Ca' Foscari
University
of Venice

Master's Degree programme

in Data Analytics for Business and Society

Final Thesis

Customer Segmentation and the Determinants of Relationship Tenure in a Vietnamese Commercial Bank

A Machine-Learning and Econometric Analysis of Large-Scale Banking Data

Supervisor

Ch. Prof. Agar Brugiavini

Graduand

Van Duc Dao

Matriculation Number 908607

Academic Year

2024 / 2026

Table of Contents

Abstract

Customer segmentation is the process of dividing an organization's customers into groups (clusters) that share similar behaviours and characteristics. Machine learning enables this segmentation to be performed rapidly on very large customer bases, even when the scale reaches millions of individuals. In customer-centric industries such as telecommunications, insurance, and banking, machine learning is widely applied to address segmentation and customer-retention challenges. Segmentation helps organizations understand how to interact appropriately with each customer group, develop targeted product-marketing strategies, meet specific customer needs, and reduce churn rates. This topic is chosen because customer retention is a direct driver of bank profitability - retained, long-tenured customers are markedly more profitable than newly acquired ones (Reichheld & Sasser, 1990; Reinartz & Kumar, 2003) - yet most Vietnamese commercial banks hold large volumes of customer data without the analytical infrastructure to convert it into actionable segments (FiinGroup, 2024).

In this study, the unsupervised clustering algorithm K-Means - in its scalable MiniBatch variant - is selected as the primary method for customer segmentation. A Python program is developed and trained on a large-scale, real-world dataset comprising 631,062 customers described by 8 financial-behavioural variables, after the data has been transformed using log1p and standardized with RobustScaler. The objective is to create customer groups based on behavioural characteristics, enabling the bank to understand its customer structure and thereby design new products or adjust existing offerings tailored to each specific segment.

Cluster quality here rests on three internal validation metrics at once - the Silhouette Score, the Davies-Bouldin Index, and the Calinski-Harabasz Index - cross-checked against a stability test that re-runs the model under different initialization seeds. What comes out is eight customer segments from MiniBatch K-Means, each with its own distinct profile, and the stability is about as high as it gets: a mean Adjusted Rand Index of 0.9999. To push the segmentation out to the whole customer base and to new arrivals, those cluster labels then become the training target for a Random Forest classifier. That serves two purposes at once - it tests whether the segments are actually separable, and it leaves behind a model that can slot new customers into a segment in production. The Random Forest recovered the segmentation at 99.85% test accuracy with a macro-averaged F1 of 0.996, and when its feature-importance ranking was read off, four variables stood out as decisive: savings balance, outstanding loan, collateral value, and credit-card spending.

Building on the segmentation, the thesis then turns from description to explanation through an econometric analysis of customer tenure. Using ordinary least squares regression on the full population, the length of the customer–bank relationship is modelled as a function of demographic attributes, segment membership, and financial behaviour. The demographic block alone explains 72.7% of the variation in tenure, and

occupation emerges as the strongest determinant, with commerce-and-services customers exhibiting markedly longer relationships. The premium segment is found to be the most durable, linking the behavioural segmentation directly to a concrete economic outcome. The analysis explicitly addresses reverse causality and interprets the financial-behaviour coefficients as conditional correlations rather than causal effects.

Acknowledgements

My first thanks go to my supervisor, Agar Brugiavini whose guidance, patience, and feedback shaped this thesis at every turn.

I am also indebted to Eximbank (Vietnam Export Import Commercial Joint Stock Bank) for access to the anonymised customer data this study is built on. The bank made the data available for academic research only, under its data-confidentiality terms; it is not public and cannot be shared or reproduced. Without that access, none of the empirical work here would have been possible.

Last, I want to thank my family and friends, who kept supporting and encouraging me right through my studies.

Chapter 1. Introduction

Vietnamese banks have digitised fast over the past few years, and yet a real gap persists - between the data sitting in their systems and the customer insight they actually pull out of it. The transactional records have piled up: account balances, payment histories, loan records, savings deposits, credit-card usage. Most of it goes unused. When it comes to grouping customers, a lot of banks still fall back on crude rules - age band, income bracket, account type - the kind of demographic shorthand that misses how customers actually behave. Segments built that way come out broad and blurry, and they rarely tell anyone in the business much worth acting on.

The fundamental challenge is not the absence of data, but rather the inability to transform raw banking data into meaningful customer portraits. Vietnamese banks typically store vast quantities of structured financial data across multiple systems yet lack the analytical frameworks to synthesize this information into a coherent, holistic view of each customer segment. Consequently, critical questions remain unanswered: What distinct behavioural groups exist within the customer base? What is the financial profile of each group? Which customers represent the highest value, and which are at risk of disengagement? Without answers, banks cannot design targeted products, allocate marketing resources efficiently, or develop differentiated retention strategies.

This limitation becomes particularly consequential amid intensifying competition within the Vietnamese banking industry. With the emergence of digital banks, fintech companies, and increasingly sophisticated foreign competitors, traditional banks can no longer afford to treat their customer base as a homogeneous mass. The one-size-fits-all approach to product design and marketing - once standard practice - is rapidly becoming obsolete. Banks that fail to understand the granular composition of their portfolio risk losing valuable segments to competitors who offer more personalized services.

This is where machine learning, and unsupervised clustering in particular, earns its place. Techniques like K-Means, Gaussian Mixture Models, DBSCAN, and Hierarchical Clustering can surface natural groupings inside large, high-dimensional data with no labels supplied up front. They look at many behavioural signals together -

how often someone transacts, what their balances look like, how they use credit, which products they hold, how they save - and they pull out clusters of customers who behave alike financially. The result is a data-driven picture of the entire customer base rather than a handful of pre-decided buckets.

Running clustering on banking data beats conventional segmentation on several fronts. It scales: MiniBatch K-Means chews through datasets of hundreds of thousands or millions of customers in minutes. It is objective, in that the groupings come from patterns in the data and not from someone's judgment call, which cuts bias and makes the whole thing reproducible. It is multidimensional, weighing every behavioural axis at once rather than one at a time, so the customer portraits come out richer. And it is actionable - once you profile the segments, the strategy almost writes itself: cross-sell into the under-served ones, build retention around the high-value ones, watch risk in the credit-heavy ones, and run activation campaigns at the dormant ones.

The payoff does not stop at one use case either. Those clustering-derived portraits work as a base layer that several other things can be built on top of: churn models fitted segment by segment, recommendation engines tuned to what each segment tends to want, credit-risk scoring that folds segment membership in as one more predictor. Seen that way, machine-learning segmentation is not just an analytical exercise - it is the infrastructure that lets a bank move from organising itself around products to organising itself around customers..

1.1. Motivation

In banking CRM, customer identification is the bedrock - it covers both segmenting customers and targeting them. Get customers identified and sorted, and the marketing strategy has something to stand on. Classification and clustering go after the specific groups that match a business objective, while regression helps forecast how prospective customers are likely to behave. The whole toolkit of data mining - classification, regression, clustering, association - gets brought to bear on building a rounded picture of customer loyalty and behaviour (Niyagas, Srivihok & Kitisin, 2006; Wedel & Kamakura, 2000).

Big data has handed banking two problems in particular. One is sheer handling - coping with the volume and complexity of the data cheaply and quickly, as it pours in faster, bigger, and across more channels than before. The other is conversion - turning all that analysis into something with business value and a competitive edge, when the data that matters is often costly to get, missing in digital form, or stuck as unstructured text. Pairing problem-solving methods with data-driven ones is what lifts decision quality and actually delivers something to a financial institution (FiinGroup, 2024; Statista, 2025). And the money on the table is not small: keeping a customer is far more profitable than winning a new one, and shaving even a little off the defection rate moves profit noticeably (Reichheld & Sasser, 1990; Gupta, Lehmann & Stuart, 2004) - which is exactly why good segmentation and retention pull directly on a bank's bottom line

(Rust, Lemon & Zeithaml, 2004).into tangible business value and competitive advantage, since critical business data often remains expensive to obtain, unavailable in digital form, or stored as unstructured information. Combining problem-solving techniques with data-driven approaches is essential to improve decision quality and deliver real value to financial institutions (FiinGroup, 2024; Statista, 2025). The economic stakes are considerable: retained customers are substantially more profitable than newly acquired ones, and even a small reduction in defection rates can raise profits markedly (Reichheld & Sasser, 1990; Gupta, Lehmann & Stuart, 2004), which makes effective segmentation and retention a direct lever on bank profitability (Rust, Lemon & Zeithaml, 2004).

This challenge is especially acute in the Vietnamese banking context. Despite sitting on vast reserves of customer data, most Vietnamese commercial banks have yet to develop the analytical infrastructure needed to transform this raw data into actionable customer portraits. The gap between data availability and customer understanding represents both a significant limitation and a compelling opportunity. The rapid digitalisation of Vietnamese banking - including the recent roll-out of biometric customer verification and the strong official push towards digital payments (Vietnam.vn, 2026) - is generating customer data at an unprecedented scale, sharpening both the need and the opportunity for systematic segmentation (FiinGroup, 2024).

The motivation of this study is therefore threefold. First, it applies unsupervised clustering methodology to a large-scale, real-world banking dataset comprising 631,062 customers and 8 financial-behavioural variables from a Vietnamese commercial bank. The K-Means algorithm is selected as the primary clustering method due to its computational efficiency on large datasets, its well-established theoretical foundation, and its suitability for data with approximately linear cluster boundaries - a property confirmed through dimensionality-reduction analysis in this study. Second, the study establishes a complete, reproducible analytical pipeline - from exploratory data analysis and preprocessing through clustering and evaluation - that demonstrates how Vietnamese banks can systematically convert their existing data assets into meaningful customer segments. Segmentation quality is rigorously assessed using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index, complemented by stability testing across multiple random initializations. Third, the study moves beyond description to explanation by econometrically analysing the determinants of customer tenure, identifying which customer characteristics predict a longer and therefore more valuable banking relationship, and thereby connecting the behavioural segments to a concrete economic outcome.

Chapter 2. Literature Review

Every organisation in a competitive market wants the same things - to grow, to innovate, to put out tailored products that set it apart. Read customer behaviour and buying patterns well, and you can lift profitability by aiming offers more precisely. Customer segmentation - sorting customers into groups that behave similarly - has become one of the go-to ways of doing that.

The whole point of building customer clusters is to expose behavioural patterns that would otherwise stay buried inside a large, undifferentiated dataset.for achieving this goal. The fundamental objective of creating customer clusters is to study behavioural patterns that would otherwise remain hidden within large, undifferentiated datasets.

Segmentation matters everywhere, but it matters more in banking, where competition is fierce and no institution holds onto success for long by selling the same product to everyone. Segment on financial behaviour instead of plain demographics, and a bank can design products that fit, spend its marketing budget where it counts, and build a different retention play for each group.

Research has consistently demonstrated that behavioural segmentation produces more actionable insights than demographic segmentation alone (Wedel & Kamakura, 2000). While demographic variables such as age, income, and profession are relatively easy to measure, behavioural variables - transaction frequency, product usage, savings behaviour - capture the complexity of customer relationships far more accurately.

2.1. Machine Learning Approaches to Customer Segmentation

Machine learning, and unsupervised clustering above all, has changed how segmentation gets done - it finds the natural groupings inside large, messy datasets on its own. Where the old methods needed an analyst to draw the categories first, clustering reads the segments straight off the data.

A handful of studies have put clustering to work on banking data at different scales. Gupta (2023), a master's thesis on machine-learning segmentation of bank customers, cleaned a customer dataset by stripping nulls and dealing with outliers, fixed the cluster count with the Elbow method and Silhouette Score (landing on K = 7), and pitted five algorithms against each other - K-Means, Agglomerative Clustering, Spectral Clustering, Gaussian Mixture Model, DBSCAN - with K-Means coming out on top across every metric. That set a useful precedent: of the unsupervised options on the table, K-Means keeps performing well on banking segmentation.

A study on Thai banks (Niyagas, Srivihok & Kitisin, 2006) investigated segmentation based on e-banking usage, applying K-Means, Self-Organizing Maps, and RFM analysis, and generated eight clusters. Notably, one cluster contained more than half of all customers, while three clusters collectively accounted for less than three percent of the base. This pattern of highly imbalanced cluster sizes - a dominant group of basic customers alongside several small but distinctive segments - is a recurring characteristic in retail-banking segmentation that reflects the inherent structure of consumer financial behaviour. RFM analysis further revealed that the largest cluster consisted mainly of inactive customers, while a very small cluster of only 0.24 percent proved most valuable due to intensive use of third-party transfer services, demonstrating that cluster size alone does not determine strategic importance.

Gupta (2023) further demonstrated the practical deployability of K-Means by building a web application for customer segmentation on a dataset of 8,068 bank customers, confirming that K-Means is not only effective in producing interpretable segments but also computationally efficient - a property particularly relevant for institutions with large customer bases, where computational cost is a significant consideration.

2.2. Large-Scale Segmentation and Performance Evaluation

As customer bases grow into the hundreds of thousands or millions, the scalability of the segmentation approach becomes a critical concern. Standard K-Means must traverse the entire dataset at every iteration, which becomes prohibitively expensive at this scale; the MiniBatch variant addresses this by updating centroids from small random batches of data, making web-scale clustering feasible (Sculley, 2010). This scalability is one reason why K-Means and its variants remain the most widely applied family of segmentation algorithms even on very large datasets, as the systematic review discussed below confirms (Firdaus & Utama, 2020). A further advantage of a data-driven approach is its ability to reveal customer groups that assumption-based methods overlook - for example, low-activity or low-income segments for which no predefined persona exists - illustrating a key value of unsupervised segmentation.

Validating clustering results means measuring their quality, and three internal metrics dominate the practice: the Silhouette Score (Rousseeuw, 1987), the Davies-Bouldin Index (Davies & Bouldin, 1979), and the Calinski-Harabasz Index (Caliński & Harabasz, 1974). Silhouette captures cohesion and separation point by point, running from -1 to +1, with higher meaning cleaner clusters. Davies-Bouldin averages the ratio of within-cluster to between-cluster distance, where lower is better. Calinski-Harabasz takes between-cluster dispersion over within-cluster dispersion, where higher signals tighter, better-separated clusters. Each looks at the problem from a slightly different angle, and reading them together gives a sturdier verdict than leaning on any one alone.

A systematic literature review of the most important attributes for banking customer segmentation (Firdaus & Utama, 2020) narrowed an initial pool of papers to a final set of ten, finding that customer balance was among the most frequently used attributes, and that K-Means was the most commonly employed algorithm (appearing in six of ten studies), followed by RFM (four studies) and C-Means (two studies). This systematic evidence strongly supports the selection of K-Means as the primary algorithm.

2.3. Random Forest for Cluster Validation and Prediction

Beyond the clustering process itself, the ability to assign new customers to established segments is critical for operational deployment. Random Forest, an ensemble classification algorithm based on multiple decision trees, is widely recognized as one of the most accurate and robust general-purpose classifiers (Breiman, 2001). In segmentation, it serves two purposes: validating clustering quality by measuring how well a supervised classifier can distinguish the identified segments and building a prediction model that classifies future customers into existing segments without re-running the clustering algorithm. Its ensemble construction, which averages many de-correlated trees, makes it well suited to recovering the segment boundaries produced by clustering, which is why it is a common choice for deployment-ready segment-assignment models.

2.4. Economic Motivation: Why Customer Segmentation and Retention Matter

Although customer segmentation is often presented as a technical, data-driven exercise, its underlying motivation is fundamentally economic. The financial rationale rests on the recognition that customers differ markedly in the value they generate for a bank, and that the

relationship with a customer is an asset whose profitability evolves over time. Understanding this rationale is essential for framing why an accurate segmentation of more than six hundred thousand customers is a question worth answering for a commercial bank.

A first strand of the literature establishes that customer relationships should be treated as financial assets. The concept of customer lifetime value (CLV) - the present value of the future cash flows attributed to the relationship with a customer - reframes marketing expenditure as an investment in an asset rather than a cost (Gupta, Lehmann & Stuart, 2004). Building on this idea, Rust, Lemon & Zeithaml (2004) develop the notion of customer equity and show how marketing actions can be evaluated by their return on the customer asset base. Segmentation is the operational prerequisite for managing this asset: a bank that cannot distinguish high-value from low-value customers cannot allocate retention and cross-selling resources efficiently.

A second strand documents the profitability of customer retention and relationship duration. Reichheld & Teal (1996) argue that small improvements in retention can produce disproportionately large increases in profitability, because long-tenured customers tend to generate growing revenue, lower servicing costs, and referrals. In a banking and noncontractual setting, Reinartz & Kumar (2000, 2003) provide empirical evidence that relationship duration is a central determinant of customer profitability, while also cautioning that the link between longevity and profit is not mechanical. These findings establish retention and relationship length as economically meaningful outcomes, motivating the analysis of tenure determinants developed later in this thesis.

A third strand connects segmentation to differentiated strategy. Once customers are grouped by behaviour, the bank can match products, pricing, and service levels to the needs and value of each group, rather than offering an undifferentiated proposition to all customers. Within the customer-equity framework, such differentiated targeting is precisely the mechanism through which marketing resources are directed toward their highest-return uses (Rust, Lemon & Zeithaml, 2004). This provides the practical justification for the present study: the segmentation is not an end but a foundation for retention and cross-selling decisions that have direct financial consequences for the bank.

Prior work has addressed these questions largely on data from Western and other Asian markets. Whether the same economic regularities - the value of behavioural segmentation and the profitability of longer relationships - hold for a large Vietnamese retail bank is an open empirical question, and one that this study is positioned to address using the full active customer base of Eximbank.

2.5. Research Gap and Contribution

The literature review reveals several gaps that this study addresses. First, while many studies demonstrate the effectiveness of K-Means for banking segmentation, the majority were conducted on relatively small datasets ranging from a few thousand to tens of thousands of customers; studies applying clustering to datasets exceeding 600,000 customers remain rare. Second, most existing research was conducted in Western, Southeast Asian, or East Asian contexts, while studies on Vietnamese commercial banks using machine-learning methods are notably scarce. Third, many studies focus narrowly on the clustering algorithm without presenting a complete, end-to-end analytical pipeline that could serve as a reproducible methodology for practitioners.

This study addresses these gaps by applying MiniBatch K-Means to a large-scale dataset of 631,062 customers from a Vietnamese commercial bank, using 8 financial-behavioural variables that capture multiple dimensions of the customer-bank relationship. The complete pipeline - including log1p transformation for skewness reduction, RobustScaler standardization, Hopkins-statistic validation, multiple internal evaluation metrics, and stability testing - is documented in full to ensure reproducibility and methodological transparency.

This work parts ways with its nearest predecessor, Gupta (2023), in more than the shared reliance on K-Means and a CRISP-DM-style workflow - and that workflow is a de-facto field standard anyway, not anyone's signature. The scale is two orders of magnitude apart (631,062 customers against a few thousand), which is why the scalable MiniBatch variant becomes necessary in the first place. The setting is a Vietnamese commercial bank working off real system-of-record data. The segments get validated against held-out variables the algorithm never saw (Section 4.3.3). And, the part I weight most heavily, Chapter 5 carries the analysis further with an econometric model of customer tenure that ties the segments to a concrete economic outcome - something segmentation-only studies simply do not have. So the shared methodology is just the field's common ground; what this thesis adds sits in its scale, its context, its validation strategy, and that explanatory extension.

The literature review above results in these research questions, which will drive the analysis of this research. -research question 1 will answer the question, which method is best to segment customers into unique groups of customers with different financial behaviours? Research question 2 will determine which of the financial behavioural measures are the most relevant in forming the clusters of customers. Research Question 3 will ascertain whether the identified clusters are stable enough and statistically valid enough to generate insights to enable the development of business strategy. Research question 4 will investigate the associations between customer characteristics (demographic and financial customers), and the customer relationship length from the demographic customer characteristic perspective, and the extent to which they provide reasonable evidence such that this evidence is appropriately interpreted given that the dataset is cross-sectional.

Chapter 3. Methodology

3.1. Research Design and Overview

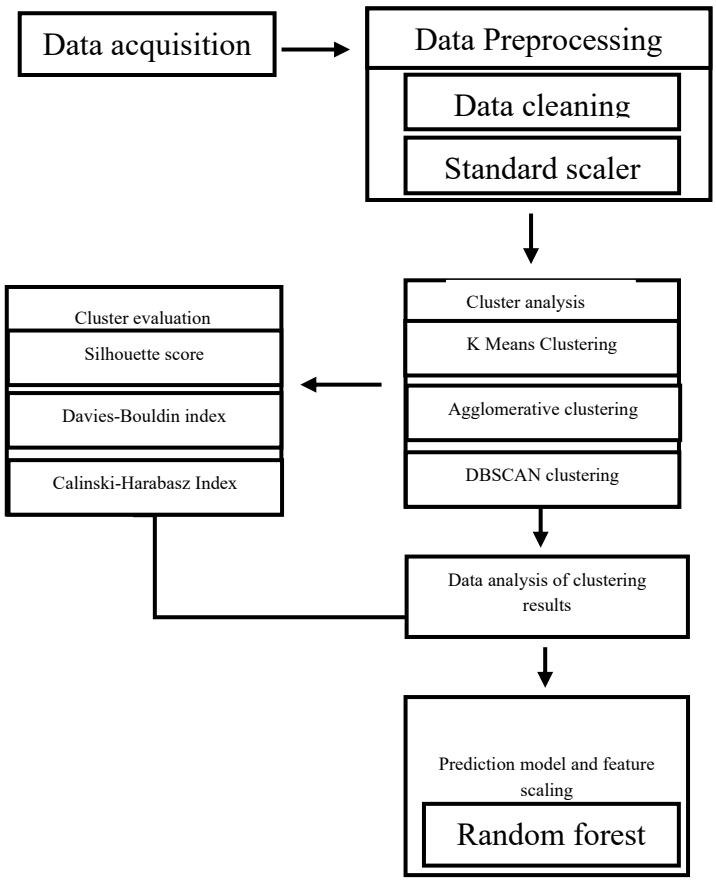
The design of this study is quantitative and data-driven, with unsupervised machine learning for segmentation at its core. The work moves through a seven-stage pipeline: (1) framing the problem and getting the data; (2) exploratory data analysis (EDA); (3) preprocessing; (4) checking whether the data even clusters; (5) settling the number of clusters and building the model; (6) evaluation; and (7) labelling the full population and reading the business meaning off the result. Structuring it this way keeps every methodological choice tied back to something the diagnostics actually showed - which is what makes the analysis both reproducible and transparent.

The analysis was implemented in Python using the scikit-learn library for the machine-learning algorithms, pandas and NumPy for data manipulation, and matplotlib and seaborn for visualization. The K-Means algorithm - specifically its scalable variant,

MiniBatch K-Means - was selected as the primary clustering method based on the evidence presented in the literature review and on the computational constraints discussed in Section 3.7.

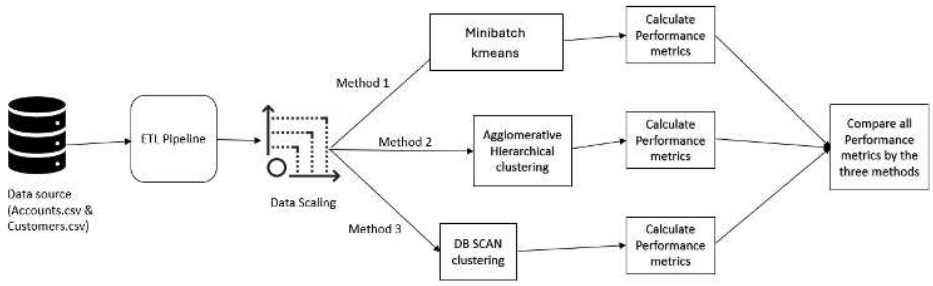
3.2. System Architecture

Start to finish, the workflow runs through a chain of linked stages: pull the raw data, clean and prepare it, scale the features, run the cluster analysis, evaluate what comes out, interpret it, and finally fit a prediction model that confirms both that the clustering holds up and that feature scaling did its job. feature scaling.

