Smart Personalised Finance Management

Candidate number: 1088093

University of Oxford

A thesis submitted in partial fulfillment of the MSc in

MSc Mathematical and Computational Finance

July 3, 2025

Acknowledgements

The completion of this dissertation would not have been possible without the exceptional support and guidance of my industry supervisors, Mr. Kal Bukovski and Mr. Chris Journeay, whose intellectual contributions proved invaluable at every stage.

I also wish to thank Dr. Leandro Sanchez-Betancourt, my departmental supervisor at the Mathematical Institute, for his careful review and helpful suggestions, which greatly refined this work.

Finally, I dedicate this dissertation to the loving memory of my Dad, whose strength and resilience continue to guide me to this day.

Abstract

This dissertation addresses limitations in holistic financial advisory by proposing an Open Finance-driven framework. It focuses on three key areas: first, augmenting limited synthetic financial data with crucial demographic and behavioral labels for comprehensive modeling; second, developing a robust model to estimate individual customer monthly cash flow components (housing, discretionary spending, savings) using these enriched profiles; and finally, leveraging the derived insights to provide personalized and actionable financial recommendations tailored to unique customer circumstances, risk preferences, and financial objectives. This integrated approach aims to demonstrate Open Finance's transformative potential in personalized financial management.

Contents

List of Figures

Chapter 1

Introduction

The financial services landscape is undergoing major changes, driven by technological advancements and changing regulatory frameworks. Open Banking and Open Finance initiatives, in particular, lead the transformations, and hold the potential to redefine the relationship between consumers, financial institutions, and innovative service providers. This dissertation aims to contribute by utilizing diverse datasets to generate optimal financial advice for customers.

1.1 The Open Banking and Open Finance Revolution

The rise of Open Banking, mandated by regulations like the revised Payment Services Directive (PSD2) [9] in Europe and the Competition and Markets Authority (CMA) Open Banking Roadmap [8] in the UK, marked an important change in the accessibility and use of financial data. It required major banks to allow third-party providers (TPPs), with customer consent, to access financial records (e.g., transaction history) and execute payments. This transformation is currently evolving into the broader concept of Open Finance. Open finance is an extension of Open Banking, moving beyond just banking transactions to include a broader range of financial products like mortgages, insurance, investments, and pensions. It allows third-party providers to access and use consented customer financial data to create personalized and innovative services, thereby fostering a more connected and customer-centric financial ecosystem [4].

The benefits of Open Finance ecosystem are substantial [6]. Firstly, Open Finance empowers consumers by allowing them a greater degree of control over their own financial data, as well as a holistic view of their financial health across multiple platforms. This unified perspective facilitates more informed decision-making and allows for one-stop access

of financial information. Secondly, it fosters better innovation and competition within the financial sector. FinTechs and challenger banks can leverage this rich, consented data to develop highly personalized products and services, which runs the gamut from intelligent budgeting tools and tailored lending solutions to dynamic investment advice. This increased competition is expected to decrease overall costs, improve service quality, and enhance financial inclusion by offering bespoke solutions to underbanked segments such as SMEs. Ultimately, the vision is a more interconnected and customer-centric financial ecosystem which adapts to individual needs in real-time.

Even with these transformative promises, the Open Finance revolution also poses significant challenges [23]. **Data security and privacy** are important issues. Strong data encryption, secure APIs, and transparent consent mechanisms are required to uphold consumer trust and prevent breaches. The regulatory landscape (especially laws governing data privacy and usage) is complex and continually evolving, necessitating vigilance from all stakeholders to ensure strict compliance. Furthermore, the **quality and standardization of data** across different financial institutions can vary significantly, posing hurdles for data aggregation and data consistency. Consumer **trust and adoption** are also key hurdles; many individuals remain hesitant to share sensitive financial information through Open Banking despite potential advantages. Finally, creating sustainable business models for TPPs and institutions which utilize this new data paradigm, while dealing with potential ethical implications inherent in AI-driven recommendations, will be crucial for long-term success.

1.2 Related Work and Literature Review

The growing field of Open Finance has since encouraged significant research and development in leveraging granular financial data for enhanced personalized financial services. Several bodies of literature are particularly relevant to this dissertation's aims:

Firstly, some proof-of-concept studies have been conducted on *personalized financial advice (PFA)* and *robo-advisory platforms*. Traditional PFA often relies on self-reported data or fragmented institutional records. However, recent research explores how real-time transaction data available through Open Finance can drastically improve the accuracy and dynamism of PFA systems, especially with credit and debt management [7]. In particular, cash flow patterns indicative of imprudent financial behavior could be flagged on their onset

prior to snowballing of financial debt.

Secondly, researchers have increasingly applied machine learning techniques on transactional data. Detailed spending habits have been used to segment customers into behavioral clusters[5], predict future financial behavior [26](e.g. savings capacity, likelihood of default), and detect fraud [3]. Their work demonstrates the potential of otherwise mundane granular transactional data, as rich insights on consumers' behavioral patterns could be inferred and acted upon.

Thirdly, data augmentation and synthetic data generation techniques are increasingly important, especially in the context of sensitive financial information [2]. Given privacy concerns and the difficulty of sharing real financial datasets, methods for generating realistic synthetic financial data that preserve real-world statistical properties are vital for research and development. The use of real-world distribution data from official sources such as Office of National Statistics (ONS) in enriching existing data, as explored in this dissertation through statistical imputation, aligns with approaches aimed at enhancing dataset utility while maintaining integrity.

Finally, some studies explore how diverse data points—from transaction history to demographic information—can be integrated to provide holistic and actionable advice. This often involves clustering techniques to identify distinct customer segments and then developing optimization frameworks to provide recommendations tailored to each segment's characteristics, risk preferences, and financial goals [25]. While a fully integrated "optimal financial recommendation" engine leveraging extensive Open Finance data remains still out of reach, the foundational components are actively being researched across these various domains.

1.3 Problem Statements

Building upon the opportunities presented by Open Finance and addressing the current limitations in holistic financial advisory, this dissertation sets out to address three key problem statements:

1. **Data Preparation and Enrichment:** To suitably enrich an initial limited synthetic dataset, containing only salary and monthly credit information, by generating additional demographic and behavioral labels (such as age group, region, and SOC

occupational code) in a sensible manner which respects real-world distribution to prepare it for comprehensive financial modeling.

