_made**: Negative correlations between Total_visits_bank and Total_calls_made, as well as Total_visits_online and Total_calls_made, imply that customers tend to use a single communication channel predominantly. Those who visit the bank or engage online are less likely to make many calls, indicating a streamlined and channel-focused approach among customers.

These correlation insights provide a nuanced understanding of customer behaviors and preferences, offering valuable inputs for targeted marketing strategies and service delivery improvements.

Model Building

Now we will explore the findings from the cluster algorithms used to determine how best to approach the problem. The models built and applied for comparative analysis were KMeans, Gaussian Mixture, and KMedoid.

Overview of Models

In the pursuit of understanding and addressing the challenges posed by the credit card customer base, three distinct cluster algorithms were employed for comparative analysis: KMeans, Gaussian Mixture, and KMedoid. Each model type contributes unique strengths to the clustering process, offering different perspectives on how best to approach the segmentation problem.

1. **KMeans**: KMeans is a partition-based clustering algorithm that assigns data points to clusters based on their proximity to the cluster centroids.
    a. Strength: It is computationally efficient and particularly effective when clusters have a spherical shape. KMeans is widely used for its simplicity and scalability, making it suitable for large datasets.
2. **Gaussian Mixture**: Gaussian Mixture Model (GMM) is a probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions.
    a. Strength: GMM accommodates clusters with different shapes and sizes, providing flexibility in capturing complex patterns within the data. It also assigns probabilities to data points belonging to each cluster.
3. **KMedoid**: KMedoid, in contrast to KMeans, uses actual data points as cluster representatives (medoids) instead of centroids. It minimizes the sum of dissimilarities between data points and their assigned medoids.
    a. Strength: Robust to outliers and less sensitive to the initial choice of cluster centers, KMedoid can be advantageous in scenarios where clusters are irregularly shaped or when dealing with noisy data.

Note: Standard Scaler was applied to standardize the numerical features of the dataset, ensuring that each variable contributes equally to the clustering process. This is crucial when variables are measured on different scales, preventing certain features from dominating the clustering algorithm due to their larger magnitudes.

Additionally, PCA was employed to reduce dimensionality while retaining the essential variance in the data. By transforming the original features into a new set of uncorrelated variables (principal components), PCA streamlines the clustering process and mitigates the curse of dimensionality.

Insights from Models

After running each model, we derived the following insights:

1. **KMeans**:
    a. **Group_0**: This segment exhibits a lower average credit limit (Mean: $12,831) with a moderate number of credit cards and a high frequency of calls made.
    b. **Group_1**: Customers in this segment demonstrate a moderate average credit limit (Mean: $33,508), a higher number of credit cards, and a substantial presence in online visits.
    c. **Group_2**: This segment comprises customers with a notably higher average credit limit (Mean: $141,040), a significant number of credit cards, and a preference for online visits.
2. **GMM**:
    a. **Group_0**: Similar to KMeans, this segment has a lower average credit limit (Mean: $12,197), fewer credit cards, and a relatively higher frequency of calls made.
    b. **Group_1**: Customers in this segment exhibit a moderate average credit limit (Mean: $33,508) and are characterized by a higher number of credit cards and a substantial online presence.
    c. **Group_2**: This segment showcases customers with a notably higher average credit limit (Mean: $136,452), a significant number of credit cards, and a preference for online visits.
3. **KMedoid**:
    a. **Group_0**: This segment stands out with a notably higher average credit limit (Mean: $70,300), a substantial number of credit cards, and a balanced distribution of visits across channels.
    b. **Group_1**: Customers in this segment have a moderate average credit limit (Mean: $16,828), a moderate number of credit cards, and a preference for in-person visits to the bank.
    c. **Group_2**: This segment is characterized by a lower average credit limit (Mean: $13,568), fewer credit cards, and a higher frequency of calls made.

Conclusion

Based on what we learned from the analysis step, the primary lesson learned here was that the diverse use cases for this data enable us to leverage each model for their respective strengths rather than rely on any single model to achieve our business objectives. Here's a breakdown of how to use each model, including some actionable steps recommended for AllLife Bank:

1. **Utilize KMeans and GMM for Comprehensive Understanding**: Leverage the insights from both KMeans and GMM to gain a comprehensive understanding of customer segmentation.
    a. KMeans and GMM exhibit consistent patterns, emphasizing the distinctions in credit limit, credit card ownership, and interaction preferences. This dual perspective enhances confidence in identified segments.
2. **Incorporate KMedoid Insights for Added Nuance**: Consider the unique insights from KMedoid as an additional layer of understanding.
    a. KMedoid introduces a distinctive perspective, highlighting a segment with notably higher credit limits and balanced interaction patterns. Integrating this viewpoint enhances the granularity of segmentation insights.
3. **Validate Results with Domain Expertise**: Collaborate with domain experts and key stakeholders to validate the interpretation of identified segments.
    a. Domain experts can provide valuable context, ensuring that the segmentation aligns with the bank's business goals, customer behavior, and market dynamics.
4. **Implement Targeted Strategies for Each Segment**: Develop tailored marketing and service enhancement strategies for each identified segment.
    a. Armed with insights from multiple models, AllLife Bank can implement targeted interventions that resonate with the unique characteristics and preferences of each customer segment.
5. **Monitor and Iterate Based on Feedback**: Establish a system for ongoing monitoring and feedback collection.
    a. Regularly assess the effectiveness of implemented strategies and gather customer feedback. Iterate the segmentation approach if necessary,

ensuring continued alignment with evolving customer behaviors and market trends.

6. **Consider Hybrid Approaches for Robustness**: Explore hybrid clustering approaches that combine the strengths of KMeans, GMM, and KMedoid.
   a. A hybrid approach may provide enhanced robustness, mitigating the limitations of individual models and offering a more resilient segmentation strategy.

By combining insights from KMeans, GMM, and KMedoid, and adopting a collaborative and iterative approach, AllLife Bank can derive actionable intelligence for targeted decision-making, fostering customer-centric strategies that drive satisfaction, engagement, and sustainable business growth.