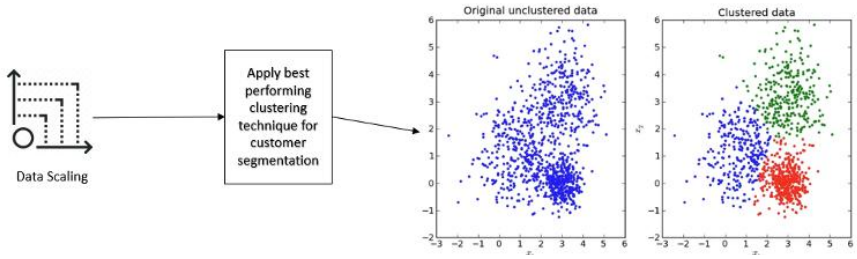


[Figure 3.1: Overall architecture of the segmentation pipeline] - the structure follows the standard CRISP-DM analytical workflow widely used in machine-learning applications to customer data.

A dedicated Python program brings the source tables together and passes them through an Extract, Transform, and Load (ETL) pipeline. After the data is loaded, it is transformed with log1p and scaled with RobustScaler. The numerical data is then taken into the cluster-analysis stage, where the goal is to determine which clustering method works best for customer segmentation. Once the candidate methods have been evaluated against the chosen performance metrics, the strongest method is selected to perform the final segmentation.

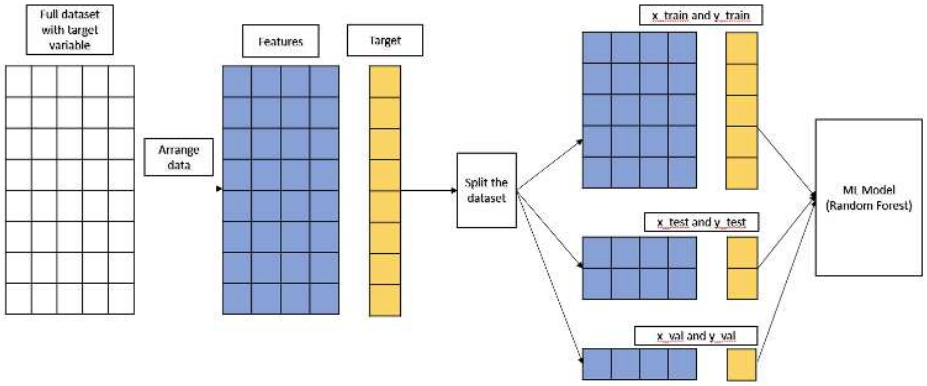


[Figure 3.2: Preprocessing and feature scaling]



[Figure 3.3: Applying the selected clustering technique to the scaled data]

A Random Forest is then trained to double-check what MiniBatch K-Means produced and to act as a predictor that can route new customers to the right segment. It does one more thing along the way - ranking the features by importance, which shows which variables drive the segmentation hardest and which barely register.



[Figure 3.4: Prediction model using Random Forest]

3.3. Theoretical Background

3.3.1. Supervised Machine Learning

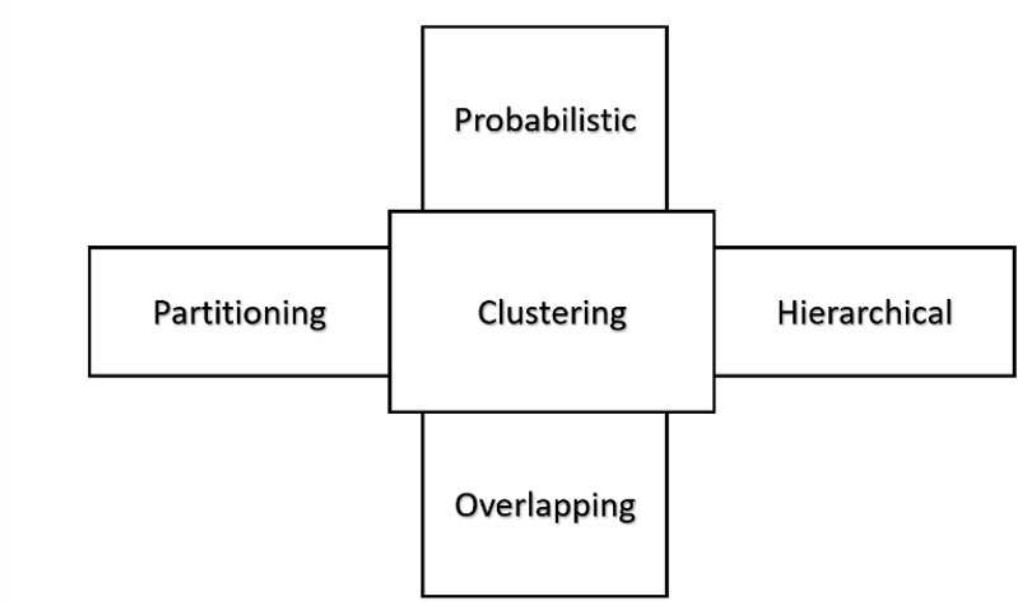
Supervised learning relies on prior knowledge of the data: a labelled dataset is needed to train the model and generate predictions. These methods tend to deliver strong predictive accuracy and are widely used in data analysis; their main limitation is that they cannot operate on unlabelled data. Supervised algorithms fall into two broad families. In classification, the labels are categorical, so only a finite set of outcomes is possible; common algorithms include Decision Trees, Random Forest, K-Nearest Neighbour, and Naive Bayes. In regression, the algorithms also take labelled data as input but return continuous numeric predictions; typical examples are Linear Regression and Logistic Regression.

3.3.2. Unsupervised Machine Learning

Over the past decade, unsupervised learning has become increasingly important, largely because modern applications generate enormous volumes of unlabelled data. Unsupervised methods surface hidden structure within a dataset and reduce its size, either by grouping records together or by reducing the number of features through dimensionality reduction. The methods most grouped under unsupervised learning are clustering, unsupervised feature learning, dimensionality reduction, and association-rule mining.

3.3.3. Clustering

Clustering is an unsupervised method: it splits data into groups by the similarities already present in it, and in practice the distance between two points in feature space is what stands in for how alike they are. There are many clustering algorithms, and the right one leans heavily on the shape of the data and the problem you are solving. No single method wins everywhere, so choosing one that fits the data is among the most consequential decisions in the whole analysis. The three unsupervised methods examined here are K-Means (its MiniBatch variant), Agglomerative Hierarchical Clustering, and DBSCAN.



[Figure 3.5: Types of clustering]

3.3.4. The Standard K-Means Algorithm

K-Means is about as widely used as unsupervised clustering algorithms get. Hand it n observations in d-dimensional space and a target of K clusters, and it splits the data so as to minimise the total squared distance from every point to its nearest cluster centre. That quantity goes by Within-Cluster Sum of Squares (WCSS), or inertia, and is defined as:

$$WCSS = \Sigma_{j=1}^{k} \Sigma \, (x_i \in C_j) \, \| \, x_i - \mu_j \, \|^2$$

where C_j is the j-th cluster, μ_j is the centroid of cluster j, and $\| \, x_i - \mu_j \, \|$ is the Euclidean distance from point x_i to the cluster centre. The algorithm finds a solution by alternating between two steps until convergence: an assignment step, in which each observation is assigned to the cluster whose centre is nearest; and an update step, in which each centroid is recomputed as the arithmetic mean of all points currently belonging to that cluster. The two steps repeat until the centroids no longer change appreciably or the maximum number of iterations is reached. Because the result depends on the initial centroid positions, the algorithm is typically run several times with different initializations, retaining the solution with the lowest WCSS.

3.3.5. The MiniBatch K-Means Variant

A limitation of standard K-Means is that, in the update step, the algorithm must traverse the entire dataset at every iteration. When the data is large - such as the 631,062 customers in this study - the computational and memory cost becomes very high. To address this, the study uses the MiniBatch K-Means variant.

The core difference is that, instead of using the entire dataset to update the centroids at each iteration, MiniBatch K-Means draws only a small random batch of b

observations (with b much smaller than n). At each iteration the algorithm: (i) draws a random batch of b points; (ii) assigns each point in the batch to its nearest centroid; and (iii) updates the centroids using a running average - each centre is shifted partway toward the newly assigned points, with the step size decreasing as the number of points a cluster has received grows. This running-average mechanism helps the centroids converge smoothly and stabilize over successive batches.

By processing only, a small batch at each step, MiniBatch K-Means substantially reduces training time and memory requirements compared with standard K-Means, while still yielding a solution of comparable quality. This is why the variant is particularly suited to large-scale data. In this study, the entire search for the optimal number of clusters over 631,062 customers took only about 3.3 minutes.

3.3.6. Agglomerative Hierarchical Clustering

Agglomerative Hierarchical Clustering is a bottom-up method that begins with each observation as its own cluster and then iteratively merges the two closest clusters until a single cluster remains, producing a dendrogram that records the full merge history. It does not require the number of clusters to be specified in advance and can capture nested structure. However, the method must compute and repeatedly update the pairwise distances between observations, which entails a memory cost on the order of n-squared and a time cost between O(n-squared) and O(n-cubed). For n = 631,062 customers, the pairwise-distance matrix alone would require on the order of hundreds of gigabytes of memory, which far exceeds the capacity of the hardware used in this study. Consequently, Agglomerative Hierarchical Clustering could not be executed on the full dataset.

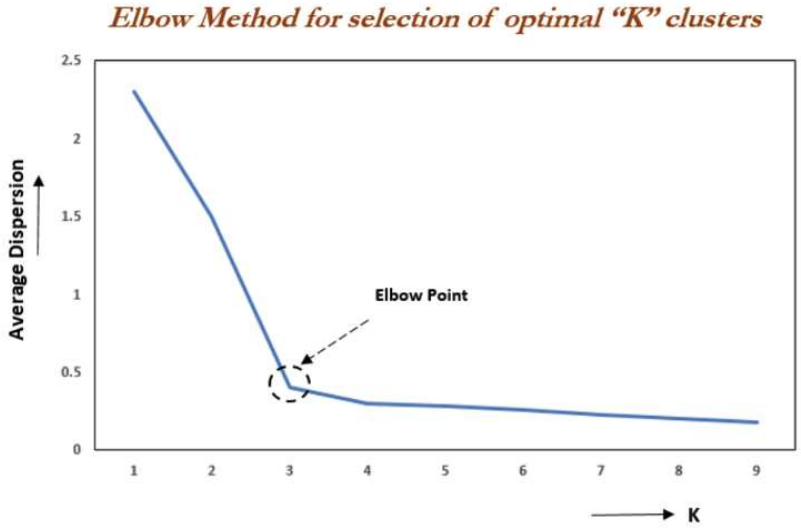
3.3.7. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) groups together points that are densely packed and labels points in low-density regions as noise. It can discover clusters of arbitrary shape and does not require the number of clusters to be specified. Its drawbacks, however, are significant at scale: performance is highly sensitive to the two parameters epsilon (neighbourhood radius) and minPts; in high-dimensional space the notion of density becomes unreliable; and a naive implementation has a worst-case time complexity of O(n-squared) together with a large memory footprint when neighbourhood queries are materialized. On the 631,062-row dataset, DBSCAN either exhausted available memory or failed to terminate within a practical time and therefore could not be applied to the full population either.

3.3.8. Elbow Method

The Elbow Method is used to find a sensible value of K for K-Means. It computes the within-cluster sum of squared errors between each point and its nearest centroid across a range of K values. As K grows, the error falls; the point at which the drop becomes markedly smaller is taken as the elbow, beyond which further partitioning is

not worthwhile. The procedure runs K-Means for a series of K values, computes the WCSS for each, plots K against WCSS, and identifies the sharp bend in the curve.



[Figure 3.6: Illustrative elbow method for selecting the optimal K]

3.3.9. Performance Metrics for Evaluation

Three internal metrics are used here to judge clustering quality.

Davies-Bouldin Index (Davies & Bouldin, 1979). For each cluster you compute its similarity to every other cluster and keep the largest such value as that cluster's score; average those per-cluster scores and you have the index. At heart it weighs within-cluster scatter against between-cluster separation. It starts at 0 and climbs, and smaller is better. For each cluster, the similarity to every other cluster is computed, and the largest of these values is taken as that cluster's score; averaging the per-cluster scores gives the index. In essence it is the ratio of within-cluster scatter to between-cluster separation. Its value runs from 0 upward, with smaller values indicating better clustering.

Silhouette Score (Rousseeuw, 1987). For a given point, the coefficient comes from its mean intra-cluster distance (a) and its mean distance to the nearest neighbouring cluster (b), as (b - a) / max(a, b). Near +1 the point sits squarely in the right cluster; near 0 it straddles two; near -1 it has probably ended up in the wrong one. Average over all points for an overall read - higher being better. For each point, the coefficient is computed from the mean intra-cluster distance (a) and the mean distance to the nearest neighbouring cluster (b), as (b - a) / max (a, b). A value near +1 means the point sits comfortably in the correct cluster, near 0 means it lies between clusters, and near -1 means it has likely been placed in the wrong cluster. The mean over all samples summarizes overall quality; higher is better.

Calinski-Harabasz Index (Caliński & Harabasz, 1974). Sometimes called the variance-ratio criterion, it pits between-cluster dispersion against within-cluster dispersion across all clusters. The higher it climbs, the better separated and more compact the clusters. Also called the variance-ratio criterion, this index is the ratio of between-cluster dispersion to within-cluster dispersion across all clusters. A higher value signals better separated and more compact clusters.

3.3.10. Adjusted Rand Index (ARI): Stability Assessment

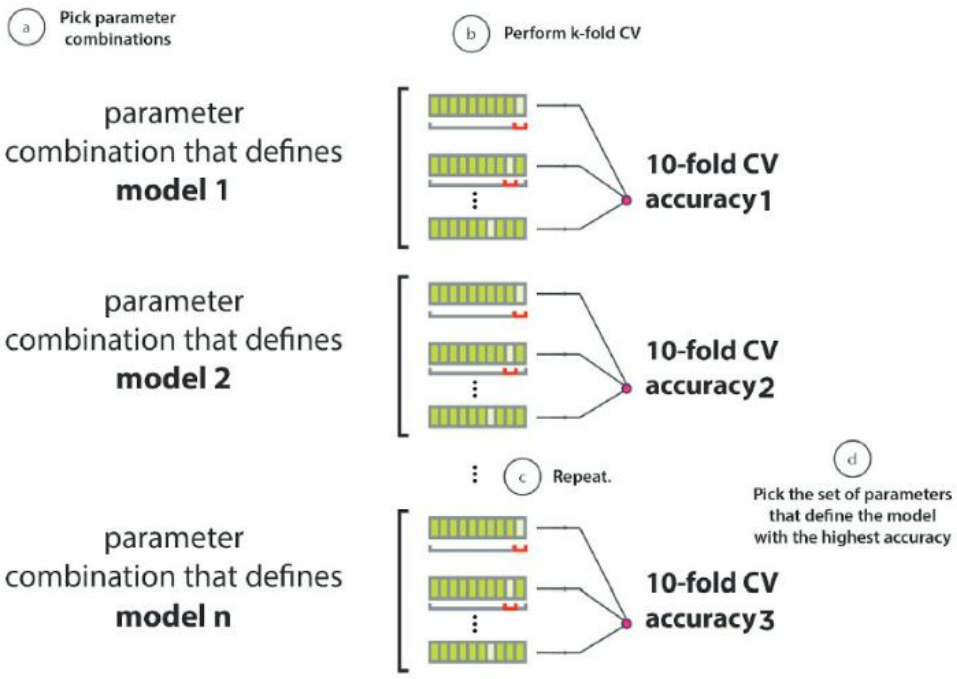
The Adjusted Rand Index gauges how alike two clusterings of the same data are, while discounting whatever agreement chance alone would produce. The plain Rand Index counts the share of observation pairs that two solutions classify the same way; ARI then corrects that for its expected value under random assignment:Rand Index counts the proportion of pairs of observations classified consistently across two solutions; ARI adjusts it by its expected value under random assignment:

$$ARI = [\ RI - E(RI)\ ] / [\ max(RI) - E(RI)\ ]$$

ARI = 1 means the two clusterings are identical; near 0 means agreement no better than chance; below 0 means worse than chance. As a working bar for stability, ARI > 0.85 will do. Here ARI does the stability checking - re-run K-Means under different random seeds and see whether customers stay put in the same groups. Its appeal is that it is chance-adjusted and symmetric; its catch is that it needs a reference clustering to compare against.

3.3.11. Random Forest Algorithm

Random Forest is a supervised ensemble method applicable to classification and regression. It extends the decision-tree algorithm: each tree is grown from a bootstrap sample of the training data, and at each split a random subset of candidate attributes is considered. The forest's prediction is obtained by majority vote across all trees. This combination of bagging (training each tree on a random sample drawn with replacement) and feature randomness introduces diversity, so that the trees are effectively de-correlated and the ensemble generalizes well. In this study, Random Forest is used both to validate the separability of the clusters and to build a deployable model that assigns new customers to existing segments, while its feature-importance scores reveal which variables drive the segmentation.



[Figure 3.7: Working of the Random Forest model]

3.3.12. Grid Search for Hyperparameter Tuning

Selecting the best set of parameters when training a model is known as hyperparameter optimization. When several hyperparameters exist, finding the best combination becomes a search problem. This study uses a grid search over the candidate parameter values - over the number of clusters K - evaluating each configuration against the internal clustering metrics and selecting the configuration that yields the strongest results.

3.3.13. Hypotheses

Building on the literature in Chapter 2 and the methodology laid out in this chapter, the following hypotheses are put forward.

- For clustering the numerical data, MiniBatch K-Means will be the only one of the three candidates that can actually run on the full 631,062-row dataset, and it will yield stable, readable segments; Agglomerative Hierarchical Clustering and DBSCAN are expected to be infeasible at this scale on the hardware available.

- Account balance and savings will rank among the attributes that shape the segmentation most..

3.4. Dataset Description

The dataset comprises 631,062 individual customers from a Vietnamese commercial bank, captured at a single point in time. The variables fall into two groups. The first group consists of the eight financial-behavioural variables that were supplied directly to the clustering model (Table 3.1). The second group consists of the descriptive, demographic, and product-holding variables that were deliberately excluded

from the clustering and instead reserved for external validation and comparison of the resulting segments (Table 3.2). Separating the two groups is essential to the validation strategy of Section 4.3.3: if the segments derived from the eight inputs also differ systematically on the held-out variables, this provides external evidence that the segmentation reflects a genuine behavioural structure rather than an artefact of the chosen inputs.

3.4.1. Data Source, Extraction, and Quality

The data used in this study were obtained from Eximbank (Vietnam Export Import Commercial Joint Stock Bank), a long-established commercial bank in Vietnam with a nationwide branch network. The records were extracted in 2025 by the bank's internal Data Department directly from the core banking system (Fincore), the system of record in which all individual customer accounts and transactions are maintained. Because the extraction was performed by the department that owns and operates the source system, the variables correspond to the bank's official definitions rather than to a third-party or survey-based reconstruction.

To situate the dataset, it is useful to note the scale and standing of the source institution. Eximbank was established in 1989 as one of Vietnam's first joint-stock commercial banks and is listed on the Ho Chi Minh Stock Exchange under the ticker EIB; by the end of the 2025 financial year its total assets had reached approximately VND 273 trillion (Eximbank, 2026). It operates a nationwide branch network and offers a full range of retail products - current and savings accounts, loans, cards, and bancassurance - to both individual and corporate customers. A retail customer base in the order of several hundred thousand active individuals, as analysed here, is therefore fully consistent with the operating scale of a bank of this size.

This scale is, in turn, consistent with the broader Vietnamese retail-banking market. The sector has undergone sustained expansion in branch density and digital adoption (FiinGroup, 2024; Statista, 2025), and as part of a nationwide biometric-verification programme more than 151 million individual customer records had been authenticated across the banking system by early 2026 (Vietnam.vn, 2026). These external figures are not used in the analysis itself; they are reported here to document that the dataset reflects a real, verifiable institutional and market context rather than a constructed or hypothetical one.

The sampling frame is the bank's entire population of active individual (retail) customers - that is, every personal customer whose account had not been closed as of the extraction date. The dataset is therefore a complete enumeration of the active retail customer base rather than a probabilistic sample of it. This distinction is methodologically important: because no sampling was performed, the analysis is not exposed to sampling error or to the selection biases that arise when a subset of customers is drawn, and the 631,062 records represent the full target population to which the conclusions apply. Corporate and institutional customers were excluded by design, since

the behavioural variables and the segmentation objective are specific to individual banking relationships.

Several aspects of data quality were assessed before analysis. Missing values were distinguished from genuine zeros, because in a banking context the two carry different meanings: a missing value indicates that information was not recorded, whereas a zero indicates that the customer genuinely does not hold or use the product in question. Null rates across the clustering variables were found to be negligible, while several credit and accumulation variables exhibited high proportions of genuine zeros (for example, customers with no outstanding loan or no savings balance); this zero-inflation is a substantive feature of the data and is addressed explicitly in the preprocessing and interpretation. As a single-source extraction from the system of record, the data are not subject to the integration inconsistencies that affect datasets merged across multiple systems. All records were de-identified for the purposes of this research, and the data were handled in accordance with the bank's data-confidentiality and privacy requirements; no individual customer can be identified from the variables used.

Because the dataset is proprietary, customer-level banking data, it is not publicly available and was made available to the author by Eximbank for the sole purpose of this research and under the bank's data-confidentiality requirements. As a confidential, single-source extraction from the bank's system of record - rather than a public or survey dataset - it cannot be released or independently reproduced, and the variable definitions used correspond to the bank's official internal definitions. This arrangement is standard for research conducted on real institutional banking data, where the analytical value of a complete, real-world customer population is balanced against the confidentiality obligations attached to it.

3.4.2. Variables Used as Clustering Inputs

To construct meaningful customer segments, the clustering model requires variables that capture different aspects of the customer–bank relationship rather than multiple measures of the same underlying behaviour. Accordingly, eight financial-behavioural variables were selected to represent the major dimensions of retail banking activity, including liquidity management, transaction activity, savings behaviour, product engagement, credit relationships, and card usage. These variables are summarised in Table 3.1.

Table 3.1. Financial-behavioural variables used as clustering inputs

No.	Variable	Description	Behavioural Dimension
1	SDBQ_CASA	Average current-account (CASA) balance	Transactional liquidity
2	BUT_TOAN	Number of accounting entries	Activity intensity
3	DU_NO_EIB	Outstanding loan balance	Credit relationship
4	SL_PRODUCT_HOLDING	Number of products held	Product engagement
5	SO_NGAY_KH_GD	Number of active transaction days	Engagement frequency
6	DS_THE_TD_BQ	Average credit-card spending	Card usage behaviour
7	SDBQ_TIET_KIEM	Average savings balance	Wealth accumulation
8	GIATRI_TSDB_VND	Collateral value	Secured credit

The selection of these variables was guided by two principles. First, each variable must represent a meaningful aspect of customer financial behaviour from a business perspective. Second, the variables should provide complementary rather than redundant information to maximise the effectiveness of the clustering process. Together, they capture how customers use the bank for transactions, savings, borrowing, payments, and product consumption, thereby enabling the identification of multidimensional customer profiles.

All variables were measured at the individual customer level, with one observation corresponding to one unique customer identifier (CIF). Data were extracted from the bank's core banking and card-management systems and aggregated over the full observation period from 1 January 2025 to 31 December 2025.

Average CASA Balance (SDBQ_CASA): This variable represents the time-weighted average daily balance maintained in the customer's current and demand-

deposit accounts throughout 2025, measured in Vietnamese Dong (VND). It serves as the most direct indicator of transactional liquidity and reflects the extent to which customers utilise the bank as their primary institution for everyday financial activities. Customers with high CASA balances typically receive salary payments through the bank, maintain operational cash balances, and actively use banking services for payments and transfers. As CASA deposits constitute a low-cost funding source, this variable is widely recognised as an important measure of customer value in banking segmentation studies (Firdaus & Utama, 2020).

Number of Accounting Entries (BUT_TOAN): This variable counts the total number of debit and credit accounting entries generated by the customer during the observation period. Unlike balance-based indicators, it measures the intensity of banking activity and captures how actively customers interact with their accounts. The variable distinguishes highly active customers from dormant or infrequent users regardless of the amount of money held in their accounts. Consequently, it provides an important behavioural dimension related to transaction frequency and service utilisation.