2. **Customer Cash Flow Modeling:** To develop a statistically robust model for estimating the monthly cash flow components of individual customers, including monthly housing costs, discretionary spending, and savings, based on their enriched financial and demographic profiles, while injecting stochasticity to reflect customer heterogeneity.

3. **Tailored Financial Recommendations:** To provide personalized financial recommendations to customers, taking into account their unique financial circumstances, derived cash flow patterns, inferred risk preferences and financial objectives.

This dissertation seeks to integrate these components into a coherent framework that demonstrates the potential of Open Finance data in revolutionizing personalized financial management.

Chapter 2

Exploratory Data Analysis and Data Enrichment

This chapter analyzes the initial synthetic dataset and explores methods for its enrichment to facilitate subsequent modeling. While the data is synthetic due to the sensitive nature of real banking information, its generation heuristics are unknown. For the purposes of this study, the dataset is treated as real-world data, and all customers are assumed to reside in the UK, given income in pounds sterling.

2.1 Exploratory Data Analysis

The initial dataset comprises 100,000 rows of customer income and credit data with the following columns:

Table 2.1: Overview of Initial Dataset Features

Category	**Feature Name**	**Data Type**
Static Data		
Customer Profile	1) Access to revolving credit products (credit card/overdraft facility)	Boolean
	2) Short-term 24-month loan status	Boolean
	3) Monthly fixed short-term loan payment amount	Numerical
	4) Credit limits for credit cards and overdrafts	Numerical
	5) Chargeable interest as percentage of outstanding balance on credit card/overdraft arrears	Numerical
12-Month Time Series Data ($\{\textbf{feature}\}_{\textbf{month}}$)		
Credit Behavior	1) Monthly balance across revolving credit lines (credit card/overdraft)	Numerical
	2) On-time payment status of previous month's revolving credit balance (assuming zero or in-full payments)	Boolean
	3) Number of late days for non-punctual payments	Numerical

Comparing our synthetic data's income distribution with the ONS 2022 Distribution of

UK Household Disposable Income [18] (Figure 2.1) reveals significant deviations. The UK population's income follows a bell-shaped curve with asymmetric tails (incomes exceeding £100,000 are grouped), whereas our synthetic data appears to be generated in three roughly uniform tranches.

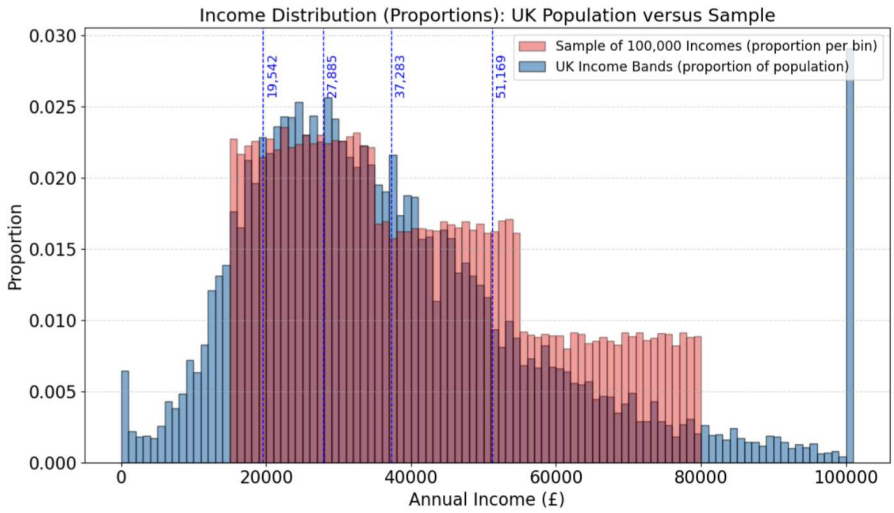

Figure 2.1: Income distribution: UK population (ONS) versus sample

A similar discrepancy is observed when overlaying the distribution of the UK population's unsecured debt-to-income (DTI) ratio from the FCA 2020 Financial Lives Survey [11] against our sample (Figure 2.2). For this analysis, unsecured debt includes outstanding credit card, overdraft balances, and short-term loan amounts.

While our synthetic data reflects some real-world segmentation with a bimodal Debt-to-Income (DTI) distribution (many having near-zero unsecured debt and a smaller group around 0.3), it is worth noting that real-world DTI data is far more asymmetric; over 60% of the population has near-zero unsecured debt, with a long tail extending beyond 0.7.

Finally, it is worth mentioning that we should not expect our data to perfectly mirror the UK population distribution, as different banks serve distinct customer segments in real life. For example, some private banks cater to wealthier clients, while building societies focus on the underbanked.

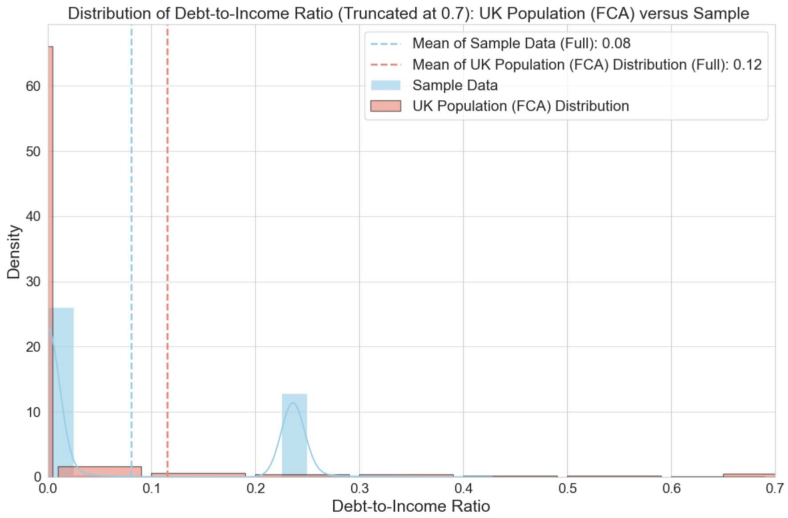

Figure 2.2: Unsecured debt-to-income distribution: UK population (FCA) versus sample

2.2 Data Enrichment

To enable customer segmentation and tailored financial advice, we'll enrich our initial dataset with additional demographic information. This is crucial as financial advice, for instance, varies significantly between a 20-year-old renter in London and a homeowner nearing retirement, even if their income and debt level are identical.