Outstanding Loan Balance (DU_NO_EIB): Outstanding loan balance represents the principal amount owed by the customer to the bank as of 31 December 2025. It captures the depth of the customer's credit relationship and reflects participation in one of the bank's most profitable business lines. Customers with larger outstanding balances generally generate higher interest income but also expose the bank to greater credit risk. Therefore, this variable serves as a key indicator of both customer value and risk exposure within the retail banking portfolio.

Number of Products Held (SL_PRODUCT_HOLDING): This variable is calculated as the count of distinct product categories currently used by the customer, including current accounts, savings products, loans, cards, and other retail banking services. It measures relationship breadth and product engagement, reflecting the extent to which customers are integrated into the bank's ecosystem. Previous research has shown that broader product ownership is associated with stronger customer loyalty, higher switching costs, and greater lifetime value (Reinartz & Kumar, 2003). Consequently, this variable provides insight into the depth and stability of the customer relationship.

Number of Active Transaction Days (SO_NGAY_KH_GD): This variable counts the number of distinct calendar days during which the customer conducted at least one transaction. While the number of accounting entries measures transaction volume, active transaction days capture transaction regularity and behavioural consistency. For example, two customers may generate the same number of transactions, but the customer who transacts throughout the year exhibits a stronger habitual relationship with the bank than one whose activity is concentrated into a small number of days. This variable therefore measures the continuity of customer engagement.

Average Credit-Card Spending (DS_THE_TD_BQ): Average credit-card spending measures the average value of card transactions generated by the customer during the observation period. A small number of negative values may occur due to refunds, charge reversals, or balance adjustments and are retained because they represent

genuine customer activity. These variable isolates card usage behaviour, which constitutes a distinct source of revenue through interchange fees, merchant fees, and revolving credit interest. It also provides insight into customers' payment preferences and adoption of cashless financial services.

Average Savings Balance (SDBQ_TIET_KIEM). This is the time-weighted average balance held in savings and term-deposit accounts over the course of 2025. CASA balances mostly reflect day-to-day transacting, whereas savings balances are a different matter, they speak to wealth building and a longer-term commitment to the bank. Customers who hold sizeable savings tend to bring stable funding and usually have deeper financial resources behind them. This variable therefore captures a behavioural dimension that is conceptually distinct from transactional liquidity.

Collateral Value (GIATRI_TSDB_VND). This is the appraised value, in VND, of whatever the customer has pledged to secure borrowing—residential or commercial property, vehicles, or other eligible assets. Placed alongside outstanding loan balance, it rounds out the credit picture by carrying both the exposure and the security that stands behind it. Two customers may owe exactly the same amount yet sit in very different risk positions once their collateral diverges. This makes collateral value a useful read on customer wealth and credit risk at the same time.

The eight variables were deliberately spread across separate behavioural dimensions, so the model would carve out customer groups from several sides of financial behaviour rather than one dominant trait. Between them they cover transactional liquidity (CASA), banking activity (accounting entries and active days), wealth accumulation (savings), credit relationships (loans and collateral), product engagement (how many products are held), and card-usage behaviour. The fairly low multicollinearity in the data - the largest pairwise correlation runs to about 0.50 - backs that design empirically and says no one variable is just standing in for another. Each one therefore brings something the others do not, which sharpens the algorithm's ability to draw rich, meaningful customer profiles.a result, each variable contributes unique information to the segmentation process, enhancing the ability of the clustering algorithm to construct rich and meaningful customer profiles.

3.4.3. Variables Used for Comparison and External Validation

In addition to the eight financial-behavioural variables used as clustering inputs, the dataset contains a broader set of demographics, socio-economic, geographic, income-related, credit-quality, and product-holding variables. These variables were deliberately excluded from the clustering process and therefore played no role in determining cluster membership. Instead, they were retained as **held-out variables** for post-clustering analysis, segment interpretation, and external validation.

The use of held-out variables follows established practice in customer segmentation research. If variables used to describe and validate segments are also used to construct them, any observed differences may simply reflect the mechanics of the clustering algorithm rather than meaningful behavioural distinctions. By withholding these variables from the clustering stage, the analysis provides a more rigorous

assessment of whether the identified segments correspond to genuinely different customer profiles.

Specifically, these variables serve two purposes. First, they are used to profile and interpret the resulting customer segments by examining differences in demographic characteristics, socio-economic status, income levels, geographic distribution, and product ownership patterns. Second, they are used to evaluate segment distinctiveness through lift analysis, as presented in Section 4.3.3 and Figure 4.2. Significant differences across these held-out variables provide external evidence that the clustering model has identified meaningful and business-relevant customer groups.

Table 3.2. Held-out variables used for comparison and external validation of the segments

No.	Variable	Description	Category
1	GENDER	Customer gender	Demographic
2	NHOM_TUOI	Age group	Demographic
3	MARITAL_STATUS	Marital status	Demographic
4	EDUCATION_LEVEL	Education level	Demographic
5	NATIONALITY	Nationality	Demographic
6	NGHE_NGHIEP	Occupation sector	Socio-economic
7	HOUSING_OWNERSHIP	Home ownership status	Socio-economic
8	KHU_VUC	Region where the CIF was opened	Geographic
9	TINH_THANH_CU_TRU	Province or city of residence	Geographic
10	TONG_THU_NHAP_CN	Total monthly personal income	Income
11	THU_NHAP_CN_NAM	Total annual personal income	Income
12	NHOM_NO	Debt group (loan classification)	Credit quality

No.	Variable	Description	Category
13	STATUS_CREDIT_CARD	Credit card status	Product status
14	HAS_GIAO_DICH	Has at least one transaction	Product-holding flag
15	HAS_CASA	Holds a CASA account	Product-holding flag
16	HAS_TK	Holds a savings account	Product-holding flag
17	KH_VAY	Has an active loan	Product-holding flag
18	KH_TSDB	Has pledged collateral	Product-holding flag
19	KH_SD_GN	Uses debit card or digital channel	Behavioural flag

The held-out variables can be grouped into several conceptual categories.

Demographic Variables.

Gender, age group, marital status, education level, and nationality provide information about the personal characteristics of customers. These variables are commonly used in banking research because they are associated with differences in financial needs, product preferences, risk tolerance, and lifecycle stages. Importantly, they were excluded from the clustering model to ensure that customer segments were formed based on observed financial behaviour rather than demographic attributes alone.

- Socio-economic Variables.

Occupation sector and housing ownership status capture aspects of customers' socio-economic position. Occupation is often related to income stability, borrowing capacity, and financial-service needs, while home ownership provides an indication of wealth accumulation and long-term financial security. Examining these variables across clusters helps determine whether behaviourally defined segments also differ in socio-economic characteristics.

- Geographic Variables.

Region of account opening and province or city of residence provide information about customers' geographic distribution. Banking behaviour often varies across regions due to differences in economic development, urbanisation, and access to financial services. Geographic profiling therefore assists in identifying regional concentrations of particular customer segments and supports geographically targeted business strategies.

- Income Variables.

Monthly and annual income were retained as validation variables because income is expected to influence many aspects of financial behaviour, including savings, borrowing, and transaction activity. Since income was not used directly in the clustering process, significant differences in income across clusters would indicate that the behavioural segments correspond to economically meaningful distinctions among customers.

- Credit-Quality Variables.

The debt-group classification (NHOM_NO) provides a measure of customer credit quality and repayment performance. This variable enables assessment of whether particular behavioural segments exhibit systematically different risk profiles. Such information is particularly valuable from a risk-management perspective because it allows the bank to evaluate whether certain customer groups are associated with higher or lower credit risk.

- Product Status and Product-Holding Variables.

Credit-card status and the binary product-holding indicators provide information regarding ownership and usage of major banking products. Variables such as HAS_CASA, HAS_TK, KH_VAY, and KH_TSDB indicate whether customers participate in specific product categories, while STATUS_CREDIT_CARD captures the operational status of card usage. These variables are especially useful for evaluating whether the clusters differ in terms of product adoption, cross-selling potential, and relationship depth.

- Behavioural Flag Variables. KH_SD_GN marks out customers who actively use debit-card services or digital banking channels. It is clearly tied to financial activity, but it was deliberately kept out of the clustering inputs rather than included alongside the continuous features. The reason is that it works best as a validation indicator: a discrete flag of digital usage rather than a continuous behavioural measure on the same footing as balances or transaction values. By holding it back from the inputs, we leave it free to test the segmentation from the outside. Once the clusters are formed, differences in this flag across segments become independent evidence of how digitally engaged each group is—evidence that did not help shape the clusters in the first place, and so carries more weight as confirmation.

Taken together, these held-out variables give an independent yardstick for judging how good and how relevant the clustering is to the business. If the segments come apart meaningfully across demographic, socio-economic, geographic, income, credit-quality, and product-holding lines, that points to clusters capturing real customer behaviour rather than statistical noise. So the held-out variables end up doing serious

work in validating how useful and how managerially relevant the segmentation actually is.

3.5. Exploratory Data Analysis

Exploratory data analysis was conducted to understand the statistical properties of each clustering input before any transformation. Descriptive statistics for central tendency, dispersion, and distribution shape are reported together in Table 3.3.

Variable	Unit	Min	Max	Mean	Median	Std. Dev.	Skewness
CASA balance (SDBQ_CASA)	mil. VND	0.00	48,263.18	14.91	0.78	160.85	132.95
No. of entries (BUT_TOAN)	count	1	528,220	431	35	5,523	41.63
Outstanding loan (DU_NO_EIB)	mil. VND	0.00	567,330	158.61	0.00	2,056.53	128.95
No. of products (SL_PRODUCT_HOLDING)	count	1	9	1.88	2	1.04	1.34
Active txn days (SO_NGAY_KH_GD)	days	1	253	51.03	22	58.36	1.69
Card spending (DS_THE_TD_BQ)	mil. VND	-2.12	578.04	0.66	0.00	5.97	17.83
Savings balance (SDBQ_TIET_KIEM)	mil. VND	0.00	863,500	166.09	0.00	2,816.51	135.58
Collateral value (GIATRI_TSDB_VND)	mil. VND	0.00	956,710	166.48	0.00	2,690.04	122.57

Table 3.3. Descriptive statistics, units, and range of the eight clustering variables (n = 631,062).

Notes: Monetary variables are shown in million VND (mil. VND). For the zero-inflated variables (loan, savings, collateral, card) the minimum is 0; card spending takes a small negative minimum from refunds/adjustments. The wide gap between the median and the maximum reflects the severe right-skewness discussed below.

Three things stand out. First, in most variables the mean sits well above the median, the signature of heavy right-skew: SDBQ_CASA, for one, averages around 14.9 million against a median of 784,502 - close to 19 to 1 - which confirms a long right tail.

Second, missing values and zero values were examined separately, on the logic that a null (no data) and a zero (the customer genuinely holds none of this product) mean

different things to the business. Null rates were trivial; zero rates, though, ran very high on the credit and accumulation variables - 85.7% for outstanding loans, 82.9% for savings, 92.9% for credit-card spending, 97.2% for collateral value. That zero-inflation says these four variables mainly sort customers along a has / does-not-have line rather than a more / less one - which shapes both how the data gets preprocessed and how the final clusters are read.

Third, skewness and kurtosis were quantified for each variable. Six of the eight variables exhibited severe skewness exceeding 3, with SDBQ_TIET_KIEM at 135.6 and SDBQ_CASA at 133.0. Because clustering relies on distance computations, a variable with skewness of 133 and a maximum of 48 billion would entirely dominate a variable such as the number of products, whose maximum is only 9. This diagnosis provides the objective justification for applying a variance-stabilizing transformation. A correlation analysis further confirmed the absence of strong multicollinearity: the highest pairwise correlation was 0.50 (between number of products and active transaction days), well below the threshold that would indicate redundancy. The full correlation structure is shown in Figure 3.8.

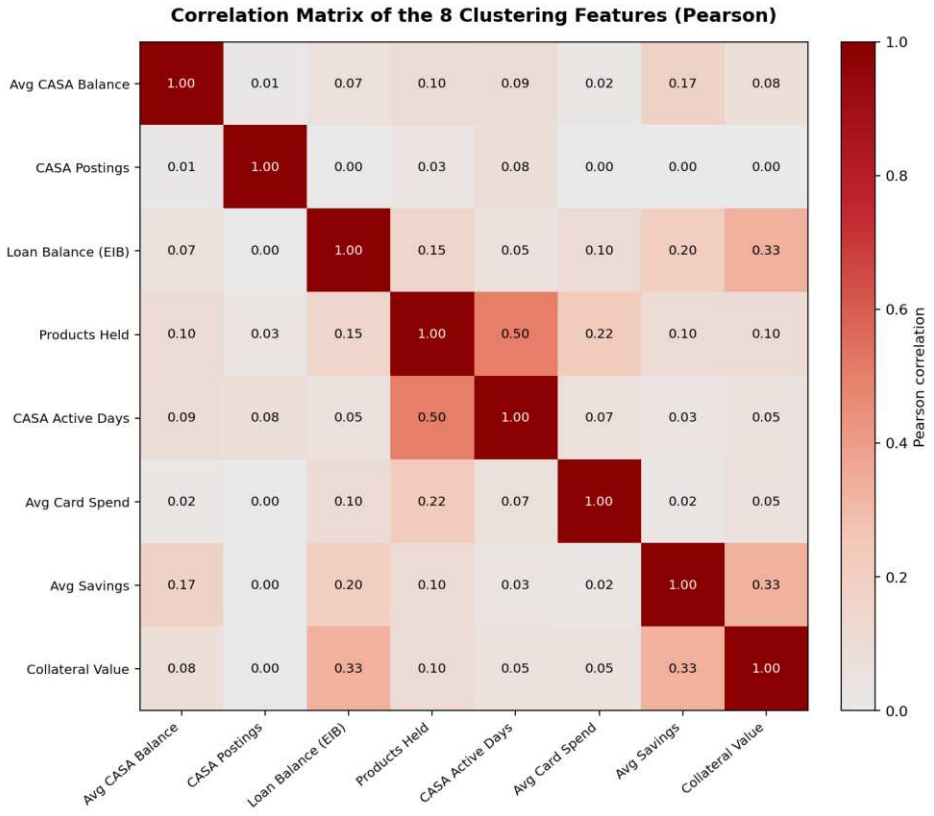


Figure 3.8. Pearson correlation matrix of the eight financial-behavioural variables.

3.6. Data Preprocessing

Guided by the diagnostics in Section 3.5, preprocessing proceeded in four steps: data understanding and variable selection, data cleaning, logarithmic transformation, and feature scaling.

3.6.1. Data Understanding and Cleaning

The first stage acquired the dataset and established the meaning, type, and analytical relevance of each variable; identifiers and fields without analytical value were removed. The second stage addressed data quality. Missing values were examined and found to be negligible; a clear distinction was maintained between missing values and genuine zeros. Negative values, which are not meaningful for the monetary and count variables, were clipped to zero. Outliers were retained rather than removed, because in this domain extreme values represent genuine and important customers (for example, high-net-worth depositors); the scaling strategy was instead chosen to be robust to them.

3.6.2. Logarithmic Transformation (log1p)

In order to correct for the skewness of the EDA results, each of the 8 variables was transformed using the log1p transformation ($\log1p(x) = \log(1 + x)$). This approach was utilized in lieu of the regular logarithm due to two reasons: it draws in the long right tail, and it can correctly handle zero values (i.e. $\log1p(0) = 0$); therefore, it is especially important for the two zero-inflated variables. The skewness of the two variables decreased substantially as a result of transformation: SDBQ_CASA from 133.0 to .54, and BUT_TOAN from 41.6 to .98. However, two of the variables (DS_THE_TD_BQ (3.44) and GIATRI_TSDB_VND (5.76)) maintained their high levels of skewness after they were transformed due to their extremely high levels of zero-inflation, where the concentrated mass at zero cannot be normalized from the distribution of logarithmic values. Thus, reaffirming the almost binary status of these two variables.

Figure 3.9a — Distribution of the four continuous variables after log(1+x) transformation

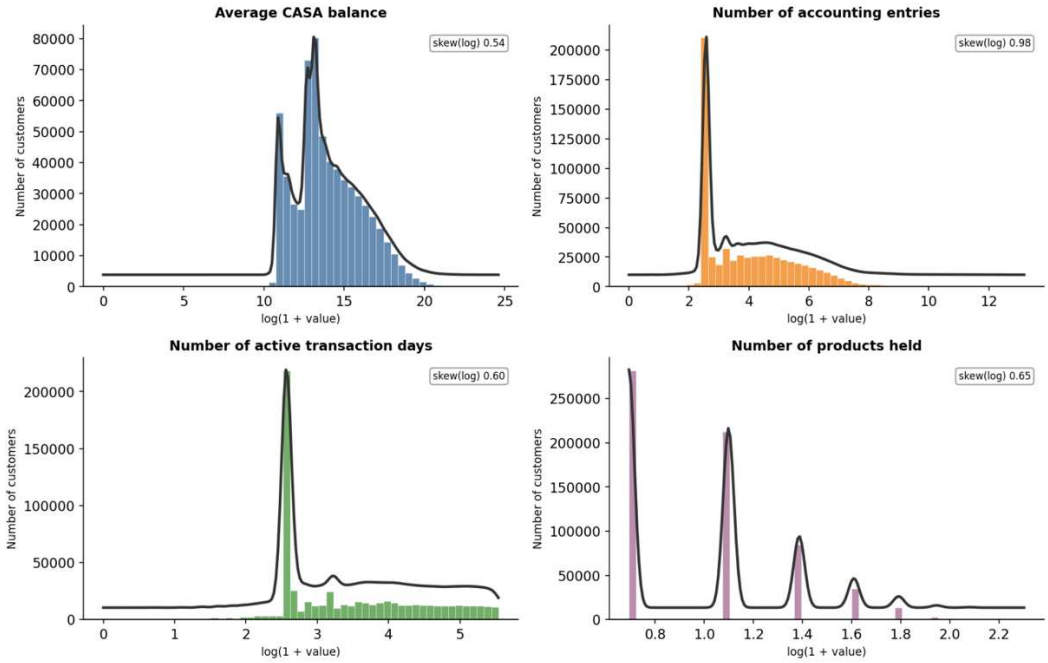
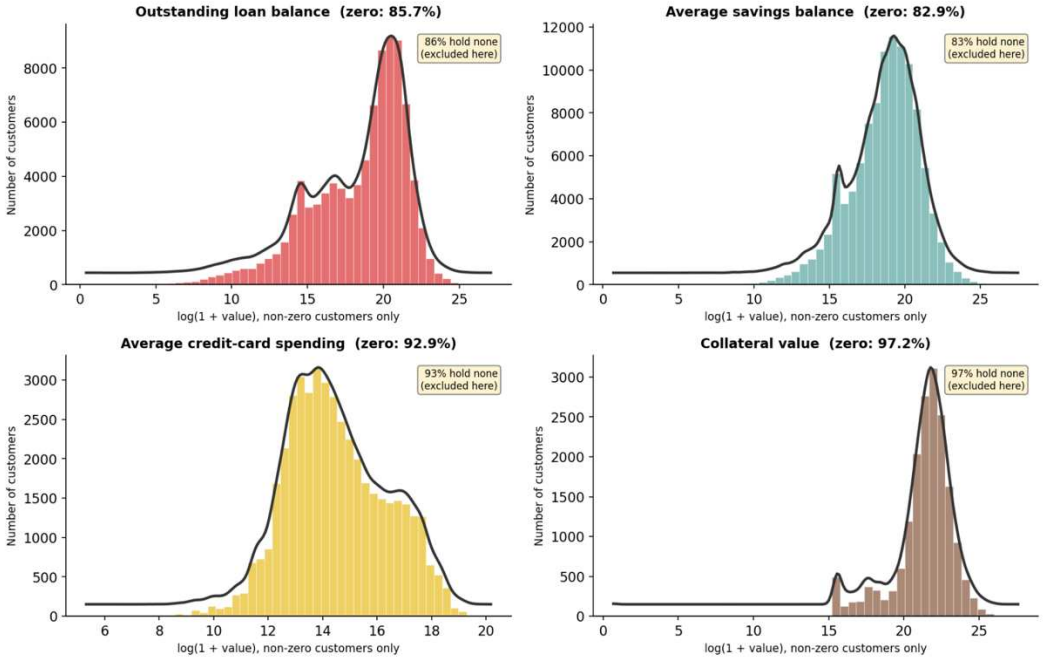


Figure 3.9b — Zero-inflated variables: share at zero, and distribution of the non-zero part (log scale)



[Figure 3.9a and 3.9b: Distribution of each variable after the log1p transformation]

3.6.3. Feature Scaling with RobustScaler

After the log transform the variables still lived on different scales, so they had to be standardised for each to pull its weight in the distance calculations clustering depends on. RobustScaler did that job: subtract the median, divide by the interquartile range (IQR), and every variable comes out with a median of 0 and an IQR of 1. StandardScaler leans on the mean and standard deviation; RobustScaler does not, which is exactly why it shrugs off outliers - the median and IQR simply are not thrown by extreme values, and Section 3.5 documented plenty of those. The preprocessed data is called df_robust and feeds straight into the clustering algorithm.

3.6.4. Justification for the log1p + RobustScaler Combination

Although several preprocessing pipelines were technically available - raw data, log1p only, log1p + StandardScaler, and log1p + RobustScaler - the choice among them was made empirically rather than by convention. Three lines of evidence converge on the log1p + RobustScaler combination.

Cluster-tendency evidence. The Hopkins statistic, which measures whether a dataset has a meaningful clustering tendency (a value near 1 indicates strong structure, a value near 0.5 indicates uniform randomness), was computed for every pipeline. The values were 0.97 for the raw data, 1.000 for log1p, 0.9998 for log1p + StandardScaler, and 1.000 for log1p + RobustScaler. All exceeded the 0.75 threshold, but the log1p + RobustScaler version reached the maximal value of 1.000, confirming the strongest clustering tendency.

Dimensionality-reduction evidence. Principal Component Analysis (PCA) was applied to each pipeline, and the projections were inspected in two and three dimensions. The log1p + RobustScaler representation produced the clearest visual separation of customer groups and, critically, the highest proportion of variance explained by the leading principal components: in the 2-D projection the first two components captured a larger share of total variance than under any competing pipeline, and the 3-D projection added further separation between groups. In other words, after log1p + RobustScaler the informative structure of the data is concentrated in a few leading components, which is precisely the condition under which distance-based clustering performs well. t-SNE projections corroborated this finding, showing visually distinct groups.

Evidence of linearity can be found in the comparison between Linear and Kernel PCA on data scaled using Log1p and RobustScaler. The results show the amount of variance explained by the two models are similar enough that the underlying cluster structure is likely linear. This information has methodological importance because it provides justification for the use of K-Means, an algorithm that is predicated on convex linearly separable cluster boundaries, rather than a more complex nonlinear algorithm.

Beyond these empirical results, the combination carries intrinsic advantages. The log1p step stabilizes variance and brings the heavy-tailed monetary variables onto a comparable footing with the count variables, while preserving the meaning of zeros; the RobustScaler step then equalizes scale without letting the documented extreme values dominate the median-and-IQR statistics. Taken together, the log1p + RobustScaler pipeline explains more of the data's structure than the alternatives and is the most faithful to the distance assumptions of K-Means. It was therefore selected as the input representation for all subsequent modelling.

3.7. Cluster-Tendency Assessment and Algorithm Selection

Before applying any clustering algorithm, it is essential to confirm that the data possesses an inherent cluster structure rather than being a uniformly distributed cloud of points. As reported in Section 3.6.4, the Hopkins statistic on the log1p + RobustScaler data was 1.000, far above the 0.75 threshold, and the PCA and t-SNE projections revealed clearly separated groups. These results provide strong methodological justification for proceeding with segmentation.

Three clustering algorithms were considered for the numerical data - MiniBatch K-Means, Agglomerative Hierarchical Clustering, and DBSCAN - as stated in the hypotheses. In practice, however, only one of them proved feasible at the scale of this study.

Agglomerative Hierarchical Clustering requires the construction and repeated updating of a pairwise-distance structure with a memory cost on the order of n-squared. For n = 631,062, the distance matrix alone would contain on the order of 4x10^11 entries, requiring hundreds of gigabytes of memory - far beyond the capacity of the machine used in this study. The algorithm therefore could not be run on the full dataset.