Useful demographic categories for modeling include:

1. *Age group*: Aged below 24; Aged 25-34; Aged 35-49; Aged 50-64; Aged over 65
2. *Region*: London; East; South East; South West; North East; North West; Wales; East Midlands; West Midlands; Yorkshire and Humber
3. *Standard Occupational Classification (SOC)*: Managers, directors and senior officials; Professional occupations; Associate professional and technical; Administrative and secretarial; Skilled trades; Caring, leisure and other services; Sales and customer service; Process, plant and machine operatives; Elementary occupations
4. *Housing tenure*: Renter, mortgage, own outright

When augmenting a dataset with new variables, say X_{n+1} to an existing set $(X_1, X_2, \ldots, X_n)$, ideally one would use the full joint distribution $P(X_1, X_2, \ldots, X_n, X_{n+1})$ to derive the conditional distribution $P(X_{n+1}|X_1, X_2, \ldots, X_n)$. This would allow for sampling new labels consistent with the underlying joint distribution.

However, access to such comprehensive joint distributions is often limited in real-world scenarios, particularly due to privacy concerns. Statistical agencies like the Office for National Statistics (ONS) typically provide only low-dimensional tables (e.g., two-way or three-way) or aggregated data, making it challenging to infer exact multi-dimensional distributions.

The problem of predicting or modeling X_{n+1} given $X_1, \ldots, X_n$ is commonly known as **imputation**. Below, we summarize several common approaches:

- **Copulas**: Sklar's theorem allows decomposing any multivariate joint distribution into its univariate marginals and a copula function capturing dependence. However, we often lack full univariate distributions when only conditional means are provided. Even with univariate distributions, modeling the copula function, especially in high-dimensional datasets (100,000 rows with multiple labels), is computationally intensive using methods like maximum likelihood estimation (MLE).

- **Iterative Proportional Fitting (IPF)**: IPF adjusts a reference joint distribution table $P_{\text{ref}}(X_1, \ldots, X_d)$ iteratively to match target marginals. For example, to impute region given tenure and age group, the algorithm cycles through each marginal until the correct distributions for tenure, age group, and region are met. While efficient for discrete variables, it requires discretizing continuous data, potentially losing information, and does not preserve higher-order dependencies.

- **Bayesian network**: Using Bayes' theorem, we can express the joint distribution as $P(X_1, \ldots, X_{n+1}) = P(X_1)P(X_2|X_1) \cdots P(X_{n+1}|X_1, \ldots, X_n)$. While obtaining all conditional likelihoods can be difficult, approximations can be made by dropping terms (e.g., approximating $P(X_3|X_1, X_2)$ with $P(X_3|X_1)$). The extreme case is the Naive Bayes approximation: $P(X_1, \ldots, X_{n+1}) \approx P(X_1)P(X_2|X_1)P(X_3|X_1) \cdots P(X_{n+1}|X_1)$. This method, despite assuming independence (which is rarely true in reality), offers flexibility. It accommodates both discrete and continuous distributions (e.g., by fitting Gaussian distributions) and allows approximating unknown conditional marginals.

Given its computational simplicity and flexibility, we will focus on the **Bayesian approach** in our subsequent work.

2.2.1 Fitting The Region Label

Our initial step involves assigning a region label, crucial for modeling region-dependent variables like home prices. From Bayes' theorem, we leverage the Naive-Bayes assumption that, given a region, income quintile is independent of unsecured debt-to-income (DTI) level:

$$\begin{aligned} P(\text{region}|\text{DTI , income quintile}) &= P(\text{region}) \times P(\text{DTI}|\text{region}) \\ &\quad \times P(\text{income quintile}|\text{DTI, region}) \\ &\propto P(\text{DTI}|\text{region}) \times P(\text{income quintile}|\text{region}) \end{aligned}$$

This independence assumption is a modeling simplification due to data constraints.

We estimate $P(\text{DTI}|\text{region})$ using a two-way table from the FCA 2020 Financial Lives Survey [11], providing average unsecured DTI for 10 UK regions. We assume each conditional distribution is Gaussian, with means from the FCA data and a single global variance estimated from FCA's grouped frequency distribution. $P(\text{income quintile}|\text{region})$ is derived from the ONS 2020 Small Area Model-based Income Estimates [17], which provides income quintile breakdowns by region. Multiplying these likelihoods and normalizing yields $P(\text{region}|\text{DTI, income quintile})$, enabling us to sample region labels for our 100,000 customers.

2.2.2 Fitting The SOC Code

We can apply a similar Bayesian approach to fit the Standard Occupational Classification (SOC) code. The ONS 2022 Annual Survey of Hours and Earnings (ASHE) [16] provides two key tables:

1. (occupation, region, income) three-way table, where each cell in the table represents the mean income for a particular (SOC occupation code, region) pair.
2. (occupation, region) two-way prior, where each cell in the table contains frequency of samples for a particular (SOC occupational code, region) pair.

Using Bayes' theorem, we can express the probability of an SOC code given region and income as:

$$\begin{aligned} P(\text{SOC}|\text{region, income}) &\propto P(\text{SOC}|\text{region}) \times P(\text{income}|\text{SOC, region}) \\ &\propto P(\text{SOC}|\text{region}) \times \exp[-\frac{(\text{income} - \mu_{\text{SOC, region}})^2}{2\sigma^2}] \end{aligned}$$

This formulation assumes income follows a **Gaussian distribution**. The conditional mean ($\mu_{\text{SOC, region}}$) is determined by the (occupation,region,income) three-way table, while the standard deviation (σ) is estimated from the ONS 2022 Distribution of UK Household Disposable Income [18]. The prior $P(\text{SOC}|\text{region})$ is derived from the (occupation, region) two-way prior. After multiplying these likelihoods, we normalize the product to obtain $P(\text{SOC}|\text{region,income})$ for sampling.