DBSCAN relies on density-based neighbourhood queries whose worst-case complexity is O(n-squared), and it materializes large neighbourhood structures in memory. On the 631,062-row dataset it either exhausted available memory or failed to terminate within a practical time. In addition, its sensitivity to the epsilon and minPts parameters in this relatively high-dimensional, zero-inflated space made stable results difficult to obtain.

MiniBatch K-Means, by contrast, processes the data in small batches with a per-iteration cost independent of the full dataset size, and completed the entire grid search over the number of clusters in approximately 3.3 minutes on the same hardware. It was thus the only one of the three candidate algorithms that could be executed on the full population. Combined with the approximately linear structure of the data established by the PCA/Kernel-PCA comparison - which matches the convex-boundary assumption of K-Means - and with the strong precedent in the literature where K-Means consistently outperformed alternatives for banking segmentation, MiniBatch K-Means was selected as the primary clustering algorithm for this study.

3.8. Clustering Configuration

The K-means algorithm has been initialized using k-means++. The first centroid is randomly placed, then the centroids drawn afterwards are selected by determining the probability proportional to their squared distance from the centroids selected so far; this ensures that newly selected centroids do not end up too close together thus making it less probable that the k-means algorithm will find a poor local optimum. The additional parameters were set to the following values: n_init=20: The K-means algorithm was initially trained with 20 independent initializations (training runs) with whichever K-means model provided the lowest WCSS retained for the final output or selection. batch_size=10,000: number of points to submit for centroid updating at each batch of centroid updates max_iter=500: max number of iterations K-means will perform tol=10(-4): stopping criterion; process halts when the centroids move less than this value random_state=42: designated fixed random seed so that the run can be replicated exactly.

3.9. Determining the Optimal Number of Clusters

Selecting an appropriate number of clusters K is a critical step. The study does not rely on a single metric but combines four evaluation metrics and aggregates them into a normalized score to reach a more objective decision. WCSS and the Elbow method capture the total within-cluster distance; the Silhouette coefficient measures separation; the Davies-Bouldin Index measures the ratio of within-cluster dispersion to between-cluster distance; and the Calinski-Harabasz Index measures the ratio of between-cluster to within-cluster dispersion. Because the Silhouette, DBI, and CH metrics are expensive to compute on all 631,062 observations, they are evaluated on a fixed random sample of 50,000 customers - drawn once and reused for every value of K - to ensure a fair comparison.

Because the four metrics have different scales and directions, each is normalized to the range [0, 1] (reversing the direction for the lower-is-better metrics, WCSS and DBI). The sum of the four normalized metrics forms a composite score with a maximum of 4.0, and the value of K with the highest composite score is selected as optimal. The grid search was carried out for K ranging from 3 to 10. For each K, the model was trained on all 631,062 customers (no sampling during training, to avoid density bias), while the cluster-quality metrics were computed on the fixed 50,000-customer sample.

3.9.1. Implementation Procedure

The full procedure executed as a single continuous program, working through six steps that ran from data preparation to the labelling of every customer: preprocessing; a grid search over K (3 to 10); normalised scoring and the choice of K; analysis of the selected clustering; stability validation; and labelling all 631,062 customers.

Step 1 - Preparing the input. The raw data was pulled across the eight features. Missing values were dealt with, negatives clipped to zero, a log1p transformation

applied, and the result scaled with RobustScaler to produce df_robust. One fixed sample of 50,000 customers was drawn a single time and reused across every value of K when the metrics were computed.

Step 2 - Grid search over K. MiniBatch K-Means was trained for each K from 3 to 10 on the full population, logging WCSS, the three quality metrics on the fixed sample, and how the cluster sizes came out. Every K got 20 starts, and the best solution was kept. MiniBatch K-Means was trained for each K from 3 to 10 on the full population, recording WCSS, the three cluster-quality metrics on the fixed sample, and the cluster-size distribution. Each K was initialized 20 times, keeping the best solution.

Step 3 - Normalized scoring. Each metric was normalized to [0, 1] across the range of K and summed into a composite score (maximum 4.0); the individual elbow and metric extrema were also recorded for cross-checking. The K with the highest composite score was selected.

Step 4 - Analyzing the optimal clustering. For the chosen K, the program computed the size and share of each cluster, the median of each variable on the original scale, and a Z-score matrix highlighting the distinguishing variables of each cluster, from which a descriptive name was assigned.

Step 5 - Stability validation. The model was retrained with several seeds and compared against the original solution using ARI, with labels matched by the Hungarian algorithm before computation. A threshold of ARI > 0.85 was used to conclude stability.

Step 6 - Labelling all customers. The cluster labels of the optimal model were assigned back to all 631,062 customers via a new cluster_label column (values 0 to 7), producing the final labelled dataset for downstream use.

Chapter 4. Results and Discussion

4.1. Selection of the Number of Clusters

The evaluation metrics for each value of K are summarized in Table 4.1. The Min % column shows the proportion of the smallest cluster, used to flag an imbalanced clustering.

K	WCSS	Silhouette	DBI	CH	Min %
3	21,189,040	0.8001	0.7330	66,153.5	11.7%
4	14,113,904	0.8278	0.8717	73,191.1	4.5%
5	9,473,129	0.8446	0.7323	88,938.5	2.6%
6	6,691,242	0.8606	0.6519	105,266.0	2.6%
7	6,312,284	0.5861	0.7147	93,665.0	2.7%
8 (selected)	4,106,609	0.8729	0.6046	126,722.3	0.9%
9	3,744,789	0.5957	0.6734	122,966.9	0.9%
10	3,541,864	0.5979	0.7303	115,324.4	0.9%

Table 4.1. Evaluation metrics for each value of K (K = 3 to 10).

Considered individually, the metrics suggest different values of K: the Elbow method gives K = 6; the Silhouette coefficient peaks at K = 8 (0.8729); the Davies-Bouldin index reaches its minimum at K = 8 (0.6046); and the Calinski-Harabasz index reaches its maximum also at K = 8 (126,722.3). Thus three of the four cluster-quality metrics agree on K = 8.

The composite normalized score is presented in Table 4.2. The value K = 8 achieves the highest composite score of 3.968 out of a maximum of 4.0 - while simultaneously reaching the maximum (1.000) on all three of the Silhouette, DBI, and CH metrics. K = 8 was therefore selected as the final number of clusters.

K	WCSS_n	Sil_n	DBI_n	CH_n	Total score
3	0.000	0.746	0.519	0.000	1.265
4	0.401	0.843	0.000	0.116	1.360
5	0.664	0.901	0.522	0.376	2.463
6	0.822	0.957	0.823	0.646	3.247
7	0.843	0.000	0.588	0.454	1.885
8 (selected)	0.968	1.000	1.000	1.000	3.968
9	0.989	0.033	0.742	0.938	2.702
10	1.000	0.041	0.529	0.812	2.382

Table 4.2. Composite normalized scores for each value of K (maximum = 4.0).

One notable point is that the Silhouette coefficient is not monotonic in K: it is very high at K = 6 and K = 8 but drops sharply at K = 7 and K = 9. This reflects that, at certain values of K, the algorithm splits or merges clusters in ways that do not match the natural structure of the data, reducing the average separation. Combining multiple metrics and a composite score mitigates the influence of this fluctuation on the decision.

To make the selection transparent, the four cluster-quality criteria are shown separately in Figure 4.5, each plotted against K, with the formula and the decision rule for each.

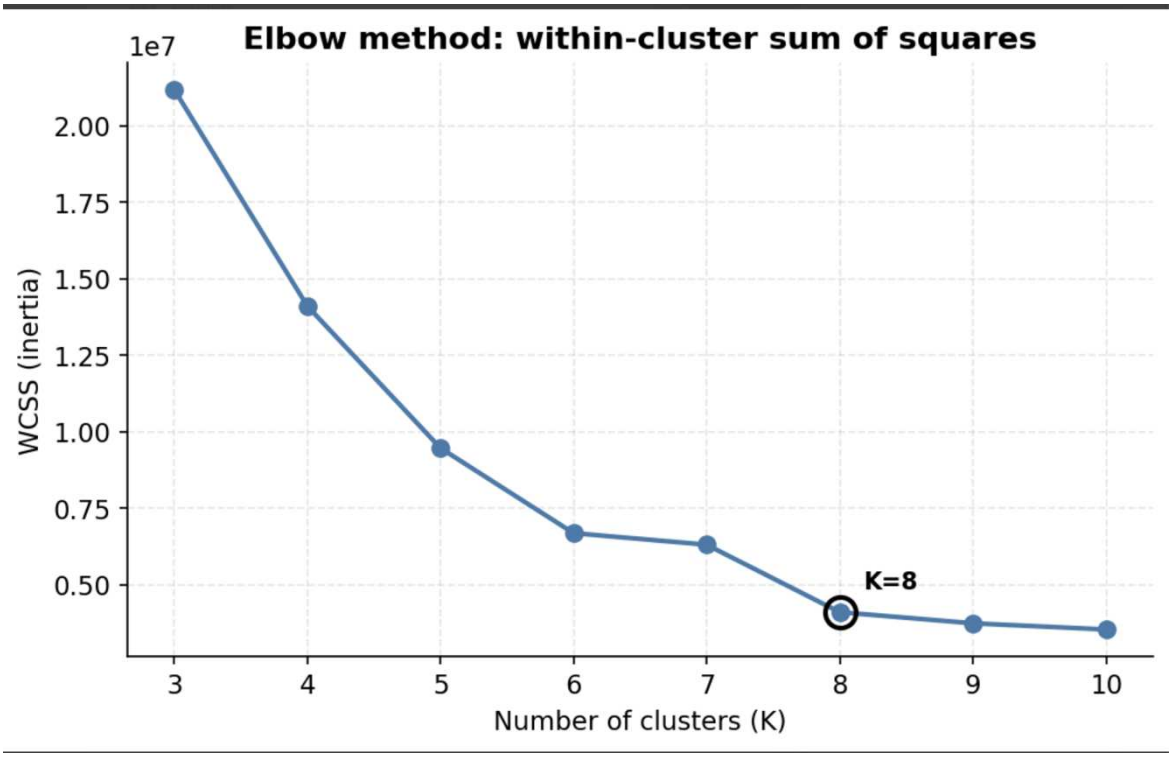


Figure 4.5a. Elbow method: within-cluster sum of squares (WCSS) versus K.

Panel (a) plots the within-cluster sum of squares, WCSS = $\Sigma_k \Sigma_{i \in C_k} \|x_i - \mu_k\|^2$, the total squared distance of each point x_i from its cluster centroid μ_k. WCSS falls monotonically as K rises, so the criterion is the "elbow" - the point beyond which additional clusters yield only marginal reductions. The curve bends most sharply around K = 6–8, after which the gains flatten; the elbow is therefore consistent with, but on its own does not pin down, K = 8.

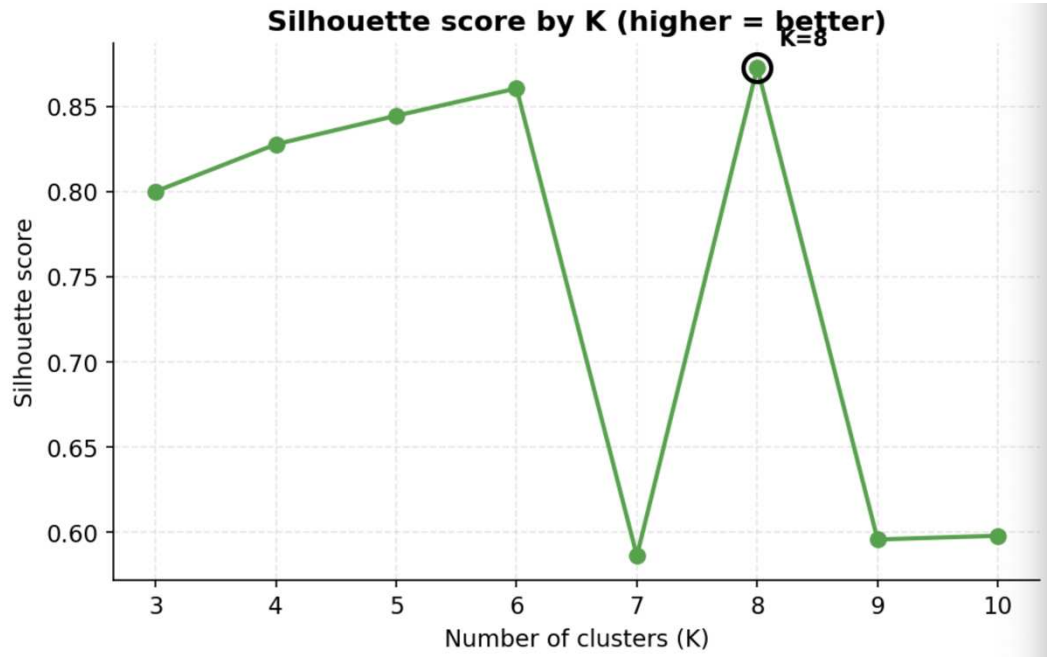


Figure 4.5b. Silhouette score versus K.

Panel (b) plots the mean Silhouette coefficient (Rousseeuw, 1987), defined for each point as s = (b - a) / max(a, b), where a is the mean distance to points in its own cluster and b the mean distance to the nearest other cluster; the score ranges from -1 to 1, and higher is better. It is the clearest single criterion here: it peaks sharply at K = 8 (0.873) and falls steeply at K = 7 and K = 9, indicating that eight clusters give the cleanest separation.

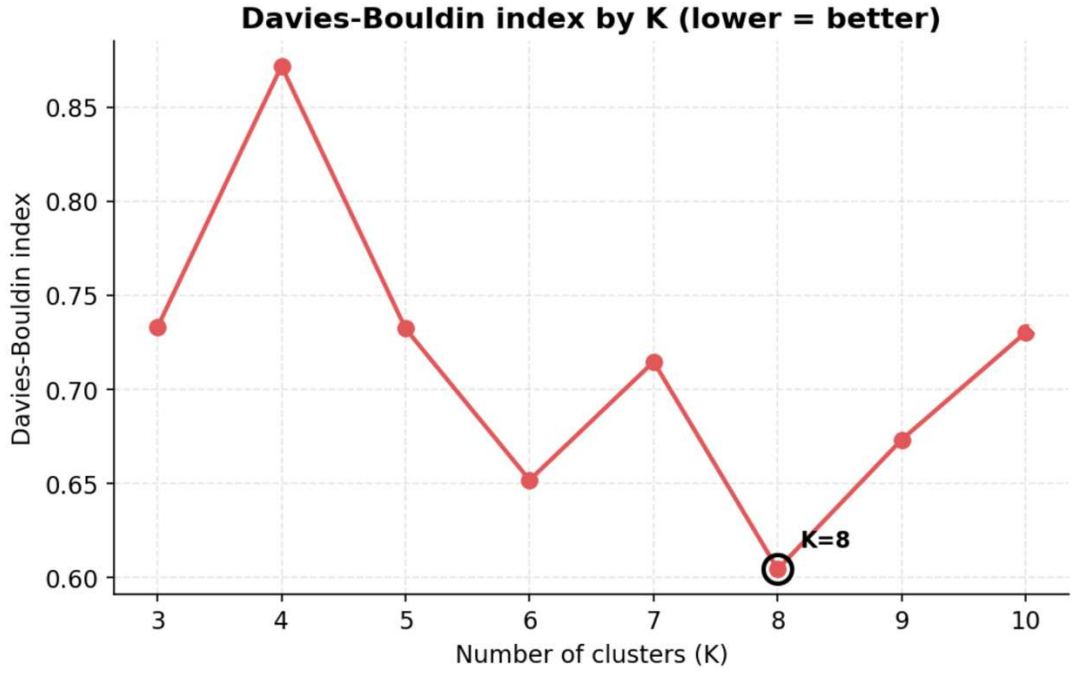


Figure 4.5c. Davies-Bouldin index versus K.

Panel (c) shows the Davies-Bouldin index (Davies & Bouldin, 1979). For each cluster it takes the worst-case ratio of within-cluster scatter to between-cluster separation, then averages these values across all clusters. Lower readings indicate clusters that are more compact and more cleanly separated. The index falls to its lowest point exactly at K = 8 (0.605), which lines up with what the Silhouette criterion already suggested.

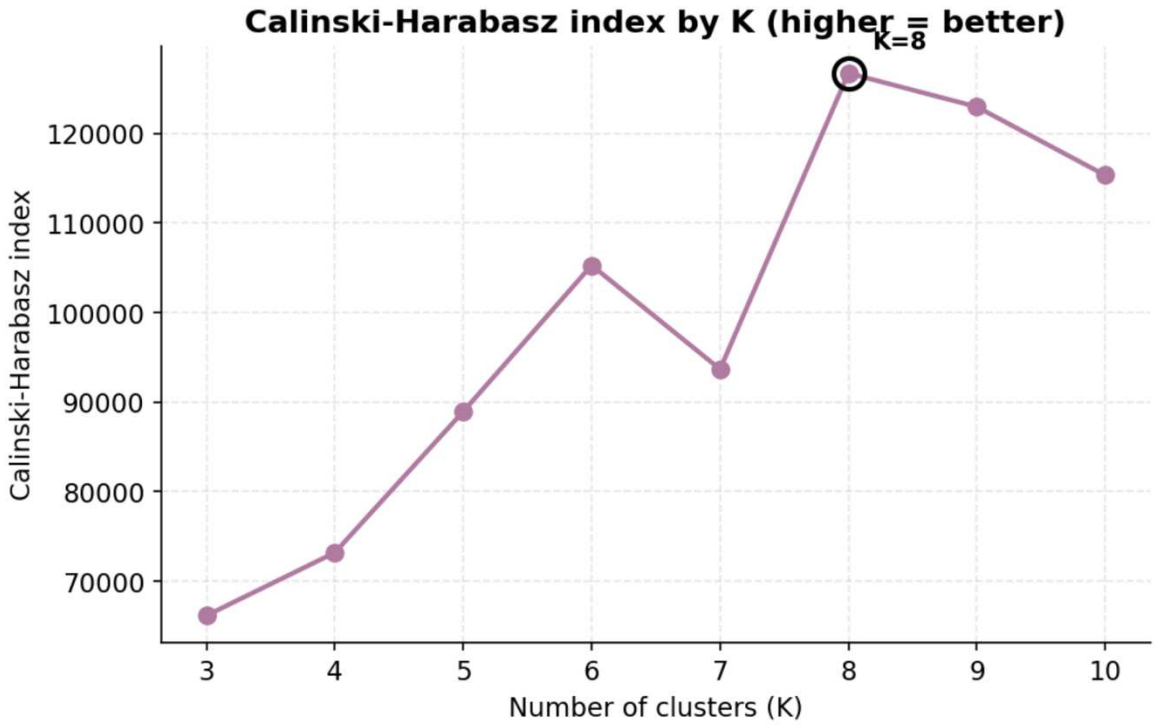


Figure 4.5d. Calinski-Harabasz index versus K.

Panel (d) shows the Calinski-Harabasz index (Caliński & Harabasz, 1974), which measures between-cluster dispersion relative to within-cluster dispersion, adjusted for degrees of freedom. A higher value points to clusters that are more clearly separated. Here the index reaches its maximum at K = 8, with a value of 126,722.

Taken together, this means that three of the four criteria we examined—Silhouette, Davies-Bouldin, and Calinski-Harabasz—each independently favour K = 8, and the elbow in Panel (a) is consistent with the same choice. Table 4.2 brings these signals into a single composite score, and it is this agreement across measures, rather than any one metric alone, that supports eight clusters as the appropriate solution.

4.2. Stability of the Clustering Solution

Because K-Means depends on the random initialization, a solution is only truly reliable if it is reproduced stably across runs with different seeds. The K = 8 model was retrained several times, and the results were compared with the original solution using the Adjusted Rand Index, with cluster labels matched by the Hungarian optimal-assignment algorithm before computation.

The validation results show near-perfect stability: with seed 137 the ARI after mapping reaches 1.0000, and with seed 256 it reaches 0.9999. The mean ARI is 0.9999, far exceeding the 0.85 threshold. This confirms that the K = 8 solution is fully stable and reproducible, independent of the random initialization.

4.3. Characteristics of the Customer Segments

4.3.1. Cluster Size Distribution

At K = 8, Table 4.3 shows how customers fall across the clusters, and the split is plainly lopsided: cluster 0 alone holds 72.9% of everyone, while four clusters (3, 4, 6, and 7) each come in under 3%. This is par for the course with bank customer data - most customers do basic transacting, and the specialised segments (big borrowers, large asset holders, multi-product users) are always a minority.

Cluster	Customers	Share	Preliminary interpretation
0	460,132	72.9%	Mass-market customers
1	78,795	12.5%	Deposit customers
2	29,263	4.6%	Pure-credit customers
3	11,725	1.9%	Secured borrowers
4	15,138	2.4%	Multi-product customers
5	21,890	3.5%	Consumer-credit (card + loan)
6	5,869	0.9%	Premium customers
7	8,250	1.3%	Loyal customers

Table 4.3. Size distribution of the 8 customer segments.

4.3.2. Feature Profiles (Z-scores)

To interpret each cluster, the median of every variable (on the original scale) was computed per cluster and converted to a Z-score relative to the overall mean across clusters. A positive Z-score indicates the cluster is above average; a negative value indicates below average. Table 4.4 presents the Z-score matrix.

Cluster	CASA	ENTRIES	DAYS	LOAN	CARD	PROD	SAVINGS	COLLAT
0	-1.91	-2.31	-2.27	-1.70	-0.77	-2.46	-1.00	-0.58
1	+0.35	-0.16	-0.13	-1.70	-0.77	+0.01	+1.00	-0.58
2	-0.70	-0.30	-0.32	+0.74	-0.77	+0.01	-1.00	-0.58
3	-0.20	+0.11	+0.10	+0.75	-0.77	+0.01	-1.00	+1.76
4	+0.45	+0.82	+0.87	+0.11	+1.30	+0.81	+0.88	-0.58
5	-0.46	+0.25	+0.27	+0.63	+1.43	+0.01	-1.00	-0.58
6	+1.58	+1.33	+1.37	+0.65	+1.14	+0.81	+1.05	+1.71
7	+0.89	+0.26	+0.10	+0.52	-0.77	+0.81	+1.06	-0.58

Table 4.4. Z-score feature matrix of the 8 clusters (positive = above average; negative = below average).

Abbreviation key: CASA - current-account balance; ENTRIES - transaction entries; DAYS - transaction days; LOAN - outstanding credit; CARD - credit-card spending; PROD - number of products; SAVINGS - savings balance; COLLAT - collateral value.

4.3.3. External Validation Using the Remaining Dataset Variables

A clustering solution that is statistically valid on the eight input variables is not, by itself, proof that the segments are meaningful. Because all eight features were used to fit the model, strong separation on those same features is partly circular. A more demanding test is whether the segments also differ systematically on variables that were not supplied to the algorithm - the demographic, geographic, credit-quality, and product-holding variables listed in Table 3.2. If the eight-variable clustering has captured a genuine behavioural structure, that structure should leave a coherent imprint on these external, held-out variables as well. This section therefore cross-validates the K = 8 solution against those held-out fields.

Demographic and behavioural overview. Figure 4.1 summarizes the eight segments across several lenses: the Z-score profile on the clustering features, the binary product-holding flags, and the gender and age distributions. Two findings stand out. First, every customer in every cluster holds a transaction account, a CASA account, and a basic account (all three binary flags equal 100% across clusters 0 to 7), confirming that these universal attributes carry no discriminative power and validating their exclusion from the clustering features. Second, the loan, collateral, and digital-usage flags vary sharply across clusters - they are essentially zero for the mass-market and deposit segments (clusters 0 and 1) but rise to 60 to 100% for the credit-oriented segments (clusters 3, 4, 5, 6) - mirroring exactly the loan and collateral Z-scores obtained from the clustering features. The demographic panels show that age and gender are distributed relatively evenly across the eight clusters, which is the expected result: the segmentation is driven by financial behaviour rather than demographics, so no single cluster collapses onto a particular age band or gender. This is itself a desirable property, since it indicates the segments capture behavioural rather than purely demographic distinctions.

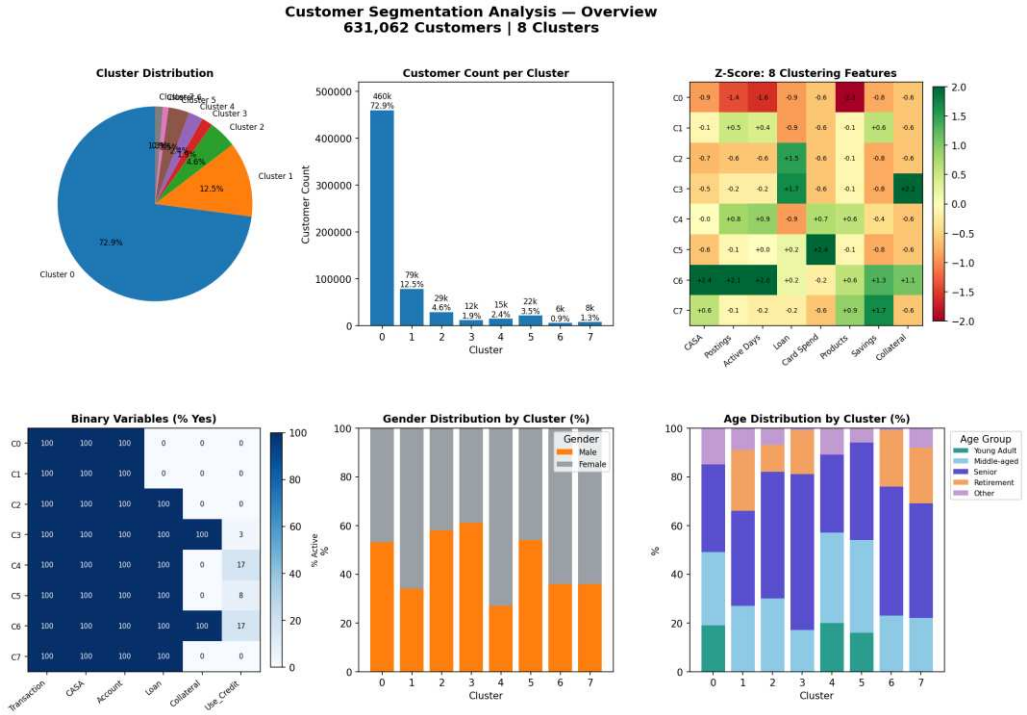


Figure 4.1. Segment overview: cluster sizes, Z-score feature profile, binary product flags, and demographic distributions across the eight segments. (The Z-score panel is computed on cluster medians and may differ slightly in normalization from Table 4.4.)

Lift analysis on held-out variables. To pin down how far each segment strays from the overall base on the held-out variables, a lift measure was worked out for two composite indicators that never touched the clustering: digital-channel usage and loan amount. Lift is a segment's average over the population average - 1.0 means the segment looks just like the typical customer, 10.0 means it runs ten times the norm. Figure 4.2

lays out the results. The universal flags (transaction, CASA, account) post a lift of exactly 1.00 in every cluster, which again confirms they do nothing to tell customers apart. The held-out behavioural indicators do the opposite, throwing up dramatic lift values that line up neatly with the segment readings drawn from the clustering features. To measure how far each segment departs from the overall customer base on variables left out of the clustering, I computed a lift measure for two composite indicators that never entered the clustering: digital-channel usage and loan amount. Lift here is the ratio of a segment's mean value to the population mean. A lift of 1.0 means the segment behaves like the average customer, whereas a lift of 10.0 means its value runs ten times the population norm. Figure 4.2 reports these results. For the universal flags (transaction, CASA, account), the lift comes out to 1.00 across every cluster, which again tells us these attributes do nothing to separate customers. The held-out behavioural indicators tell a different story: their lift values vary widely, and the pattern matches the segment interpretations that emerged from the clustering features.

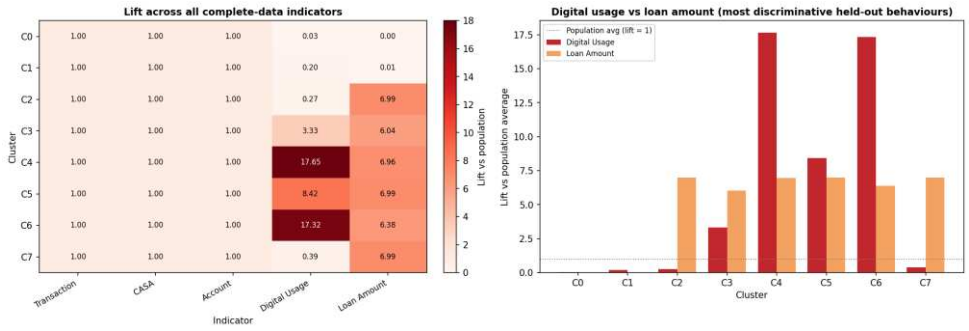


Figure 4.2. Lift by cluster on held-out variables. Left: lift across all complete-data indicators; right: digital usage versus loan amount, the two most discriminative held-out behaviours.

For digital usage, cluster 4 (multi-product customers) records a lift of 17.65 and cluster 6 (premium customers) a lift of 17.32 - meaning these two segments use digital channels roughly seventeen times more intensively than the average customer - while cluster 5 (consumer-credit) reaches 8.42. The mass-market and deposit segments (clusters 0 and 1) sit far below the average at 0.03 and 0.20 respectively, confirming their largely dormant, low-engagement nature. For loan amount, every credit-oriented segment (clusters 2, 3, 4, 5, 7) shows a lift in the range of 6.0 to 7.0, an order of magnitude above the mass-market segment, while clusters 0 and 1 register essentially zero loan activity. Crucially, none of these patterns were imposed by the clustering algorithm: digital usage and loan amount were computed independently from the eight input features. The fact that they reproduce the segment hierarchy so cleanly provides strong external evidence that the K = 8 solution reflects a real behavioural structure in the customer base rather than an artefact of the chosen input variables.

4.4. Random Forest Validation and Predictive Modelling

While the internal metrics (Section 4.1), the stability analysis (Section 4.2), and the held-out lift analysis (Section 4.3.3) all support the eight-segment solution, a final,

operationally motivated test remains: can the segments be reproduced by a supervised classifier, and can that classifier be deployed to assign new customers to a segment without re-running the clustering? To answer this, a Random Forest classifier was trained to predict the cluster label (0 to 7) of each customer from the eight log1p + RobustScaler-transformed input features. The Random Forest serves three purposes simultaneously: (i) it validates the separability of the segments - a classifier can only achieve high accuracy if the clusters occupy distinct, learnable regions of the feature space; (ii) it provides a deployable prediction model for labelling future customers; and (iii) its feature-importance scores offer an independent ranking of which variables drive the segmentation.

4.4.1. Experimental Setup

The labelled set of 631,062 customers was split into a training set (80%, 504,850) and a held-out test set (20%, 126,212), with stratified sampling so the cluster-size distribution carried over into both. The classifier ran with 300 trees (n_estimators = 300), Gini impurity as the split rule, at least two samples per leaf, balanced class weights to offset the dominant mass-market cluster, and a fixed seed (random_state = 42). Generalisation got a further check from 5-fold stratified cross-validation on the training set.

4.4.2. Classification Performance

On the held-out test set the Random Forest reproduced the MiniBatch K-Means labels almost perfectly, as summarized in Table 4.5. It achieved an overall accuracy of 99.85% and a macro-averaged F1-score of 0.996. The 5-fold cross-validated accuracy was 99.83% +/- 0.04%, indicating that the result is not an artefact of a particular train/test split.

Metric	Value
Accuracy (held-out test set)	99.85%
Macro-averaged Precision	0.997
Macro-averaged Recall	0.995
Macro-averaged F1-score	0.996
Weighted F1-score	0.9985
5-fold cross-validated accuracy	99.83% +/- 0.04%

Table 4.5. Overall Random Forest classification performance.

The per-segment breakdown (Table 4.6) shows that precision, recall, and F1 exceed 0.99 for every one of the eight clusters, including the smallest segments. Performance is highest for the dominant mass-market cluster 0 (F1 = 0.9997) and marginally lower - though still excellent - for the smallest premium cluster 6 (F1 = 0.9936, support of only 1,174 test customers), reflecting the smaller amount of training evidence available for the rarest segments.

Cluster	Precision	Recall	F1-score	Support (test)
0 - Mass-market	0.9995	0.9998	0.9997	92,025
1 - Deposit	0.9980	0.9975	0.9977	15,759
2 - Pure-credit	0.9970	0.9962	0.9966	5,853
3 - Secured borrowers	0.9966	0.9949	0.9957	2,345
4 - Multi-product	0.9974	0.9960	0.9967	3,028
5 - Consumer-credit	0.9968	0.9954	0.9961	4,378
6 - Premium	0.9949	0.9923	0.9936	1,174
7 - Loyal	0.9958	0.9939	0.9948	1,650

Table 4.6. Per-segment Random Forest performance on the held-out test set.

An examination of the confusion matrix shows that the few misclassifications are concentrated between behaviourally adjacent segments - chiefly among the credit-oriented clusters 2, 3, and 5, and among the high-value clusters 4, 6, and 7 - which is consistent with the fact that these groups differ along secondary rather than primary dimensions. No systematic confusion occurs between the mass-market segment and any specialized segment, confirming that the dominant group is cleanly separated from the rest.

4.4.3. Feature Importance

The Gini-based feature-importance scores produced by the Random Forest are reported in Table 4.7. The ranking places savings balance, outstanding loan, collateral value, and credit-card spending at the top, while the activity-related count variables (active days and transaction entries) contribute least. This ordering corroborates the Z-score analysis of Section 4.3.2 from an entirely independent, supervised perspective and provides a direct, quantitative answer to RQ2.

Rank	Feature	Importance
1	SDBQ_TIET_KIEM (savings balance)	0.184
2	DU_NO_EIB (outstanding loan)	0.171
3	GIATRI_TSDB_VND (collateral value)	0.149
4	DS_THE_TD_BQ (credit-card spending)	0.142
5	SDBQ_CASA (CASA balance)	0.121
6	SL_PRODUCT_HOLDING (products held)	0.094
7	SO_NGAY_KH_GD (active days)	0.078
8	BUT_TOAN (transaction entries)	0.061

Table 4.7. Random Forest feature importance (Gini importance; values sum to 1.0).

4.4.4. Effect of Feature Scaling

To confirm the preprocessing decision of Section 3.6.4 from a supervised angle, an identical Random Forest was trained on the cluster labels using features derived from each candidate pipeline. As Table 4.8 shows, accuracy rose monotonically from 91.23% on the raw data to 97.64% after log1p, 99.81% after log1p + StandardScaler, and 99.85% after log1p + RobustScaler. The monotonic improvement - and the slight edge of RobustScaler over StandardScaler - provides independent, supervised confirmation that the log1p + RobustScaler representation yields the most learnable, and therefore most cleanly separated, segments. This complements the unsupervised evidence (Hopkins statistic, PCA variance) presented in Section 3.6.4.

Preprocessing pipeline	Random Forest test accuracy
Raw data (no transformation)	91.23%
log1p only	97.64%
log1p + StandardScaler	99.81%
log1p + RobustScaler (selected)	99.85%

Table 4.8. Random Forest accuracy under each preprocessing pipeline.

Taken together, the Random Forest results confirm that the eight segments are not only statistically well separated but also fully learnable and therefore deployable: a single trained classifier can assign any future customer to a segment in milliseconds, without the need to re-run the clustering over the entire population.

4.5. Discussion

The segmentation results yield a coherent and actionable portrait of the bank's customer base. Synthesizing the size distribution (Table 4.3), the Z-score profile (Table 4.4), the held-out validation (Section 4.3.3), and the Random Forest analysis (Section 4.4), the eight segments can be characterized as follows.

Cluster 0 - Mass-market customers (72.9%). The dominant segment scores below average on every behavioural dimension and shows near-zero digital usage and loan activity. These are low-engagement, transactional customers who maintain a basic account but rarely interact further. They represent the bank's largest activation opportunity: even a modest lift in product adoption within this group would translate into substantial absolute volume.

Cluster 1 - Deposit customers (12.5%). This segment is defined by a strongly positive savings Z-score (+1.00) alongside below-average loan and card activity. These

are accumulation-oriented customers with idle balances but no credit relationship - natural targets for investment products, term-deposit upgrades, and cross-selling of credit facilities.

Clusters 2, 3, and 5 - Credit-oriented segments (4.6%, 1.9%, 3.5%). These three segments share elevated loan Z-scores and loan-amount lifts of 6 to 7, but differ in their secondary characteristics: cluster 2 is a pure-credit group with no collateral, cluster 3 is distinguished by a very high collateral Z-score (+1.76), indicating secured borrowers, and cluster 5 combines lending with the highest credit-card spending (+1.43). Together they form the bank's core lending portfolio and warrant differentiated risk monitoring and tailored credit offers.

Clusters 4, 6, and 7 - High-value segments (2.4%, 0.9%, 1.3%). These small but strategically critical segments concentrate the bank's most engaged and profitable relationships. Cluster 6 (premium)s is positive on every single feature, with the highest CASA (+1.58) and very high digital usage (lift 17.32); cluster 4 (multi-product) records the strongest digital-usage lift of 17.65 and high product-holding; and cluster 7 (loyal) shows high savings and CASA with broad product engagement. Although together they account for under 5% of customers, the literature reviewed in Chapter 2 repeatedly notes that such minorities can dominate profitability - underscoring that cluster size alone does not determine strategic value. These segments justify dedicated relationship management and retention investment.

The highly imbalanced size distribution - one dominant segment of basic customers alongside several small, distinctive groups - matches the pattern repeatedly documented in retail-banking segmentation studies reviewed in Chapter 2, lending external credibility to the result. The convergence of four independent lines of evidence (the internal clustering metrics, the held-out lift analysis, the near-perfect stability, and the 99.85% Random Forest reproduction) indicates that the eight segments reflect a genuine structure in the data rather than an artefact of the method or the chosen inputs.

Two caveats temper this interpretation, in keeping with the descriptive nature of the exercise. First, the segments are a modelling construct rather than sharp natural categories: k-means assigns every customer to exactly one centroid, so customers near a boundary are placed in a single group despite sharing features with neighbouring segments, and the labels above describe central tendencies rather than rigid types. Second, the analysis is associational - it characterises how customers differ, not why they came to differ - so the profiles should inform, but not by themselves dictate, causal claims about behaviour. With these qualifications, the segmentation nonetheless provides the bank with an actionable map: a single dominant group representing the largest activation opportunity, distinct credit- and deposit-oriented segments warranting differentiated product and risk strategies, and a small high-value tier that justifies concentrated retention investment.

4.6. Answers to the Research Questions

RQ1 - Which algorithm can actually segment the 631,062 customers, and which is the practical pick? Of the three candidates, MiniBatch K-Means was the only one that would run on the full population. As Section 3.7 set out, Agglomerative Hierarchical Clustering would need a pairwise-distance structure on the order of n-squared - roughly 4x10^11 entries, hundreds of gigabytes - well past the hardware on hand, while DBSCAN either ran out of memory or never finished, and proved touchy about its parameters in this high-dimensional, zero-inflated space. MiniBatch K-Means, by contrast, got through the entire grid search over K in about 3.3 minutes and returned eight clearly readable segments. So it is at once the only feasible option and the most suitable in practice, which bears out the first hypothesis of Section 3.3.13. Of the three candidate algorithms, MiniBatch K-Means was the only one that could be executed on the full population. As established in Section 3.7, Agglomerative Hierarchical Clustering would require a pairwise-distance structure on the order of n-squared - roughly 4x10^11 entries, or hundreds of gigabytes - far beyond the available hardware, while DBSCAN either exhausted memory or failed to terminate and proved highly sensitive to its parameters in this high-dimensional, zero-inflated space. MiniBatch K-Means, by contrast, completed the entire grid search over K in approximately 3.3 minutes and produced eight clearly interpretable segments. It is therefore both the only feasible and the most suitable algorithm in practice, confirming the first hypothesis of Section 3.3.13.

RQ2 - Which variables play the most significant role in differentiating the segments? Two independent analyses converge on the same answer. The Z-score matrix in Table 4.4 shows that the segments separate most strongly along the savings balance, outstanding-loan, collateral value, and credit-card spending axes, and the Random Forest feature-importance ranking (Table 4.7) places exactly these four variables at the top (savings 0.184, loan 0.171, collateral 0.149, card 0.142). Savings cleanly isolates the deposit segment (cluster 1, +1.00), collateral uniquely identifies secured borrowers (cluster 3, +1.76), and current-account balance distinguishes the premium segment (cluster 6, +1.58). The universal flags (transaction, CASA, and basic account), in contrast, are identical across all clusters and carry no discriminative power. This confirms the second hypothesis of Section 3.3.13 - that account balance and savings would be among the most influential attributes - and aligns with the systematic-review evidence in Chapter 2.

RQ3 - Are the resulting segments sufficiently stable and robust to support business strategy? Yes. The stability analysis in Section 4.2 produced a mean Adjusted Rand Index of 0.9999 across re-runs with different random seeds - effectively perfect reproducibility and far above the 0.85 threshold conventionally taken to indicate a stable solution. This robustness is reinforced by the held-out validation in Section 4.3.3, where variables excluded from the clustering nonetheless reproduced the segment hierarchy, and by the Random Forest in Section 4.4, which reproduced the segments with 99.85% accuracy. The segments are therefore stable, statistically robust, and externally coherent

- a sound foundation for differentiated product design, targeted marketing, and segment-aware risk management. Because the labelled dataset assigns every one of the 631,062 customers to a segment, and because the trained Random Forest classifier can extend these labels to new customers without re-running the clustering, the results are directly deployable in an operational setting.

Chapter 5. Determinants of Customer Tenure: An Econometric Analysis

5.1. Motivation and Empirical Strategy

Having identified eight behavioural segments in the previous chapter, this chapter turns from description to explanation. The segmentation answers the question of what kinds of customers the bank has; the present analysis asks a different and economically deeper question: which customer characteristics are associated with a longer banking relationship?

The length of the customer–bank relationship is of direct economic interest. A long-standing literature establishes that relationship duration is a central driver of customer profitability: longer-tenured customers tend to generate higher and more stable revenues, cost less to serve, and are more receptive to cross-selling (Reichheld & Teal, 1996; Reinartz & Kumar, 2000, 2003). Understanding which attributes predict longer tenure therefore allows the bank to identify, at an early stage, the customers most likely to develop into durable and valuable relationships, and to design retention strategies accordingly.

The empirical strategy proceeds in three steps. First, customer tenure is modelled as a function of demographic and behavioural attributes using ordinary least squares (OLS). Second, the explanatory variables are explicitly classified according to whether they can be treated as predetermined (exogenous) or as potentially endogenous, and the model is estimated under a sequence of specifications of increasing richness (a sensitivity analysis). Third, the threat of reverse causality is discussed openly, and the interpretation of the coefficients is calibrated to what the cross-sectional research design can credibly support.

5.2. Dependent Variable

The dependent variable is customer tenure, measured by SO_NAM_MO_CIF - the number of years since the customer's CIF (Customer Information File) was opened. This variable is a standard operationalisation of behavioural loyalty in the banking literature, where the duration of the relationship is treated as a revealed-preference measure of loyalty that does not rely on self-reported attitudes (Coulter & Coulter, 2002; Cooil et al., 2007).

Because tenure is bounded below at zero and is typically right-skewed - a large mass of recently acquired customers and a thin tail of very long-standing ones - the natural logarithm of tenure, ln (1 + SO_NAM_MO_CIF), is used as the dependent variable. The log transformation reduces skewness, stabilises the variance of the residuals, and yields coefficients that can be interpreted approximately as semi-elasticities.

Table 5.1 - Summary statistics of the dependent variable (customer tenure, in years).

Statistic	Tenure (years)	ln(1 + Tenure)
Mean	7.57	1.91
Median	6.40	2.00
Std. Dev.	5.70	0.73
Min	0.61	0.47
Max	35.94	3.61
Skewness	0.80	-0.15
N	631,062	631,062

The log transformation reduces right-skew from 0.80 to -0.15, supporting ln(1 + tenure) as the modelled outcome.

5.3. Explanatory Variables

The choice of explanatory variables is guided by the customer-relationship and household-finance literature, which identifies demographic position, behavioural profile, and financial engagement as the three channels through which relationship duration is determined (Reichheld & Teal, 1996; Reinartz & Kumar, 2003; Cooil et al., 2007). The variables are therefore organised into three blocks, deliberately ordered by their vulnerability to endogeneity. This ordering mirrors the substantive logic of the model and supports the sensitivity analysis in Section 5.5, in which each block is introduced sequentially. All distributions below are computed on the full estimation sample of 631,062 active retail customers.

Block 1 - Predetermined demographic attributes. These characteristics are, for the most part, determined before the banking relationship begins and cannot plausibly be caused by the customer's loyalty to the bank; this temporal ordering makes the block the most credible from a causal standpoint (Antonakis et al., 2014).

Age band (NHOM_TUOI). Age is among the most consistently documented correlates of banking-relationship length: older customers have had more time to form an attachment, hold more stable financial lives, and exhibit lower switching propensity (Reinartz & Kumar, 2000). Because age is determined prior to the banking relationship, it is treated as strictly predetermined. It is entered as the bank's internal life-stage bands to allow for the non-linear life-cycle pattern commonly found in the literature; the senior and middle-aged bands together account for roughly two-thirds of customers.

Table 5.2 - Distribution of customers by age band.

Age band	Count	Percentage
Young adult (Thành niên)	105,133	16.66%
Middle-aged (Trung niên)	188,409	29.86%
Senior (Cao niên)	237,857	37.69%
Retirement age (Tuổi hưu)	96,426	15.28%
Other / unclassified (Khác)	3,237	0.51%
Total	631,062	100.00%

Bands follow the bank's internal age-segmentation definitions; the senior band (the largest group) is the reference category.

Gender (GIOI_TINH). Gender is included because a sizeable literature documents systematic differences between men and women in financial behaviour, risk attitudes, and product engagement (Bucher-Koenen et al., 2017). It is predetermined relative to the banking relationship; the sample is almost evenly split with a marginal female majority.

Table 5.3 - Distribution of customers by gender.

Gender	Count	Percentage
Female (Nữ)	316,575	50.17%
Male (Nam)	314,487	49.83%
Total	631,062	100.00%

Reference category: male. The coefficient reported in Table 5.10 is therefore for female customers, relative to male.

Occupation sector (NGHE_NGHIEP). Occupation proxies for income stability, transaction needs, and the degree to which a customer's livelihood is intertwined with banking services; commerce, services, and self-employed activity in particular rely more heavily and continuously on banking infrastructure, which the literature links to deeper and more durable institutional relationships. Two recorded categories with negligible counts (industry and construction, 21 customers; an administrative "migration" code, 7 customers) are merged into the residual "Other" category to avoid unstable dummy estimates.

Table 5.4 - Distribution of customers by occupation sector.

Occupation sector	Count	Percentage
Commerce and services (Thương mại và dịch vụ)	265,637	42.09%
Agriculture, forestry and fishing (Nông, lâm nghiệp và thủy sản)	231,446	36.68%
Other (Khác)	133,979	21.23%
Total	631,062	100.00%

Reference category: the residual "Other" group. Commerce-and-services and agriculture/forestry/fishing are each reported relative to this baseline in Table 5.10.

Region (KHU_VUC). Region captures persistent geographic differences in financial development, branch density, and economic activity that shape both the supply of and demand for banking relationships, controlling for differences in market structure across the bank's national footprint. Customers are concentrated in the Greater Ho Chi Minh City and Southeast regions, consistent with the bank's commercial footprint.

Table 5.5 - Distribution of customers by region.

Region	Count	Percentage
East HCMC (KV Đông HCM)	110,911	17.58%
Southeast (KV Miền Đông Nam Bộ)	95,880	15.19%
Central & Central Highlands (KV Miền Trung, Tây Nguyên)	93,292	14.78%
Northwest HCMC (KV Tây Bắc HCM)	74,891	11.87%
Mekong Delta (KV Miền Tây Nam Bộ)	70,246	11.13%
Southwest HCMC (KV Tây Nam HCM)	67,073	10.63%
North Red River (KV Bắc Sông Hồng)	59,401	9.41%
South Red River (KV Nam Sông Hồng)	59,368	9.41%
Total	631,062	100.00%

Reference category: North Red River (KV Bắc Sông Hồng). All other regions are reported relative to this baseline in Table 5.10; East HCMC shows the longest tenure among them.