It's important to note that, unlike previous sections, we **do not assume independence** between variables here. The ONS data provides sufficient conditionals for a more accurate estimation.

2.2.3 Fitting The Age Group

Unlike the region and SOC label assignments, fitting the age group utilizes richer information from our original dataset, specifically capturing financial prudence through late payment flags. The FCA's 2024 recontact survey [12] provides data on the percentage of the population, by age group, who missed credit payments in the last six months, allowing us to estimate $P(\text{late}|\text{age group})$.

For our 12-month time series data, we define a "late payment flag" for individuals with three or more late payments across any credit product (credit cards, overdraft, short-term loan) within a 12-month period. This threshold of three was chosen empirically to best fit the FCA's $P(\text{late}|\text{age group})$ data, and to reflect FCA's intent of modeling genuine financial difficulty rather than minor payment tardiness in our synthetic data. (see Appendix D).

By Bayes' theorem, the probability of age given other variables is:

$$
\begin{aligned}
P(\text{age}|\text{SOC, region, income, late flag, DTI}) &\propto P(\text{age}|\text{SOC, region}) \\
&\quad \times P(\text{income}|\text{region, age, SOC}) \\
&\quad \times P(\text{late flag}|\text{income, region, age, SOC}) \\
&\quad \times P(\text{DTI}|\text{late flag, income, region, age, SOC}) \\
&\quad \propto P(\text{age}|\text{SOC, region}) \times P(\text{income}|\text{region, age}) \\
&\quad \times P(\text{late flag}|\text{age}) \times P(\text{DTI}|\text{age})
\end{aligned}
$$

This derivation incorporates Naive-Bayes assumptions:

1. Given region and age, income is independent of SOC.
2. Given age, the lateness flag is independent of SOC, region, and income.
3. Given age, the unsecured debt-to-income ratio is independent of late flag, region, income, and SOC.

We obtain $P(\text{age}|\text{SOC, region})$ from the ONS 2021 census three-way frequency table [21]. $P(\text{late flag}|\text{age group})$ is from the FCA survey [12]. $P(\text{income}|\text{region, age})$ comes from the ONS ASHE table [16]. Finally, $P(\text{DTI}|\text{age})$ is sourced from the FCA 2020 Financial Lives Survey [11]. The product of these prior and likelihoods, followed by normalization, yields the required conditional probability, from which we sample the age group label for our 100,000 customers.

2.2.4 Fitting The Tenure Label

Our final step is to model a customer's housing tenure: renting, owning outright, or having a mortgage.

Following our previous Naive-Bayes approach and assuming independence between variables, we have:

$$P(\text{tenure}|\text{age group, income quintile, region, DTI, late flag})$$

$$\begin{aligned}
&\propto P(\text{tenure}) \times P(\text{age group}|\text{tenure}) \\
&\times P(\text{income quintile}|\text{tenure}) \\
&\times P(\text{region}|\text{tenure}) \times P(\text{DTI}|\text{tenure}) \\
&\times P(\text{late flag}|\text{tenure})
\end{aligned}$$

The 2022-2023 English Housing Survey [14] provides three two-way frequency tables (region,tenure), (age group,tenure), and (income quintile,tenure), from which we derive $P(\text{age group}|\text{ tenure})$, $P(\text{income quintile}|\text{ tenure})$, $P(\text{region}|\text{ tenure})$, and the prior $P(\text{tenure})$.

The FCA 2020 Financial Lives Survey [11] offers average unsecured debt-to-income levels for each of the three tenure groups, allowing us to deduce $P(\text{DTI}|\text{ tenure})$. Additionally, the FCA 2024 Recontact Survey [12] provides late payment probabilities for each tenure group, enabling estimation of $P(\text{late flag}|\text{ tenure})$.

By performing standard normalization and sampling, we impute the housing tenure label for all 100,000 customers.

2.3 Robustness Checks

Given the absence of a real-world labeled dataset, assessing the accuracy of our imputed labels against actual data is challenging. Therefore, we have implemented several robustness checks to validate our imputation process.

2.3.1 Checks Against Known Marginal Distributions

Our primary check involves comparing the marginal distributions of our fitted labels against established data from official sources like the ONS.

We do not expect a perfect match due to the inherent bias in our initial synthetic income and unsecured debt-to-income data. However, any significant discrepancies should be explainable and justifiable.

In the following figures, the orange bars represent distributions from actual datasets, while the blue bars (labeled "robust" for these checks) represent our imputed dataset.

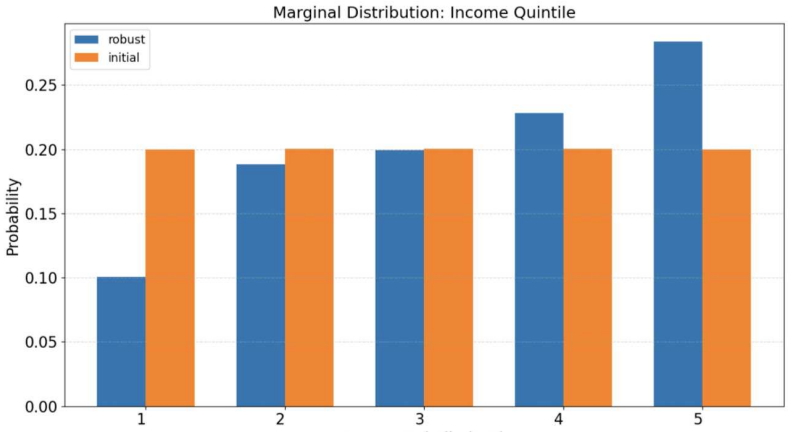

Figure 2.3: Income quintile: UK population (ONS) versus sample

Figure 2.3 confirms the known bias in our dataset towards higher income quintiles. It underweights the lowest income quintile by up to 10 percentage points (ppt) and overweights

the highest income quintile by over 5 ppt.