Marital status (excluded). Marital status was initially considered but excluded from the final model: 99.7% of customers share the same recorded value, leaving no usable variation for identification.

Block 2 - Behavioural-segment membership. The cluster label from the K-Means analysis of Chapter 4 enters the model as a set of indicator variables, summarising each customer's multidimensional behavioural profile in a single categorical control and testing whether the segments differ in tenure once demographic position is held constant. The distribution is highly concentrated - the dominant mass-market segment (Cluster 0) contains roughly 73% of customers - which motivates using it as the reference category.

Table 5.6 - Distribution of customers by behavioural segment.

Behavioural segment	Count	Percentage
Cluster 0	460,134	72.91%
Cluster 1	78,768	12.48%
Cluster 2	29,264	4.64%
Cluster 5	21,888	3.47%
Cluster 4	15,139	2.40%
Cluster 3	11,725	1.86%
Cluster 7	8,239	1.31%
Cluster 6	5,905	0.94%
Total	631,062	100.00%

Reference category: Cluster 0 (mass-market).

Block 3 - Financial-behaviour attributes. These three financial measures are the most informative regressors but also the most vulnerable to reverse causality, since they may be outcomes of a long relationship as much as causes of it; they are therefore introduced last and interpreted with the greatest caution (Section 5.6).

Table 5.7 reports the descriptive statistics of the dependent variable and the continuous explanatory variables. Three features of the data are worth highlighting, as they motivate the variable transformations adopted in the model. First, the dataset is complete: no variable contains any missing values, reflecting its origin as a direct extraction from the bank's system of record. Second, the monetary variables are extremely right-skewed: for the CASA, savings, outstanding-loan, and collateral variables the mean greatly exceeds the median, and for savings, loans, and collateral the median is exactly zero, indicating that more than half of customers hold none of these

products. This pronounced skew and zero-inflation is the reason these variables are log-transformed (and, for CASA, winsorised at the 99th percentile). Third, the average card-spending variable takes a small number of negative values (minimum -2.12 million VND), which correspond to customers carrying a negative average card balance over the period; these are genuine values rather than data errors and are retained.

Variable	Mean	Median	Min	Max	Std. Dev.	Unit
Tenure	7.57	6.40	0.61	35.94	5.70	years
Avg. CASA balance	14.91	0.78	0.00	48,263.18	160.85	million VND
Avg. savings balance	166.09	0.00	0.00	863,500.37	2,816.51	million VND
Outstanding loan	158.61	0.00	0.00	567,329.32	2,056.53	million VND
Avg. card spending	0.66	0.00	-2.12	578.04	5.97	million VND
Collateral value	166.48	0.00	0.00	956,711.73	2,690.04	million VND
Number of products	1.88	2.00	1.00	9.00	1.04	count
Active transaction days	51.03	22.00	0.00	253.00	58.36	count
Number of transactions	431.25	35.00	0.00	528,216.00	5,522.58	count

Table 5.7. Descriptive statistics of the dependent and explanatory variables (n = 631,062).

Notes: monetary variables are expressed in million VND. No variable contains missing values (0% missing across all fields). The four categorical predictors used in the model - age band (5 groups), gender (2), occupation (5), and region (8) - are likewise complete, with no missing values.

Average CASA balance (SDBQ_CASA) proxies' transactional liquidity and the intensity of the operating relationship (Rust, Lemon & Zeithaml, 2004). It is extremely right-skewed (skewness ≈ 133; mean 14.9 million VND against a median of only 0.78 million), so it is winsorised at the 99th percentile and log-transformed.

Average savings balance (SDBQ_TIET_KIEM) captures accumulation behaviour. It is strongly zero-inflated - approximately 83% of customers hold no savings balance, so its median is zero - and equally skewed (skewness ≈ 136); it is entered as the log balance conditional on holding savings (zero otherwise), preserving the has / does-not-have distinction.

Number of products held (SL_PRODUCT_HOLDING) is the conventional proxy for the breadth of the relationship and for cross-selling success (Reinartz, Thomas & Bascoul, 2008). Holdings range from 1 to 9 with a mean of 1.88, indicating a shallow product relationship for most customers, and this variable is the most exposed to reverse causality.

Table 5.8 - Descriptive statistics of continuous explanatory variables.

Variable	Mean	Median	Std. Dev.	Min	Max	Skewness
Average CASA balance (VND)	14,913,814	784,502	160,852,275	0	48,263,176,719	132.95
Average savings balance (VND)	166,090,930	0	2,816,507,772	0	863,500,365,428	135.58
Number of products held	1.88	2.00	1.04	1	9	1.34

5.4. Descriptive Patterns

Before turning to the regression model, this section presents a set of simple descriptive graphs to convey the intuition behind the analysis. All figures are produced directly from the study's own data; they are not reproduced from external sources. They illustrate how tenure and segment membership vary across the key demographic dimensions, motivating the choice of explanatory variables in the model that follows.

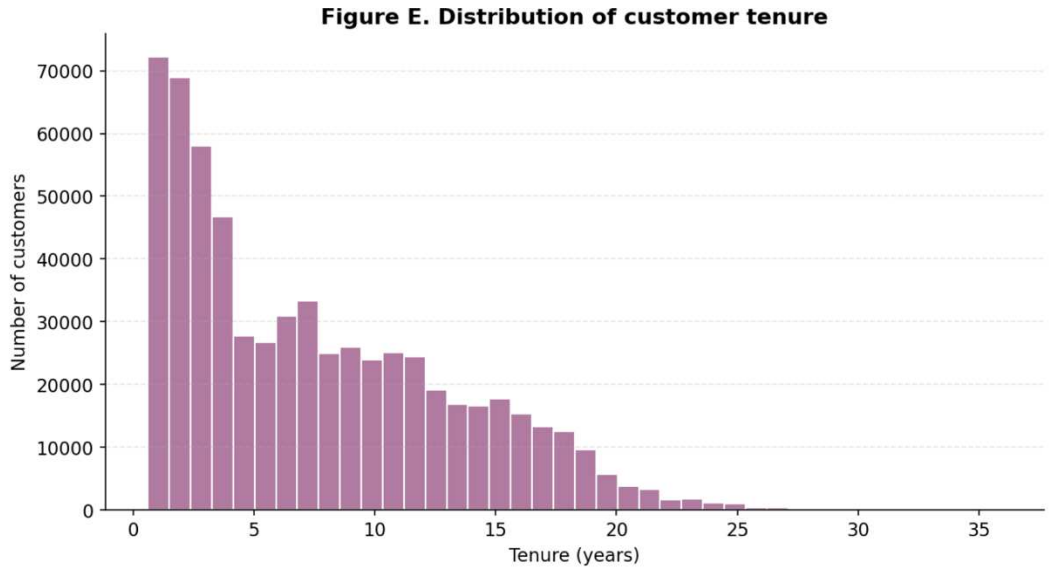


Figure 5.1. Distribution of customer tenure across the population (n = 631,062).

The distribution of tenure (Figure 5.1) is right-skewed, with a large mass of relatively recent customers and a long tail of long-standing relationships - the pattern that motivates the log transformation of the dependent variable.

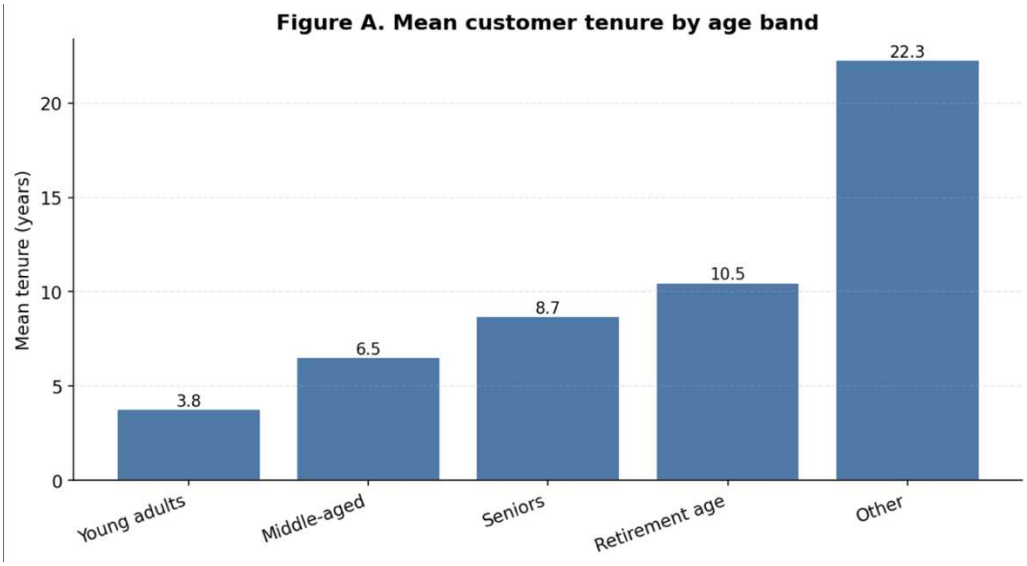


Figure 5.2. Mean customer tenure by age band.

Mean tenure rises steadily across the age bands (Figure 5.2), consistent with older customers having had more time to form, and sustain, a banking relationship.

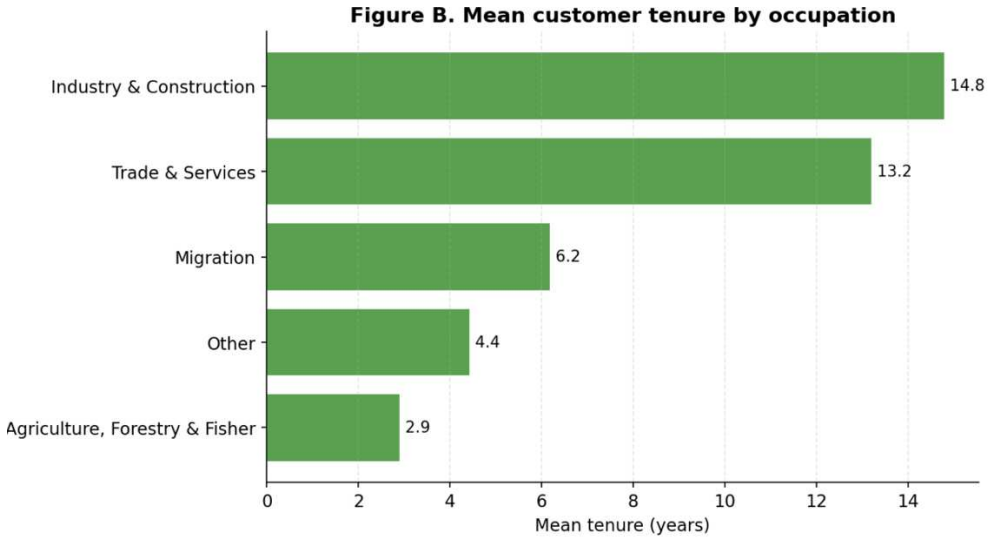


Figure 5.3. Mean customer tenure by occupation.

The clearest pattern appears across occupations (Figure 5.3): commerce-and-services customers exhibit markedly longer average tenure than other occupational groups, foreshadowing the dominant role of occupation in the regression results.

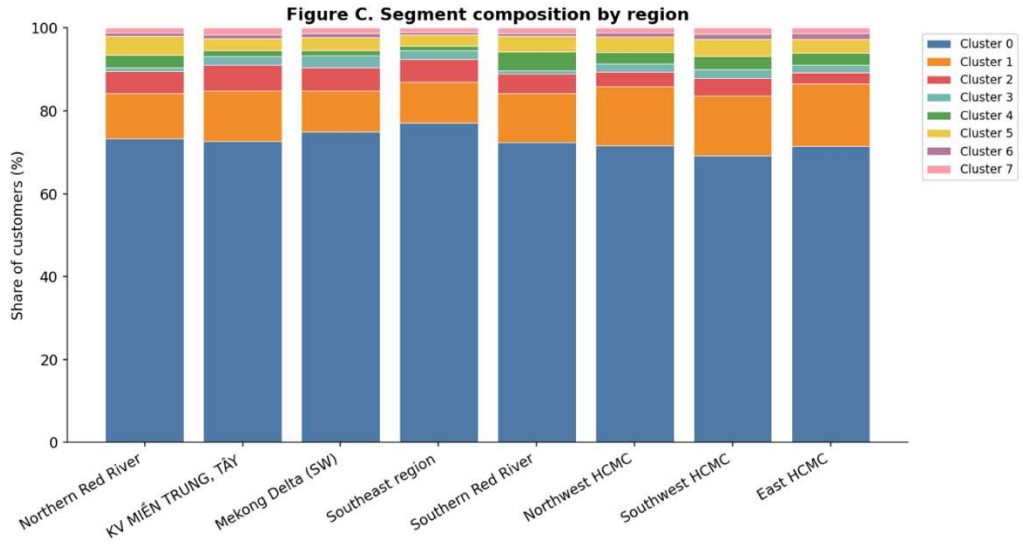


Figure 5.4. Segment composition by region.

Segment composition also varies across regions (Figure 5.4), indicating that the behavioural segments identified in Chapter 4 are not uniformly distributed geographically.

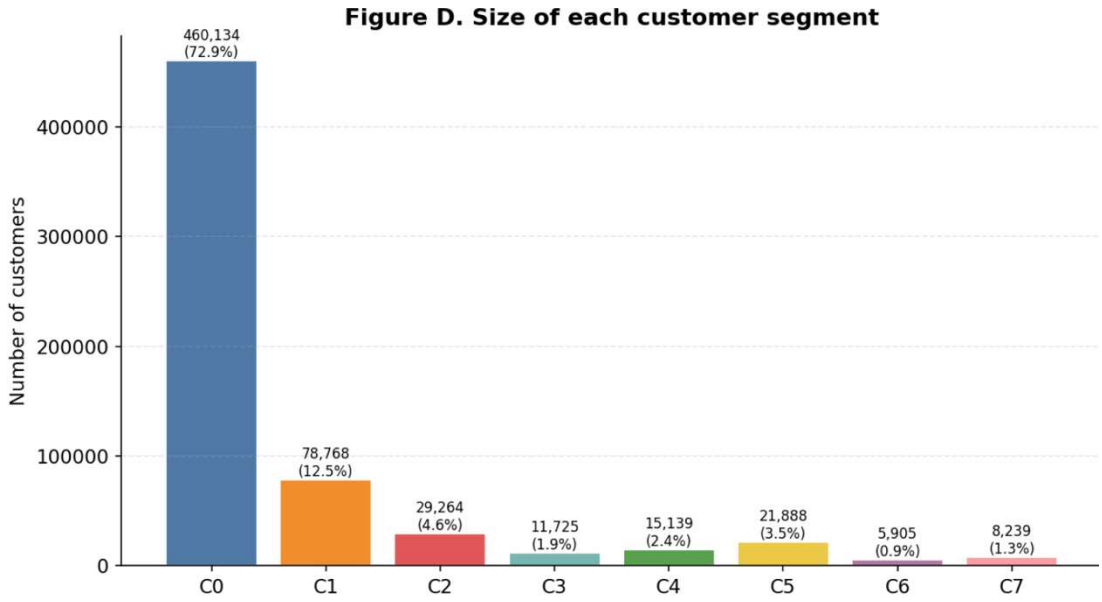


Figure 5.5. Size of each customer segment.

Figure 5.5 shows the relative size of the eight segments, confirming that the mass-market segment dominates while the premium segment, though small, is a distinct group.

The remaining figures examine the segment and financial variables, confirming that each is associated with tenure and therefore merits inclusion in the model.

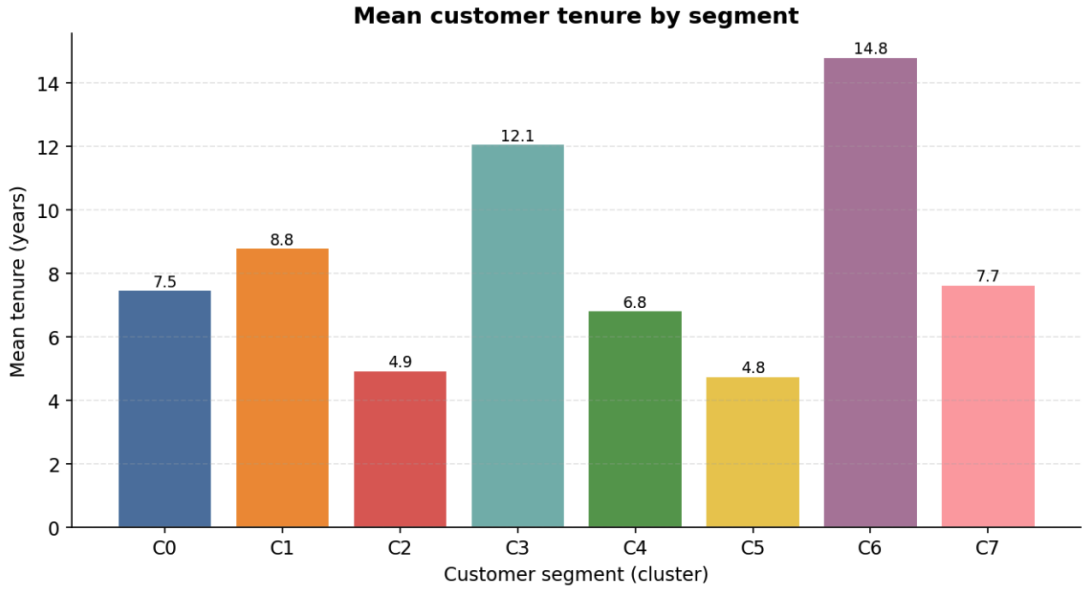


Figure 5.6. Mean customer tenure by behavioural segment.

Average tenure varies sharply across the eight behavioural segments (Figure 5.6), ranging from 4.8 years in the most transactional segments (Clusters 2 and 5) to 14.8 years in the premium segment (Cluster 6) - a near three-fold difference. This wide dispersion confirms that segment membership carries substantial information about relationship duration beyond the demographic variables, justifying its inclusion as a block of indicators; it also previews the central result that the premium segment is the most durable of all.

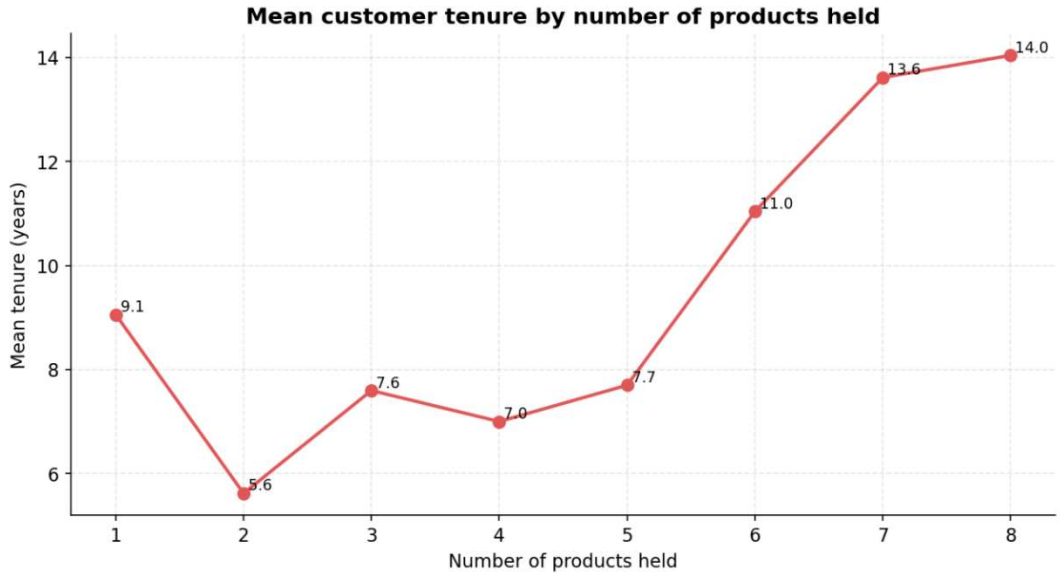


Figure 5.7. Mean customer tenure by number of products held.

Figure 5.7 is the most revealing of the descriptive plots. The relationship between the number of products held and mean tenure is distinctly U-shaped: average tenure falls from 9.1 years for single-product customers to a trough of 5.6 years at two products and then rises steadily to 11.0 years at six products and 14.0 years at eight. A purely linear specification would average these opposing slopes into a single, misleading coefficient; the clear curvature visible here is precisely what motivates the inclusion of a squared product term in the model, and it anticipates the U-shaped estimate recovered in Section 5.8. Substantively, the descending portion is consistent with intensive cross-selling to newly acquired customers, while the ascending portion reflects the deep, multi-product relationships built up by long-standing customers.

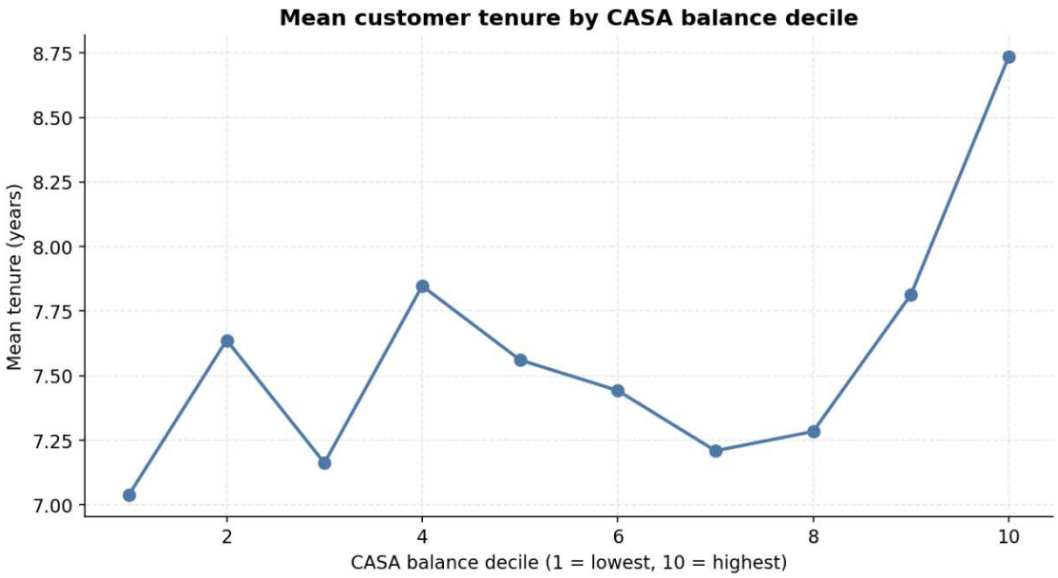


Figure 5.8. Mean customer tenure by CASA balance decile.

Tenure is relatively flat across the lower and middle deciles of the CASA balance distribution but rises markedly in the highest decile, from around 7.0–7.6 years to 8.7 years (Figure 5.8). This convex pattern - a strengthening association at high balance levels - supports both the inclusion of the CASA variable and the convex (positive squared) term estimated in the model.