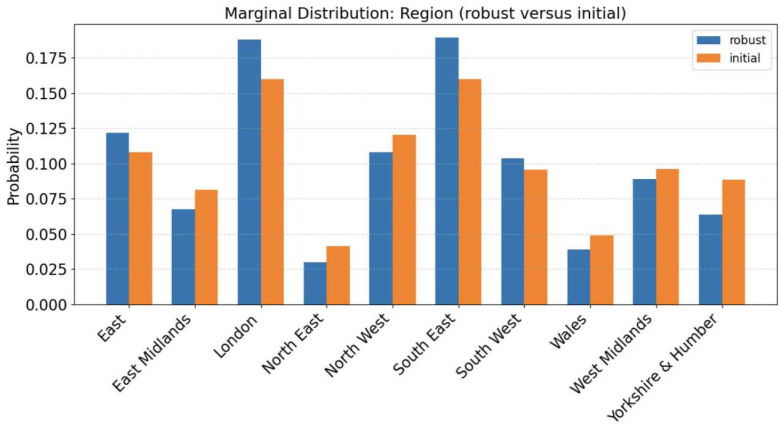

Figure 2.4: Region: UK population (ONS) versus sample

Figure 2.4 illustrates that our imputed region labels skew towards wealthier regions like London, East, and the South. This trend is unsurprising, given the higher income bias observed in our initial dataset.

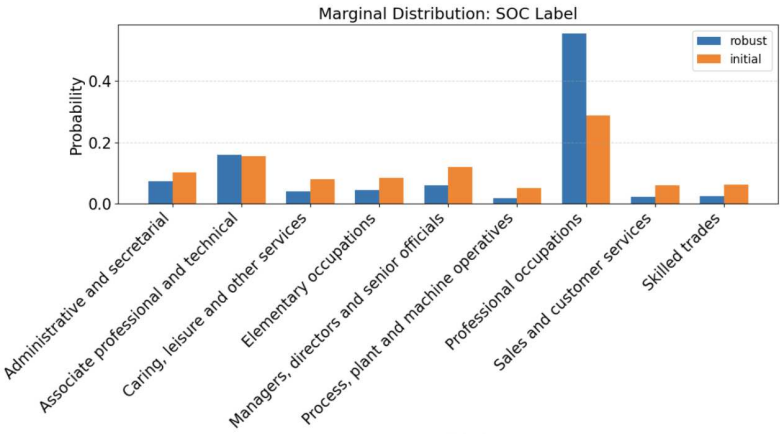

Figure 2.5: SOC label: UK population (ONS) versus sample

Figure 2.5 indicates a strong bias in our imputed SOC labels towards professional occupations. While this is already the modal SOC class in the UK, its representation in our dataset is disproportionately higher, exceeding the baseline by over 20 percentage points. This skew is likely attributable to the influence of both income and regional distributions used for SOC label imputation; it is expected that wealthier demographics in urban centers are more prone to hold professional occupations.

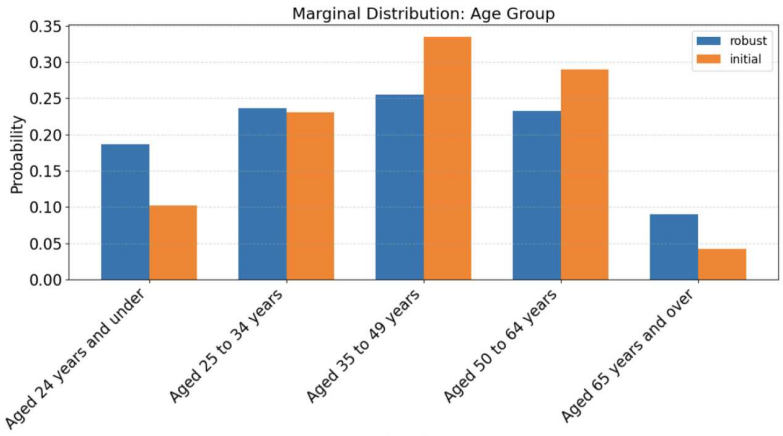

Figure 2.6: Age group: UK population (ONS) versus sample

Figure 2.6 reveals a surprising tilt in our imputed age group labels towards both ends of the age demographic (24 years and under, and 65 years and over). This is unexpected, as one might assume the dataset's higher income distribution would favor the 35-49 age group, typically in their prime earning years. Upon closer inspection, this tilt is primarily driven by $P(\text{DTI}|\text{age group})$. Our initial dataset's average unsecured debt-to-income (DTI) ratio is 0.08, which is 4 percentage points below the FCA national average. Among the five age groups, the DTI ratio is lowest for both the youngest (DTI: 0.06) and the most elderly (DTI: 0.04) cohorts, explaining the observed tilt. The youngest demographic likely hasn't built a sufficiently strong credit profile yet, while the most elderly cohort would have generally paid down outstanding debt.

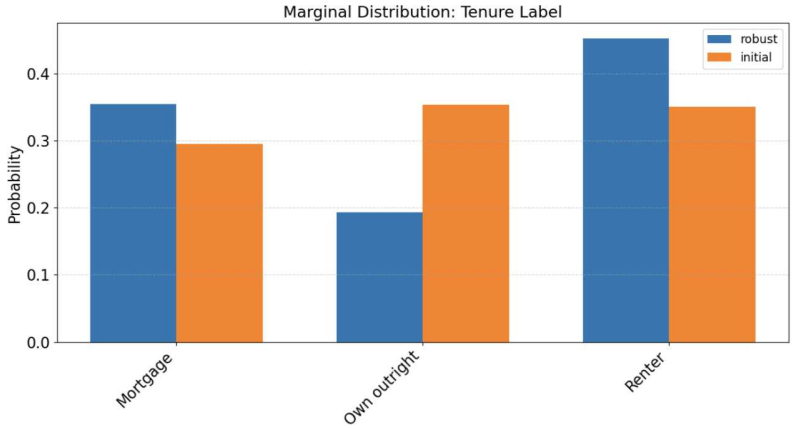

Figure 2.7: Housing tenure: UK population (EHS) versus sample

Figure 2.7 illustrates a significant lean in our imputed **tenure label** towards **renters** (+10 ppt against baseline), which aligns well with the younger demographic prevalent in

our sample. Conversely, the proportion of **outright home owners** sees a notable decrease (close to -15 ppt against baseline), reasonable given both the younger demographic and the lower home ownership rates in large urban centers such as London and the South.