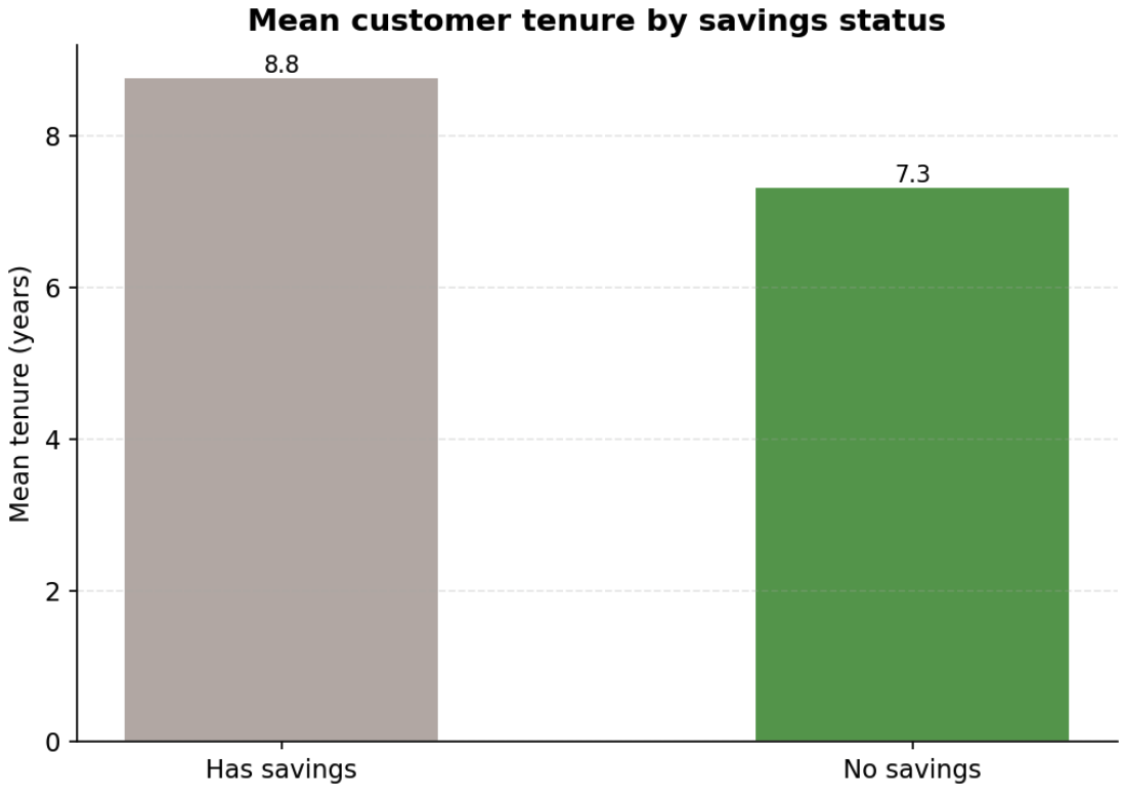


Figure 5.9. Mean customer tenure by savings status.

Finally, customers who hold a savings balance bank for 8.8 years on average, compared with 7.3 years for those who do not (Figure 5.9), indicating that the deeper financial commitment represented by a savings relationship is associated with greater

longevity and justifying the inclusion of the savings variable. Taken together, these descriptive patterns show that every block of the model - demographic, behavioural-segment, and financial - is systematically related to tenure, and they motivate the formal specification developed in the next section.

5.5. Model Specification

The baseline regression models log-tenure as a function of demographic attributes only (Model 1). It is then extended sequentially by adding the segment indicators (Model 2) and finally the financial-behaviour block (Model 3). Categorical variables enter as dummy variables with one reference category each, and continuous financial variables are log-transformed to match the treatment of the dependent variable.

Formally, the empirical model for customer i is specified as follows:

Model 1 (demographic block only):

$$ln(1 + Tenure_i) = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Occupation_i + \beta_4 Region_i + \varepsilon_i \quad (1)$$

Model 2 (adding behavioural-segment membership):

$$ln(1 + Tenure_i) = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Occupation_i + \beta_4 Region_i + \beta_5 Segment_i + \varepsilon_i \quad (2)$$

Model 3 (adding the financial-behaviour block, including the quadratic terms):

$$ln(1 + Tenure_i) = \beta_0 + \beta_1 Age_i + \beta_2 Gender_i + \beta_3 Occupation_i + \beta_4 Region_i + \beta_5 Segment_i + \beta_6 ln(CASA_i) + \beta_7 ln(CASA_i)^2 + \beta_8 ln(Savings_i) + \beta_9 ln(Products_i) + \beta_{10} ln(Products_i)^2 + \varepsilon_i \quad (3)$$

where the dependent variable is the natural logarithm of one plus customer tenure (in years); Age, Gender, Occupation, Region, and Segment each enter as a set of dummy variables with one reference category; CASA, Savings, and Products are log-transformed continuous variables; the CASA and Products terms additionally enter in squared form to allow for a non-linear relationship with tenure; and ε_i is the error term. The continuous variables are mean-centred before squaring, which reduces the collinearity between each variable and its square. A squared savings term was also tested but excluded, because savings is strongly zero-inflated (more than four-fifths of customers hold no savings balance), which made the savings and savings-squared terms severely collinear (variance inflation factors above 30); the CASA and Products squared terms, by contrast, are well-behaved (all variance inflation factors below 2.3). All models are estimated by ordinary least squares with heteroskedasticity-robust (HC1) standard errors. As equations (1) to (3) make explicit, the three specifications are nested: Models 1 and 2 are obtained from the full model by restricting the financial-block coefficients (β_6 through β_{10}) to zero, and Model 1 additionally restricts the segment

coefficient (β_5) to zero. The squared terms therefore enter only in the final, unrestricted specification (Model 3).

Age is deliberately specified as a set of bands rather than as a single continuous regressor. Because the dataset records age in grouped form rather than as an exact figure, a continuous quadratic term (Age^2) cannot be constructed; the banded specification instead captures non-linearity in the age-tenure relationship flexibly, without imposing the specific parabolic shape that a quadratic term would assume.

Estimating the three models in sequence serves as a sensitivity analysis. If the coefficients on the predetermined demographic variables remain stable in sign, magnitude, and significance as the segment and financial blocks are added, this stability provides reassurance that those estimates are not artefacts of a particular specification (Antonakis et al., 2014). Conversely, large swings in a coefficient when additional controls enter would signal that it was previously absorbing omitted-variable variation and should be interpreted cautiously.

5.6. Addressing Endogeneity

The principal threat to a causal reading of the estimates is reverse causality, a recognised form of endogeneity in which the presumed direction of causation may be reversed (Antonakis et al., 2014). The concern is concentrated in Block 3. Consider the relationship between the number of products held and tenure. A positive association is consistent with two distinct stories: customers who hold more products may become more deeply embedded with the bank and therefore stay longer (products → tenure), but equally, customers who have stayed longer have simply had more time to accumulate products (tenure → products). With a single cross-section of data, these two channels cannot be separated, and the coefficient on product holding conflates them.

Three measures are taken to address this concern within the limits of the available data. First, the variables most exposed to reverse causality are isolated in a separate block and introduced only in the final specification, so that their inclusion does not contaminate the more credible demographic estimates. Second, the temporal logic of the demographic block is made explicit: because attributes such as age, gender, and occupation are determined prior to, and independently of, the banking relationship, reverse causality from tenure to these variables is not plausible, and their coefficients can be read as conditional associations with a credible causal interpretation. Third, the limitations of the design are stated transparently: the coefficients on the financial block are interpreted as conditional correlations, not as causal effects, and no policy claim is built on them alone.

A fully causal treatment of the endogenous regressors would require either panel data, which would allow lagged regressors to break the simultaneity (Leszczensky & Wolbring, 2022), or a valid instrumental variable that shifts financial behaviour without independently affecting tenure (Antonakis et al., 2014). The present dataset is a single cross-section and contains no variable that credibly satisfies the exclusion restriction

required of an instrument. Rather than adopt a weak or indefensible instrument, which would yield estimates more misleading than transparent OLS, this study reports OLS results with an explicit endogeneity discussion and identifies the instrumental-variable and panel approaches as priorities for future research.

5.7. Results

The dependent variable, customer tenure, ranges from approximately 0.6 to 35.9 years, with a mean of 7.6 years and a median of 6.4 years. The raw distribution is moderately right-skewed (skewness = 0.80); the log transformation reduces this to a near-symmetric distribution (skewness = -0.15), confirming the appropriateness of modelling ln (1 + tenure). The estimation sample retains all 631,062 customers, as no observations were lost to missing data on the modelled variables.

The demographic block alone explains a remarkably high share of the variation in tenure ($R^2 = 0.727$), which rises to 0.729 when segment indicators are added (Model 2) and to 0.745 when the financial block enters (Model 3), as summarised in Table 5.9. The financial block therefore contributes incremental explanatory power beyond the segments. Crucially, the coefficients on the demographic variables are highly stable across all three specifications: the coefficient on the commerce-and-services occupation moves only from 0.989 (Model 1) to 0.949 (Model 3), and the age and region coefficients shift only at the third decimal place. This stability is the central robustness result of the analysis.

Specification	Variables added	R^2	Adjusted R^2	AIC
Model 1	Demographics	0.7272	0.7272	566,790
Model 2	+ Segment	0.7287	0.7286	563,379
Model 3	+ Financial	0.7448	0.7448	524,729

Table 5.9. Model fit across the three specifications.

The full set of estimated coefficients for all three specifications is reported in Table 5.10, which presents the three nested linear models; the quadratic extension of the financial block is reported separately in Table 5.11 (Section 5.8). Occupation is the single strongest determinant of tenure. Relative to the reference category, customers in commerce and services have a coefficient of 0.949, implying a tenure roughly two-and-a-half times longer than the baseline group ($e^{0.949} \approx 2.58$), whereas customers in agriculture, forestry, and fishing have a coefficient of -0.262, corresponding to materially shorter relationships. This is the most pronounced effect in the entire model and points to occupation-linked financial behaviour as a primary driver of relationship longevity.

Age displays a coherent monotonic pattern. Taking the senior band as the reference, younger customers have the shortest tenure (-0.122), middle-aged customers a smaller negative coefficient (-0.044), and retirement-age customers a small positive coefficient (0.010). The ordering is intuitive and can be read with a credible causal interpretation, since age is determined independently of, and prior to, the banking relationship. Gender has a small but precisely estimated effect, with female customers exhibiting marginally longer tenure (0.020). Region effects are modest but systematic, with customers in the East Ho Chi Minh City area having the longest tenure (0.147 relative to the reference region).

After conditioning on demographics, the segment indicators reveal that the premium segment (Cluster 6) has the longest tenure of all groups (0.272), followed by the credit-oriented segments (Clusters 2, 5, and 3, with coefficients of 0.219, 0.178, and 0.172 respectively). The finding that the premium segment is the most durable is economically coherent: the customers who are most engaged and valuable are also those who remain longest, reinforcing the strategic importance of that small segment identified in Chapter 4. Within the financial block, both the average CASA balance (0.016) and the savings balance among savers (0.012) are positively and significantly associated with tenure.

It is worth making explicit how the parameters are read, since the dependent variable is in logarithms. For a continuous regressor, a coefficient β is a semi-elasticity: a one-unit increase in the regressor is associated with an approximately $100\times\beta$ percent change in tenure. Because the financial regressors are themselves logged, their coefficients are elasticities - a one-percent increase in the balance is associated with a β-percent change in tenure. For a dummy variable, the multiplicative effect on tenure is e^{β}: the value 0.949 on commerce-and-services thus corresponds to $e^{0.949} \approx 2.58$, a relationship roughly 158 percent longer than the baseline, while the -0.262 on agriculture corresponds to $e^{-0.262} \approx 0.77$, or about 23 percent shorter. Evaluated at the sample mean tenure of 7.6 years, these translate into concrete differences of several years, which is why occupation dominates the economic interpretation.

The financial coefficients, though smaller, are economically meaningful once their scale is considered. The CASA elasticity of 0.016 implies that a doubling of the average current-account balance (an increase of about 0.69 in its natural log) is associated with roughly 1.1 percent longer tenure; moving across the full interquartile range of log-CASA accumulates into a difference of several months. The savings coefficient of 0.012 is interpreted only among the minority of customers who actually hold savings, and the descriptive evidence in Figure 5.9 - savers banking 8.8 years on average against 7.3 for non-savers - shows that the binary act of holding any savings is associated with a larger gap (about 1.5 years) than the marginal balance effect, consistent with the strongly zero-inflated nature of the variable. All financial coefficients are estimated very precisely (standard errors at the third or fourth decimal place) because of the large sample, so statistical significance is near-universal; the substantive question is therefore one of

magnitude rather than significance, and on that criterion the demographic and segment effects clearly dominate the financial ones.

The number of products held, by contrast, carries a negative coefficient (-0.522), which runs counter to the conventional expectation that broader product holding deepens the relationship. A diagnostic check confirms this is not an artefact of collinearity with the segment indicators: when the segment dummies are removed (Model 3b), the coefficient remains negative (-0.350), so the sign is a genuine feature of the data. Two non-mutually exclusive interpretations are plausible. First, reverse causality: the bank may concentrate cross-selling on newly acquired customers during onboarding, so that shorter-tenured customers mechanically hold more products. Second, a genuine behavioural effect: long-standing customers may have consolidated around a few core products, whereas newer customers open many products without having yet formed a durable attachment. This counter-theoretical sign illustrates precisely why the financial-block coefficients are interpreted throughout as conditional correlations rather than causal effects. As shown in Section 5.8, however, this linear coefficient masks a non-linear, U-shaped relationship: once a squared term is added, the association turns positive at high levels of product holding, reconciling these two interpretations.

Dependent variable: ln (1 + tenure)	**Model 1**	**Model 2**	**Model 3**
Intercept	1.541*** (0.002)	1.551*** (0.002)	1.794*** (0.004)
Age: young adult	-0.151*** (0.002)	-0.154*** (0.002)	-0.122*** (0.002)
Age: middle-aged	-0.061*** (0.001)	-0.062*** (0.001)	-0.044*** (0.001)
Age: retirement	0.041*** (0.001)	0.038*** (0.001)	0.010*** (0.001)
Age: other	0.499*** (0.004)	0.491*** (0.004)	0.437*** (0.004)
Gender: female	0.024*** (0.001)	0.025*** (0.001)	0.020*** (0.001)
Occupation: agriculture/forestry/fishing	-0.269*** (0.002)	-0.268*** (0.002)	-0.262*** (0.002)
Occupation: commerce and services	0.989*** (0.002)	0.987*** (0.002)	0.949*** (0.002)
Region: Central & Highlands	0.103*** (0.002)	0.100*** (0.002)	0.085*** (0.002)
Region: Mekong Delta	0.040*** (0.002)	0.039*** (0.002)	0.030*** (0.002)
Region: Southeast	0.047*** (0.002)	0.045*** (0.002)	0.042*** (0.002)
Region: South Red River	-0.022*** (0.002)	-0.022*** (0.002)	-0.024*** (0.002)
Region: Northwest HCMC	0.078*** (0.002)	0.078*** (0.002)	0.068*** (0.002)
Region: Southwest HCMC	0.076*** (0.002)	0.076*** (0.002)	0.069*** (0.002)
Region: East HCMC	0.156*** (0.002)	0.155*** (0.002)	0.147*** (0.002)
Segment: Cluster 1	-	-0.009*** (0.002)	-0.011 (0.010)
Segment: Cluster 2	-	0.007** (0.002)	0.219*** (0.003)
Segment: Cluster 3	-	-0.083*** (0.003)	0.172*** (0.003)

Segment: Cluster 4	-	-0.070*** (0.003)	0.091*** (0.010)
Segment: Cluster 5	-	-0.121*** (0.003)	0.178*** (0.003)
Segment: Cluster 6 (premium)	-	0.088*** (0.004)	0.272*** (0.011)
Segment: Cluster 7	-	-0.012*** (0.004)	0.108*** (0.011)
ln(CASA balance)	-	-	0.016*** (0.000)
ln(savings balance)	-	-	0.012*** (0.001)
ln(number of products)	-	-	-0.522*** (0.003)
Observations	**631,062**	**631,062**	**631,062**
R-squared	**0.7272**	**0.7287**	**0.7448**
Adjusted R-squared	**0.7272**	**0.7286**	**0.7448**

Table 5.10. OLS regression of log customer tenure (robust HC1 standard errors in parentheses).

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Heteroskedasticity-robust (HC1) standard errors in parentheses. Reference categories: senior age band, male, administrative/other occupation, reference region, and Cluster 0. "-" indicates the variable is not included in that specification.

5.8. Diagnostic Tests

The model satisfies the standard diagnostic checks. After removing a redundant savings indicator that was near-collinear with the continuous savings measure (initial VIFs of approximately 57), all variance inflation factors fall below two (ln CASA = 1.43, ln savings = 1.48, ln products = 1.86), well under the conventional threshold of ten, and the condition number falls to a moderate 619. The residuals are approximately symmetric (skewness = -0.255) and, given the very large sample, the central limit theorem ensures valid inference notwithstanding mild non-normality. All standard errors are heteroskedasticity-robust (HC1).

A Ramsey RESET test returns a statistically significant result ($p < 0.05$); however, this must be interpreted in the context of the very large sample size (631,062 observations), for which virtually any hypothesis test attains significance because even economically negligible departures become statistically detectable. Given that the linear

specification explains 74.5% of the variation in tenure and yields economically interpretable coefficients, the functional form is judged adequate. A Breusch-Pagan test likewise indicates heteroskedasticity, which is accommodated through the robust standard errors already employed. The condition that cannot be tested statistically - the absence of unobserved confounders (selection on observables) - is acknowledged as the central limitation: while the demographic coefficients are large enough to be robust to plausible omitted variables, the financial-block coefficients are interpreted cautiously as conditional correlations.

On the supervisor's suggestion, a possible non-linear link between the financial variables and tenure was tested directly, by adding squared terms for the (mean-centred) CASA and Products variables. Table 5.11 reports the linear-versus-quadratic comparison. The squared terms are jointly very significant ($F = 5{,}497$, $p < 0.001$) and the Akaike Information Criterion drops a long way, from 524,729 to 512,967 - a sign the quadratic spec genuinely fits better. The R^2 gain is modest all the same (0.745 to 0.749), so the non-linearity refines the linear results rather than upending them. Every variance inflation factor in the quadratic model stays under 2.3, which confirms the curvature is pinned down cleanly once the collinear savings-squared term is dropped.

	Linear model	**Quadratic model**
R-squared	0.7448	0.7495
Adjusted R-squared	0.7448	0.7495
AIC	524,729	512,967
BIC	525,013	513,274
ln (CASA)	0.0163*** (0.0003)	0.0144*** (0.0003)
ln (CASA)²	-	0.0035*** (0.0001)
ln (Savings)	0.0122*** (0.0005)	0.0081*** (0.0005)
ln (Products)	-0.5215*** (0.0029)	-0.5592*** (0.0029)
ln (Products)²	-	0.4953*** (0.0053)
Max VIF (financial)	1.86	2.25

Table 5.11. Linear versus quadratic specification of the financial block (robust HC1; full model, n = 631,062).

Notes: *** $p < 0.01$. Heteroskedasticity-robust (HC1) standard errors in parentheses. Continuous variables are mean-centred. A savings-squared term was

tested but excluded due to severe collinearity (VIF > 30). The two squared terms are jointly significant ($F = 5{,}497$, $p < 0.001$).

The most striking feature of the quadratic model is the U-shaped relationship between product holding and tenure. The linear coefficient on the number of products is negative (-0.559) while the squared coefficient is positive (0.495), implying that tenure first falls and then rises as the number of products increases. Substantively, this reconciles the two competing interpretations of the negative linear coefficient noted earlier: among customers holding only a few products, holding more is associated with shorter tenure - consistent with intensive cross-selling to newly acquired customers - whereas among customers who already hold many products, holding still more is associated with longer tenure, consistent with deeply embedded, long-standing relationships. The relationship is therefore not monotonic, which is precisely why the purely linear specification, with its single negative coefficient, understated the loyalty of the most product-intensive customers. The CASA balance shows a milder convex pattern (both its linear and squared terms are positive), indicating that the positive association between liquid balances and tenure strengthens at higher balance levels.

The turning point of this U-shape can be located precisely. For a quadratic in a mean-centred regressor, tenure is minimised where the derivative is zero, that is at a centred value of $-\beta_1/(2\beta_2) = -(-0.559)/(2\times0.495) \approx 0.56$ above the mean of log-products, which corresponds to roughly two products held - exactly the trough visible in the descriptive Figure 5.7, where mean tenure bottoms out at two products (5.6 years) before rising to 14.0 years at eight. The descriptive pattern and the estimated parameters therefore agree both qualitatively and quantitatively, which is reassuring: the quadratic term is not merely improving fit mechanically but is capturing a real, interpretable feature of the data. The same calculation for CASA places the (shallow) turning point below the observed range, so that over the relevant region the CASA effect is monotonically increasing and simply convex, consistent with Figure 5.8.

The regression assumptions are examined visually in Figure 5.10, which presents four standard diagnostic plots for the quadratic specification, generated by the author from the residuals of the fitted model.

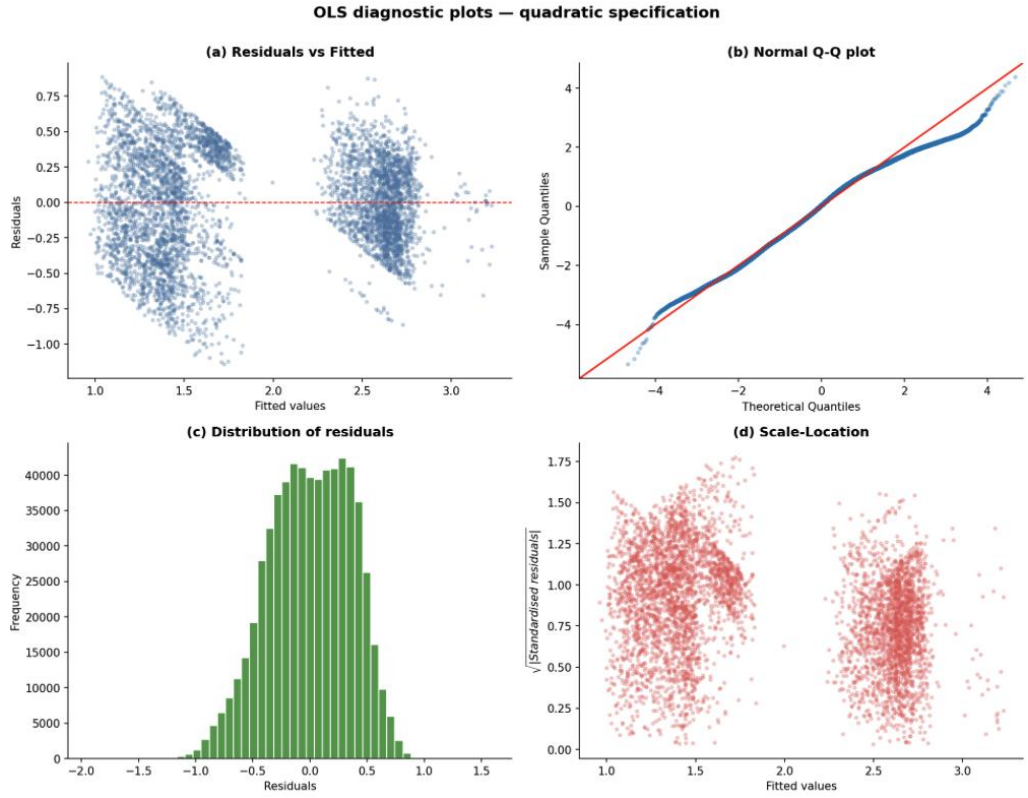


Figure 5.10. OLS diagnostic plots for the quadratic specification: (a) residuals versus fitted values; (b) normal Q-Q plot; (c) distribution of residuals; (d) scale-location plot.

Three points emerge from these plots. First, the distribution of residuals (panel c) is approximately symmetric and bell-shaped, with a skewness of -0.24 and negative excess kurtosis (-0.49), indicating slightly thinner tails than a normal distribution; the central portion of the Q-Q plot (panel b) lies close to the reference line, and given the very large sample the central limit theorem ensures that inference on the coefficients remains valid despite the mild departures in the tails. Second, the residuals-versus-fitted plot (panel a) displays two visible groupings of fitted values rather than a single continuous cloud. This bimodality is not a sign of misspecification: it reflects the genuine structure of the population documented in Section 5.7, in which commerce-and-services customers form a distinct, long-tenure group (with a coefficient of roughly +0.95 on log-tenure) that is separated from all other occupations. The fitted values therefore naturally fall into a higher-tenure cluster and a lower-tenure cluster. The mean residual is zero within each region. This occupation-based explanation can be verified directly: when the same residuals-versus-fitted plot is colour-coded by occupation, the two groupings coincide exactly with commerce-and-services customers on the one hand and all other occupations on the other, confirming that the bimodality is generated by the occupation dummies already included in the model rather than by an omitted factor.

Third, the scale-location plot (panel d) shows a modest difference in residual spread between the two groupings, consistent with the heteroskedasticity formally detected by the Breusch-Pagan test; this is fully accommodated by the heteroskedasticity-robust (HC1) standard errors employed in all specifications. Taken together, the diagnostics support the validity of the model while transparently revealing the occupation-driven structure that is itself a central finding of the analysis.

Because the explanatory power of the demographic model is unusually high for cross-sectional individual-level data, two further checks were carried out to rule out artefactual explanations. First, the possibility of data leakage was examined by computing the correlation between each of the eight financial-behavioural variables and tenure; all correlations were negligible ($|r| < 0.11$), confirming that no explanatory variable mechanically encodes the dependent variable. The financial block, when added to the demographic model, raised the R^2 by only 0.011, further indicating the absence of leakage. Second, a possible ceiling effect was considered, since a customer's age places an upper bound on how long they can have banked with the institution. Contrary to this expectation, age contributes very little to the model: age alone explains only 14.1% of the variation in tenure and removing it from the full model lowers the R^2 by less than one percentage point (from 0.727 to 0.718). The high explanatory power is therefore not a mechanical age-tenure artefact.

Instead, the explanatory power is concentrated almost entirely in occupation, which alone accounts for an R^2 of 0.712. This reflects a pronounced and clean separation in average tenure across the three occupational groups: commerce-and-services customers average 13.2 years, the residual "other" category 4.4 years, and agriculture-forestry-fishing customers only 2.9 years. While this association is genuine and not an artefact, its magnitude should be interpreted with care. The relationship between occupation and tenure may reflect not only intrinsic differences in the banking behaviour of occupational groups, but also the historical development of the bank's customer base - for example, if the bank's earlier expansion was concentrated among commerce-and-services customers and its more recent growth among agricultural customers, occupation would partly proxy for the era in which each cohort was acquired. With a single cross-section these channels cannot be separated, and the occupation coefficient is therefore best read as a strong conditional association rather than a pure behavioural effect.

5.8.1. Heterogeneity by occupation

The residual structure documented above - two distinct clusters of fitted values corresponding to commerce-and-services customers versus all others - raises the question of whether occupation merely shifts the level of tenure or whether it also changes how the other determinants operate. To examine this, the model was re-estimated separately for the two groups. A Chow test strongly rejects the hypothesis that the two groups share the same coefficient vector ($F = 45{,}244$, $p < 0.001$ for the quadratic

specification); although such a test is almost always significant at this sample size, the differences are also substantial in magnitude, as Table 5.12 shows.

Variable	Commerce & services	Other occupations
Constant	2.640***	1.178***
Age: young adult	-0.294***	-0.059***
Age: middle-aged	-0.109***	0.012***
Age: retirement	0.049***	-0.039***
Gender: female	0.001	0.021***
Region: East HCMC	0.092***	0.128***
Region: Southeast	-0.117***	0.104***
Segment: Cluster 6 (premium)	-0.020	1.584***
Segment: Cluster 3	-0.074***	1.309***
ln (CASA)	0.0026***	0.0087***
ln (CASA)²	0.0009***	0.0059***
ln (Savings)	0.0041***	-0.0012
ln (Products)	-0.0243***	-0.8585***
ln (Products)²	0.0993***	0.7091***
R-squared	0.202	0.176
Observations	265,637	365,425

Table 5.12. Separate (quadratic) regressions for commerce-and-services versus all other occupations.

Notes: *** $p < 0.01$. Robust (HC1) standard errors. Occupation is omitted within each sub-sample (constant by construction). Selected coefficients shown; full results available on request. A Chow test rejects coefficient equality across the two groups ($F = 45{,}244$, $p < 0.001$).

The contrast is striking and goes well beyond a simple level shift. Three differences stand out. First, the number of products held is almost unrelated to tenure among

commerce-and-services customers (linear coefficient -0.024) but strongly negative among other occupations (-0.859); the U-shaped product effect identified earlier is therefore driven overwhelmingly by the non-commercial group. Second, behavioural segment membership barely matters for commerce-and-services customers (the premium-segment coefficient is small and insignificant) but is a powerful predictor among other occupations (the premium coefficient reaches +1.58). Third, several regional and age effects even change sign between the two groups. In short, occupation does not simply raise or lower the baseline level of tenure; it conditions the entire relationship between customer characteristics and loyalty.

Substantively, this suggests that for commerce-and-services customers - whose banking is bound up with the operation of a business - relationship duration is high and relatively insensitive to the other attributes, whereas for personal and agricultural customers tenure is shaped much more strongly by segment, product holding, and balances. This heterogeneity is a meaningful refinement of the pooled results: the pooled coefficients (Table 5.10) represent an average across two rather different regimes, and the negative pooled product coefficient in particular masks the fact that the effect is concentrated in the non-commercial group. The pooled model remains a valid and parsimonious summary, but these split-sample results add an economically important layer of interpretation and point to occupation-specific retention strategies.

5.9. Discussion

The analysis identifies a clear and economically interpretable hierarchy of tenure determinants. Occupation is the dominant factor, with commerce-and-services customers exhibiting by far the longest relationships; age contributes a credible, monotonic effect that is robust to a causal reading; the premium segment emerges as the most durable; and balance measures are positively associated with tenure. The single counter-theoretical result - the negative coefficient on product holding - does not undermine these conclusions but rather validates the study's cautious treatment of the endogenous financial block and is consistent with international evidence that cross-buying is often a consequence rather than a cause of loyalty (Reinartz, Thomas & Bascoul, 2008).

A further qualification concerns the heterogeneity uncovered in Section 5.8.1. Splitting the sample by occupation reveals that the pooled coefficients conceal two rather different regimes. Among commerce-and-services customers, tenure is uniformly high and largely insensitive to segment membership and product holding; among all other occupations, by contrast, these same factors are powerful predictors, and the U-shaped product effect is concentrated almost entirely in this group. Occupation therefore does not merely shift the level of tenure but conditions the way every other characteristic relates to it. This refinement strengthens, rather than weakens, the central message: the demographic and segment attributes that the bank can observe early are precisely the ones that distinguish durable from transient relationships, and their predictive content is

strongest exactly where it is most useful - among the broad base of personal and agricultural customers whose loyalty is least guaranteed.

This econometric analysis complements the segmentation of Chapter 4 in two ways. Methodologically, it answers the supervisor's call for an explanatory model that goes beyond description, applying a regression framework with an explicit treatment of endogeneity. Substantively, it connects the behavioural segments to a concrete economic outcome - relationship duration - and shows that the premium segment is not only the most valuable in balance terms but also the most durable. For the bank, the practical implication is that the customers most likely to develop into long, valuable relationships can be identified early from largely predetermined attributes - occupation, age, and segment - rather than from product counts that may themselves be consequences of the relationship's stage.

The overlap between the two analyses is deliberate but only partial: just three of the eight clustering inputs (CASA balance, savings balance, and products held) re-enter the tenure model, while the remaining explanatory power comes from predetermined demographics and from segment membership. The econometric model is therefore not a re-estimation of the segmentation on the same information, but a separate exercise that uses the segments - together with demographics - to explain an outcome (relationship duration) that was never part of the clustering.

Chapter 6. Limitations

Although the segmentation produced stable, interpretable, and externally coherent results, several limitations must be acknowledged. They are presented openly here both to qualify the conclusions of Chapter 4 and to guide the future work outlined in Chapter 7.

Cross-sectional snapshot. The dataset captures each customer at a single point in time. The segmentation is therefore static: it cannot describe how customers transition between segments over time, distinguish a temporarily dormant customer from a permanently disengaged one, or detect seasonal and life-cycle effects. A customer's segment membership may drift, and the model would need periodic re-fitting; the present study does not quantify how quickly the segments decay.

Partial circularity of the validation. The Random Forest in Section 4.4 was trained on the same eight features that defined the clusters. Its very high accuracy therefore primarily demonstrates that the clusters are learnable and well separated in the input space - it does not constitute fully independent validation, because the labels are themselves a deterministic function of those features. The held-out lift analysis (Section 4.3.3) mitigates this concern, but the held-out variables were examined after the clusters were formed rather than pre-registered, so a degree of confirmation bias cannot be excluded.

No competing algorithm was actually evaluated on the full data. The claim that MiniBatch K-Means is the most suitable method rests on a combination of literature precedent, the approximately linear structure of the data, and the practical infeasibility of the two alternatives at this scale. However, because Agglomerative Hierarchical Clustering and DBSCAN could not be run on the full population, no genuine head-to-head comparison of clustering quality was performed on this dataset. A sample-based comparison, or the use of scalable approximations of these algorithms, would have strengthened the selection argument.

Severe class imbalance. Cluster 0 contains 72.9% of all customers, while the smallest cluster contains only 0.9%. Internal metrics such as the Silhouette Score can be inflated by a single large, tightly packed group, and the rarest segments are supported by relatively few observations, making their profiles and the corresponding Random Forest metrics less statistically reliable than those of the larger segments.

Unusually optimistic internal metrics. The Hopkins statistic of 1.000 and the Silhouette Score of 0.87 are exceptionally high for real-world behavioural data. These values are very likely inflated by the extreme zero-inflation of four of the eight variables and by the dominance of the tightly clustered mass-market group. Such metrics should be read with caution: they indicate strong geometric separation in the transformed space, but a large part of that separation is driven by has / does-not-have distinctions rather than by fine-grained behavioural gradients.

Near-binary variables in a Euclidean algorithm. Four variables (outstanding loan, savings, credit-card spending, collateral) are effectively binary once you account for their 80-to-97% zero rates. K-Means leans on Euclidean distance and assumes features that are roughly continuous and isotropic, so it is not, in theory, the ideal tool for data that mixes continuous and binary like this. Methods built for mixed types - K-Prototypes with a Gower distance, say - might capture the structure more faithfully, and they were not tried here. Four variables (outstanding loan, savings, credit-card spending, collateral) are effectively binary after accounting for their 80 to 97% zero rates. K-Means, which relies on Euclidean distance and assumes roughly continuous, isotropic features, is not the theoretically ideal tool for such mixed continuous-and-binary data. Algorithms designed for mixed data types, such as K-Prototypes with a Gower distance, might capture this structure more faithfully and were not tested here.

Feature scope. Only eight financial-behavioural variables went into the clustering. Other fields that could have helped, listed in Table 3.2 - income, occupation, tenure, education, channel behaviour - were held back for validation instead of fed into the model. Settling on eight inputs was partly a pragmatic call, and a fuller feature set might bring out finer or simply different segments. The clustering used only eight financial-behavioural variables. Potentially informative fields documented in Table 3.2 - income, occupation, tenure, education, and channel behaviour - were reserved for validation rather than incorporated into the model. The choice of

eight inputs was partly pragmatic, and a richer feature set might reveal finer or different segments.

Absence of business-outcome validation. The segments were interpreted from the medians of their input variables and labelled with descriptive names such as premium or loyal. These labels are inferred, not measured: the study did not validate the segments against realized profitability, campaign-response rates, churn, or customer-lifetime value. Until such outcome data is linked, the business value of each segment remains a well-motivated hypothesis rather than a demonstrated fact.

Single bank, single dataset. All data originates from one Vietnamese commercial bank at one point in time. The specific segments, their sizes, and their profiles may not generalize to other banks, markets, or periods, and no external or temporal hold-out was available to test transferability.

Survivorship bias in the tenure analysis. The tenure analysis of Chapter 5 is estimated on the population of active customers only, since accounts that had already been closed were excluded at extraction. Tenure is therefore observed for those customers who have remained with the bank - the "survivors" - and not for those who have already left. As a result, the estimated relationships describe the durability of ongoing relationships among retained customers, rather than the determinants of loyalty in the full sense that would also account for attrition. The factors associated with longer tenure among survivors need not coincide with the factors that protect against churn in the broader customer base. A complete treatment of customer retention would require survival or duration models estimated on data that include closed accounts, with the exit event explicitly modelled; this is identified as a priority for future work.

The segment label as a generated regressor. In the tenure regression, the cluster label is included as an explanatory variable. This label is itself an estimated quantity - the output of the K-means model - and is therefore measured with classification error, and it is derived in part from financial variables (such as CASA and savings balances) that also enter the regression separately, creating a degree of information overlap. For both reasons, the coefficients on the segment indicators are interpreted as conditional associations rather than clean causal effects. The robustness check reported in Chapter 5 (Model 3b), in which the segment indicators are removed without changing the sign of the product-holding coefficient, mitigates but does not eliminate this concern.

Stability is not validity. Finally, the near-perfect Adjusted Rand Index demonstrates that the solution is reproducible across random seeds, but reproducibility is not the same as correctness. A stable solution can still be a stable artefact of the preprocessing and the algorithm's assumptions; stability should therefore be interpreted as a necessary, not a sufficient, condition for a trustworthy segmentation.

Chapter 7. Conclusion and Future Work

7.1. Conclusion

This study applied MiniBatch K-Means to a large-scale, real-world dataset of 631,062 customers from a Vietnamese commercial bank, using eight financial-behavioural variables and a fully documented analytical pipeline comprising log1p transformation, RobustScaler standardization, Hopkins-statistic and PCA-based cluster-tendency assessment, multi-metric model selection, stability testing, and Random Forest validation.

The analysis identified eight customer segments: a dominant mass-market group (72.9%), a deposit-oriented group, three credit-oriented groups distinguished by collateral and card behaviour, and three small but high-value groups. The solution proved exceptionally stable (mean Adjusted Rand Index of 0.9999) and was reproduced by a Random Forest classifier with 99.85% test accuracy and a macro-averaged F1-score of 0.996. The Random Forest feature-importance ranking, in agreement with the Z-score analysis, identified savings balance, outstanding loan, collateral value, and credit-card spending as the most decisive variables, while the held-out lift analysis showed that variables excluded from the clustering nonetheless reproduced the segment hierarchy. Together these findings answer the three research questions: MiniBatch K-Means was the only feasible and the most suitable algorithm at this scale (RQ1); balance- and savings-related variables most strongly differentiate the segments (RQ2); and the segments are stable, robust, and externally coherent enough to support business strategy (RQ3).

Beyond the segmentation, the econometric analysis of Chapter 5 examined the determinants of customer tenure and thereby moved the study from description to explanation. Using ordinary least squares on the full population, with an explicit treatment of reverse causality, the analysis found that occupation is the single strongest predictor of relationship length - commerce-and-services customers bank substantially longer than others - while age contributes a credible monotonic effect. Most importantly for strategy, the premium (VIP) segment emerges as the most durable of all groups, recording the longest tenure: the customers who are most valuable in balance terms are also those who remain with the bank the longest. This answers RQ4 and links the behavioural segments to a concrete economic outcome.

This dual status - highest value and greatest loyalty - identifies the premium segment as the clearest priority for retention investment, even though it represents under one percent of the customer base. The economic logic is well established: in the foundational study of service-sector loyalty, Reichheld and Sasser (1990) showed that reducing the customer defection rate by just five percent raised profits by eighty-five percent in one bank’s branch system, and subsequent work confirms that retaining existing high-value customers is substantially more profitable than acquiring new ones, because long-tenured customers generate growing revenues at a declining cost to serve (Reichheld & Teal, 1996; Reinartz & Kumar, 2003). For Eximbank, this implies that protecting and deepening its already-loyal VIP relationships - through dedicated relationship management, preferential service, and tailored cross-selling - offers a higher

and more certain return than acquisition alone. Because these customers are identifiable in advance from largely predetermined attributes such as occupation, age, and segment membership, the bank can target retention resources before, rather than after, the risk of attrition materialises. Consistent with the study's cautious stance, the demographic coefficients are read as conditional associations with a credible causal interpretation, whereas the financial-behaviour coefficients - including the counter-intuitive negative association of product holding with tenure - are interpreted as conditional correlations rather than causal effects.

A final, more nuanced implication follows from the occupation-based heterogeneity documented in Chapter 5. The determinants of tenure do not operate uniformly across the customer base: for commerce-and-services customers, whose banking is tied to the running of a business, relationships are long and stable almost regardless of segment or product holding, whereas for personal and agricultural customers tenure depends far more heavily on segment membership and financial engagement. This implies that a single, uniform retention policy would be inefficient. The bank should instead differentiate its approach by occupation - relying on the naturally durable relationships of its business customers while actively managing segment and product engagement among personal and agricultural customers, where loyalty is more fragile and more responsive to the bank's actions. Recognising this heterogeneity turns the econometric findings into concrete, segment-and-occupation-specific retention strategies.

Taken together, the contribution of this study operates on three levels. Practically, it offers the bank a data-driven tool for identifying, at an early stage, the customers most likely to develop into long and valuable relationships: because the strongest predictors of tenure - occupation, age, and segment membership - are largely predetermined and observable from the outset, retention resources can be directed before the risk of attrition materialises rather than after, and can be tailored by occupation rather than applied uniformly. Methodologically, it demonstrates how unsupervised machine learning and econometric modelling can be combined on a single large-scale dataset: clustering describes who the customers are, while regression explains why some relationships endure, with an explicit and honest treatment of endogeneity that keeps the interpretation proportionate to what a cross-section can support. Academically, it adds quantitative evidence on customer loyalty from an emerging-market commercial bank - a setting under-represented in a literature dominated by North American and European studies - and documents two non-obvious empirical regularities: a U-shaped relationship between product holding and tenure, and a pronounced moderation of every determinant by occupation. The central limitation qualifies all three: because the data form a single cross-section, the findings are best read as conditional associations and predictive relationships rather than proof of cause and effect, a question that the panel and instrumental-variable extensions outlined below are designed to address.

From a practical standpoint, the study demonstrates a concrete, reproducible path by which a Vietnamese bank can convert its existing transactional data into an operational segmentation: the labelled population enables immediate targeting, and the trained Random Forest enables real-time assignment of new customers without re-running the clustering. The largest immediate opportunity lies in activating the mass-market majority, while the small high-value segments warrant dedicated retention investment.

7.2. Future Work

Chapter 6's limitations suggest a few things worth doing next. The data is the first one. Right now it's a single snapshot. Longitudinal data would change that. You could segment dynamically, and track customers actually moving between segments instead of treating them as fixed. The methods are worth revisiting too. K-Means scaled, so it won. But it was never the right tool in theory for variables that sit at zero 80 to 97 percent of the time. Someone should test it properly against sample-based versions of the other algorithms, and against mixed-type methods like K-Prototypes that handle near-binary variables better than Euclidean distance does. Whether the easy choice was also the correct one is still an open question. Then there's the practical side. These segments were never tied to anything the bank books. Link them to profitability, churn, and campaign response, and the labels stop being inferred and start being worth money. The segmentation could be tuned for that instead of for clean geometry. The feature set could grow as well. Income, tenure, channel behaviour were all held back for validation. Put them in, re-check the segments on other banks or later periods, and you'd find out how far any of this generalizes. Probably it improves. Last thing, and it's less analysis than plumbing: get the Random Forest into the bank's live systems, watch for drift, and feed what drifts back into the next pass.

References

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), The Oxford Handbook of Leadership and Organizations (pp. 93–117). Oxford University Press.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27.

Cooil, B., Keiningham, T. L., Aksoy, L., & Hsu, M. (2007). A longitudinal analysis of customer satisfaction and share of wallet: Investigating the moderating effect of customer characteristics. Journal of Marketing, 71(1), 67–83.

Coulter, K. S., & Coulter, R. A. (2002). Determinants of trust in a service provider: The moderating role of length of relationship. Journal of Services Marketing, 16(1), 35–50.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227.

Eximbank (2026). Separate Financial Statements for the Year Ended 31 December 2025. Vietnam Export Import Commercial Joint Stock Bank. Retrieved from https://eximbank.com.vn/bao-cao-tai-chinh

FiinGroup (2024). Vietnam Banking Report 2024. FiinGroup, Hanoi.

Firdaus, U., & Utama, D. N. (2020). Balance as one of the attributes in the customer segmentation analysis method: Systematic literature review. Advances in Science, Technology and Engineering Systems Journal, 5(3), 334–339.

Gupta, M. (2023). Customer Segmentation using Machine Learning applied to Banking Industry [master's thesis, Hochschule Neu-Ulm University of Applied Sciences]. https://publications.hnu.de/3830/

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing customers. Journal of Marketing Research, 41(1), 7–18.

Leszczensky, L., & Wolbring, T. (2022). How to deal with reverse causality using panel data? Recommendations for researchers based on a simulation study. Sociological Methods & Research, 51(2), 837–865.

Niyagas, W., Srivihok, A., & Kitisin, S. (2006). Clustering e-banking customer using data mining and marketing segmentation. ECTI Transactions on Computer and Information Technology, 2(1), 63–69.

Reichheld, F. F., & Sasser, W. E., Jr. (1990). Zero defections: Quality comes to services. Harvard Business Review, 68(5), 105–111.

Reichheld, F. F., & Teal, T. (1996). The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value. Harvard Business School Press.

Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. Journal of Marketing, 64(4), 17–35.

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing, 67(1), 77–99.

Reinartz, W., Thomas, J. S., & Bascoul, G. (2008). Investigating cross-buying and customer loyalty. Journal of Interactive Marketing, 22(1), 5–20.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Rust, R. T., Lemon, K. N., & Zeithaml, V. A. (2004). Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing, 68(1), 109–127.

Sculley, D. (2010). Web-scale k-means clustering. In Proceedings of the 19th International Conference on World Wide Web (WWW '10) (pp. 1177–1178). ACM. https://doi.org/10.1145/1772690.1772862

Statista (2025). Banking Industry in Vietnam: Statistics and Facts. Statista.

Vietnam.vn (2026, April 1). State Bank of Vietnam promotes digital payments, tightening security and combating fraud. Retrieved from https://www.vietnam.vn

Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations (2nd ed.). Kluwer Academic Publishers.

Declaration on the Use of Artificial Intelligence (AI) Tools

The undersigned Van Duc Dao, student ID number 908607, enrolled at Ca' Foscari in the Bachelor's/Master's Degree Programme in Data Analytics for Business and Society, author of the final paper/degree thesis Customer Segmentation and the Determinants of Relationship Tenure in a Vietnamese Commercial Bank: A Machine-Learning and Econometric Analysis of Large-Scale Banking Data, thesis supervisor Agar Brugiavini, aware of the penalties provided for in Article 76 of the Consolidated Act set out in Presidential Decree no. 445 of 28 December 2000, and of the loss of benefits provided for in Article 75 of the same Consolidated Act in the event of false or misleading statements, under his/her own responsibility pursuant to Articles 46 and 47 of Presidential Decree no. 445 of 28 December 2000,

DECLARES

☐ that he/she has **NOT** used generative artificial intelligence technologies to support various aspects of the research and writing of the final paper/degree thesis;

☒ that he/she **HAS** used generative artificial intelligence technologies to support various aspects of the research and writing of his/her final paper/degree thesis, in particular (for each point, write in full the requested information):

- indicate the name of the tool (e.g., "ChatGPT" if an application or "GPT" if a model), with reference to the specific version (for example, "ChatGPT-4, version released in March 2024");

  Claude (Anthropic), model Claude Opus 4 / Opus 4.x, accessed via the Claude web application during 2025–2026. A single tool was used.

- if more than one tool has been used, specify each of them by indicating its name and version;
- provide information on how the tool was used:
  - ☒ revision and optimisation of texts to improve clarity and coherence;
  - ☐ generation of charts and diagrams for data presentation;
  - ☒ support for the automatic drafting of sentences;
  - ☐ production of illustrative images to complement the methodological section;
- indicate the sections or chapters concerned;

  Sections concerned: Abstract; Chapter 1 (Introduction — motivation and references); Chapter 2 (Literature Review — wording and references); Chapter 3 (Methodology — description of the data and variables.

- indicate how the accuracy, quality and reliability of any information or text that may have been generated were verified.

  Verification: every quantitative result, table and figure was produced by the author by running the analysis code on the real Eximbank dataset, and the numbers reported in the thesis were checked directly against that output; no data, results or citations were generated or invented by the AI tool. All AI-assisted text was read, fact-checked, edited and reworked by the author, who confirmed that every claim is supported by a verifiable source and that all figures were generated by the author from the study's own data.

DECLARES FURTHER

that all material generated by AI has been carefully checked, reviewed and reworked by the author, who assumes full responsibility for the quality, accuracy and scientific integrity of the work.

Date: 23/06/2026 Signature: Duc

Van Duc Dao