From this analysis, we can sketch a profile of a typical customer in our dataset: they likely **earn above the national average**, work as a **young professional**, and **rent in major urban areas** across the UK.

2.3.2 More Thorough Checks Against Known Conditional Distributions

While checks against known marginal distributions provide some insight, they are insufficient to guarantee the robustness of our imputation. In this subsection, we conduct more thorough checks by comparing the conditional distributions of our dataset against baseline data, providing explanations for any significant deviations.

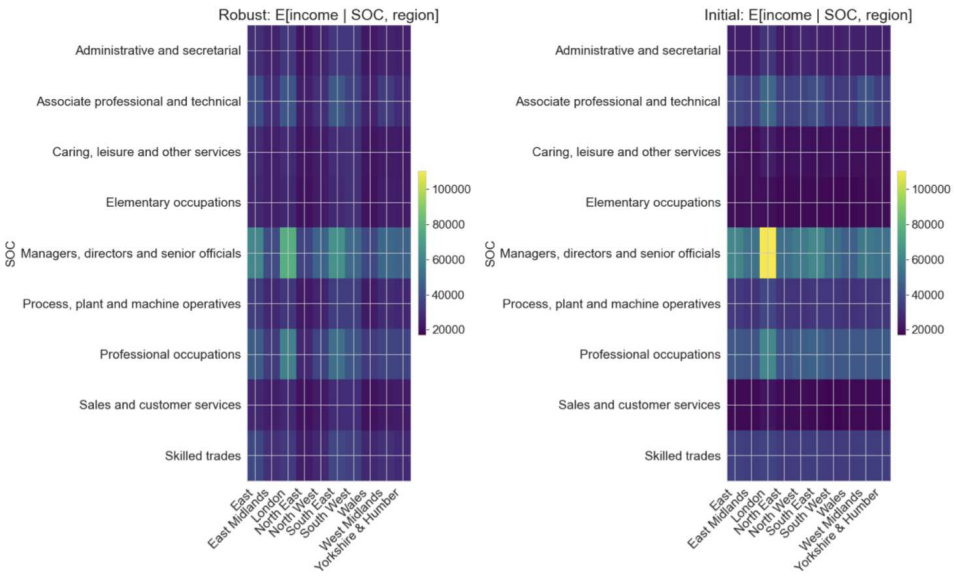

Figure 2.8: Average income given SOC and region: UK population (ONS) versus sample

Figure 2.8 indicates a largely consistent color gradient across both conditional probability matrices, suggesting our model effectively imputes region and SOC labels based on income. A notable deviation exists, however: ONS data shows managers and directors based in London earning over £100,000 per annum, compared to a more modest £80,000 per annum in our dataset. This is expected, given our income data appears to be drawn from a uniform distribution, lacking any heavy tail.

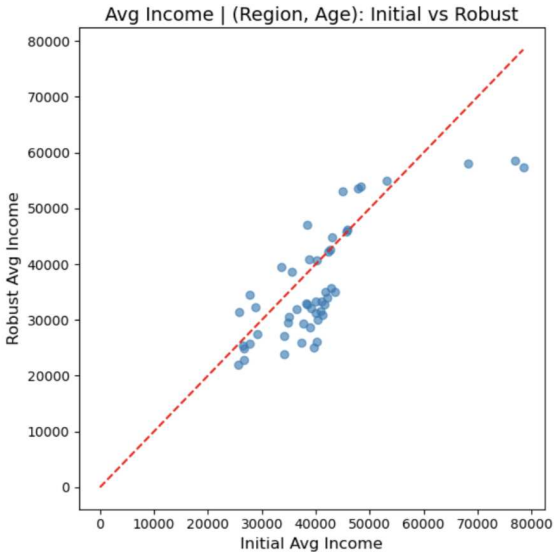

Figure 2.9: Average income given region and age: UK population (ONS) versus sample

Figure 2.9 presents a scatter diagram comparing the conditional mean income given region and age from our dataset against ONS data. The **tight clustering of bubbles along the $y = x$ line** indicates a very good fit, with a **Pearson correlation coefficient of 0.79**. While three bubbles notably deviate from the trend line (showing higher ONS income but lower in our dataset), this aligns with our earlier explanations regarding data characteristics.

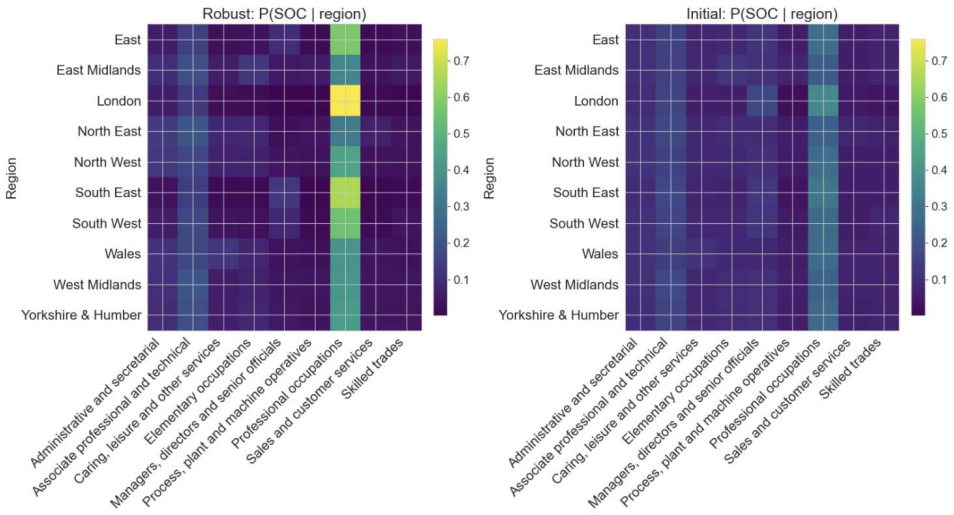

Figure 2.10: Distribution of SOC given region: UK population (ONS) versus sample

Figure 2.10 shows that the color gradient for P(SOC label| region) appears largely consistent between ONS and our dataset, with professional occupations being the sole excep-

tion. This suggests a near-uniform reduction in probability across other SOC labels, with the additional probability channeled towards the professional occupation label.

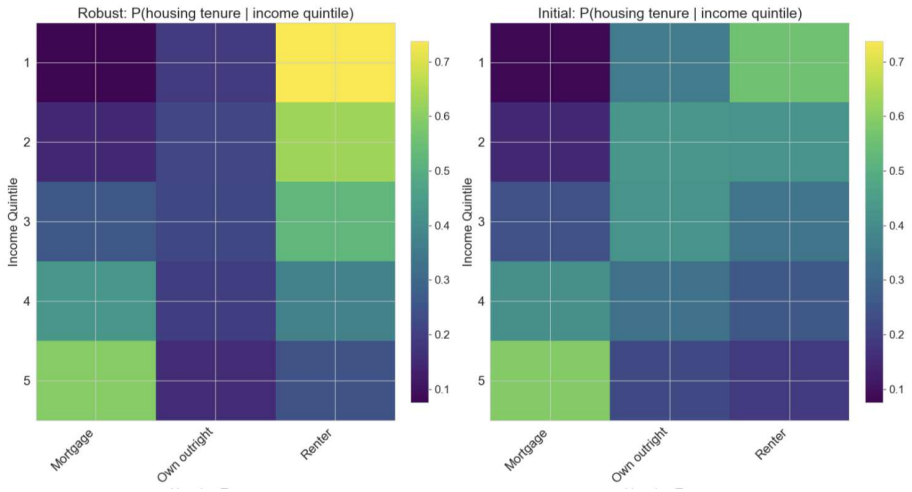

Figure 2.11: Distribution of tenure label given income quintile: UK population (EHS) versus sample

Figure 2.11 shows that, given a fixed income quintile, our dataset assigns fewer 'Own outright' and more 'Renter' labels, consistent with observations of the tenure label marginal. Apart from this distinction, the color gradient across both matrices appears consistent.

We include more such robustness checks in Appendix D.

2.3.3 Checks Against Alternative IPF Approach

As a final sanity check, we repeat the imputation for the tenure label using the iterative proportional fitting (IPF) approach. We then verify that both methods yield similar distributional properties for the imputed labels. More details are available in Appendix E.

Chapter 3

Cash Flow Modeling

3.1 Modeling

With the necessary labels imputed, our next step is to perform month-by-month cash flow modeling to realistically represent a customer's financial position, accounting for housing costs, essential expenditures, debt payments, discretionary spending, cash allocated to savings and investments, bank account balances, investment account balances, and housing asset values over a 12-month period.

3.1.1 Amount Spent on Housing Cost

This section models the proportion of income allocated to housing costs, defined as monthly mortgage payments for homeowners with outstanding mortgages and monthly rent for renters.

The 2022-2023 English Housing Survey [14] provides two key tables:

1. **(Region , Tenure Label)**: Average proportion of income spent on housing and observation frequency for each (region, tenure label) pair.
2. **(Income Quintile , Tenure Label)**: Average proportion of income spent on housing and observation frequency for each (income quintile, tenure label) pair.

We define p as the proportion of income spent on housing. Our prior density is modeled using the (region , tenure) table:

$$P(p|\text{ region}) \sim \text{Beta}(\alpha_r, \beta_r)$$

Here, α_r and β_r are estimated via the method of moments: $\alpha_r = p_r \times N_r$ and $\beta_r = (1 - p_r) \times N_r$. N_r is the total sample size for a region and p_r is the empirical mean share.

Data from the (income quintile , tenure table) serves as our likelihood. We derive an effective sample size N_q and empirical mean share p_q for each income quintile. This is treated as observing N_q Bernoulli trials with unknown success probability p, yielding $k_q = p_q \times N_q$ "successes" (households spending on housing) and $N_q - k_q$ "failures." While continuous spending is not strictly binary, this Binomial approximation is used due to the lack of granular data for direct α, β estimation.

Thus, the likelihood $L(p)$ is $L(p) \sim p^{k_q}(1-p)^{N_q - k_q}$, which is the kernel of a Beta$(k_q + 1, N_q - k_q + 1)$ distribution. Given that the Beta distribution is conjugate to the Binomial likelihood, the posterior distribution remains a Beta distribution:

$$\begin{aligned} P(p|\text{ region, income quintile}) &\propto p^{\alpha_r - 1}(1-p)^{\beta_r - 1} \times p^{k_q}(1-p)^{N_q - k_q} \\ &= p^{\alpha_r + k_q - 1}(1-p)^{N_q + \beta_r - k_q - 1} \end{aligned}$$

Our posterior is therefore Beta$(\alpha_r + k_q, N_q + \beta_r - k_q)$, which is supported on $p \in [0, 1]$, consistent with regional and income conditionals. We sample from this posterior to obtain the proportion of income spent on housing for all 100,000 customers.

3.1.2 Amount Spent on Other Essentials (Ex-Housing)

We now model customer spending on non-housing essentials (food, transport, energy). This amount depends on both income (e.g., wealthier customers' grocery choices) and region (e.g., Londoners possibly spending less on other essentials due to high housing costs).

From the ONS 2022-2023 Family Spending Survey [19], we obtain the average proportion of income spent on other essentials as a multiple of housing costs (denoted as p), allowing estimation of $\mathbb{E}$(p| income decile). Similarly, another ONS Family Spending Survey source [20] provides $\mathbb{E}$(p| region).

Using Iterative Proportional Fitting (IPF)[1], we generate a two-way table for $\mathbb{E}$(p| region, income decile), ensuring it respects all marginals. For each customer[2], this ratio is then multiplied by monthly housing cost to determine the baseline monthly essential spending.

[1] See Appendix E for a thorough discussion of this technique

[2] This works well for customers who are renters or mortgagors, whose housing cost is non-zero. For outright owners whose housing cost is 0, we just impute this amount using the respective (income decile, region) average

3.1.3 Amount Spent on Debt Payment

Monthly debt payments distinguish between **revolving credit** (payment due equals previous month's balance) and **fixed credit** (e.g., STL, with a set monthly amount). Payments within a 5-day grace period incur no interest, while late payments accrue interest based on their respective Annual Percentage Rates (APRs).

To model monthly debt payments, we utilize the following data for each customer i:

1. Time series $X_{i,k,1}, \cdots, X_{i,k,12}$ representing debt incurred across three credit products: credit card 1 (cc1), credit card 2 (cc2), and overdraft.
2. Monthly short-term loan (STL) payment amount, R_i^{STL}.
3. Time series $Y_{i,k,1}, \cdots, Y_{i,k,12}$ indicating on-time (0) or late (1) payment status for four credit products: cc1, cc2, overdraft, and STL.
4. Initial balance for revolving credit $X_{i,k,0}$ for cc1, cc2, and overdraft. (Note: $X_{\text{i,overdraft,0}} = 0$ for all customers initially).
5. Time series $Z_{i,k,1}, \cdots, Z_{i,k,12}$ detailing the number of late days for non-punctual payments across all four credit products.

The total debt payment for customer i in month m is:

$$\text{DebtPaid}_{i,m} = \sum_{k \in \{\text{cc1, cc2, overdraft, STL}\}} [\text{On time component}_{i,k,m} + \text{Late component}_{i,k,m}]$$

where:

$$\text{On time component}_{i,k,m} = \begin{cases} 0, & \text{if customer } i \text{ does not have product } k \\ (1 - \text{flag}_{i,k,m}) \times \text{Base amount}_{i,k,m} & \text{if they have product } k \end{cases}$$

1. $\text{Flag}_{i,k,m} = 0$ if $Z_{i,k,m} \leq 5$, otherwise 1, reflecting the 5-day grace period.
2. $\text{Base amount}_{k,i,m}$ is:
   (a) For STL: the fixed monthly payment R_i^{STL}.
   (b) For revolving credit (cc1, cc2, or overdraft):

$$\text{Base amount}_{i,k,m} = \begin{cases} X_{i,k,0}, & \text{if } m = 1 \text{ and } k \in \{\text{cc1, cc2}\} \\ 0, & \text{if } m = 1 \text{ and } k = \text{overdraft} \\ X_{i,k,m-1} & \text{if } m > 1 \end{cases}$$

3. Late amount$_{k,i,m}$ is calculated as:

$$= \sum_{l=L_{\min}}^{L_{\max}} \left[\mathbb{1}_{\text{late days}_{k,i,m-l+\delta}>5} \times \text{BalanceOrRepay}_{k,i,m-l} \times \left(1 + \text{APR}_{k,i} \frac{\text{late days}_{k,i,m-l+\delta}}{365}\right)\right]$$

where:

(a) For STLs: $\delta = 0$, and BalanceOrRepay$_{\text{STL,k,m-l}} = R_i^{\text{STL}}$.

(b) For revolving credit (cc1, cc2, overdraft): $\delta = +1$, and BalanceOrRepay$_{i,k,m-l} = X_{i,k,m-l}$.

(c) l represents the lookback period, from 2 to 8 for revolving credit and 1 to 7 for STLs.[3]

(d) A term is included only if 'late days' for that month strictly exceeds 5.

(e) In practice, 'late days' are binned into intervals (e.g., $[0, 30, 60, ...]$) to ensure each payment is tabulated uniquely.

3.1.4 Amount Allocated Towards Discretionary Spending, Cash Savings, and Risky Investments

Leveraging prior models for housing costs, essential ex-housing costs, and debt payments, we now model monthly residual cash flow. If positive, this flow is allocated to discretionary spending, savings, and investments. If negative, customers deplete savings or rely on external support.

The monthly cash flow for each customer i is:

$$\begin{aligned}\text{ResidualCashFlow}_{i,m} = \text{Monthly income}_i - \text{Monthly housing cost}_i \\ - \text{Essential spending}_{i,m} + \min(0, \text{New debt drawn}_{i,m} - \text{DebtPaid}_{i,m})\end{aligned}$$

Monthly income is annual net income divided by 12. Monthly housing cost is from Subsection 3.1.1. To introduce stochasticity, Essential spending$_{i,m}$ is sampled from a Gaussian distribution with the mean from Subsection 3.1.3 and a standard deviation of 10 ppt.

The $\min(0, \text{New debt drawn}_{i,m} - \text{DebtPaid}_{i,m})$ term accounts for debt dynamics. If debt paid exceeds new debt drawn, it reduces the available free cash flow for the month, but also results in concurrent reduction of debt liabilities. Conversely, if new debt drawn exceeds debt paid, this is treated as an expense (assuming it covers a shortfall), not as positive

[3] L_{max} is a global maximum, determined by scanning the maximum observed late days across all credit products in the provided data.

residual cash flow available for savings or investment.

Should residual cash flow be negative, a customer's bank account savings are reduced assuming that the balance is being used to cover the shortfall. If savings are depleted, we assume additional external support (state benefits, family/friends) is obtained. While the modeling of such support is outside this dissertation's scope, the negative cash flow is recorded for potential future feature engineering.

For positive residual cash flow, customers first allocate funds to discretionary spending, then split the remainder between cash savings and riskier investments. We model discretionary spending indirectly via the savings rate, through $\text{disc spend}_m = (1 - \text{savings rate}) \times \text{cash flow}_m$. We use Iterative Proportional Fitting (IPF) to create a joint distribution of savings rate given age group and income decile, based on grouped marginals from the 2022 National Institute UK Economic Outlook Box Article [15] and Statista [22].

To introduce heterogeneity, each customer's savings rate is sampled from Gaussian distribution with mean determined by the IPF marginal and standard deviation of 10 %.

Finally, we model the stochastic split between cash deposits and investments using a Dirichlet distribution. The expected proportions, $\mu^{\text{save}}_{\text{inc,age}}$ for savings and $\mu^{\text{invest}}_{\text{inc,age}} = 1 - \mu^{\text{save}}_{\text{inc,age}}$ for investments, follow a glide path reflecting age and wealth-dependent risk appetites[4]. To simulate variability, the allocation proportions $(p_{\text{save}}, p_{\text{invest}})$ are sampled from this Dirichlet distribution, whose concentration parameters are $\alpha_{\text{save}} = \mu^{\text{save}}_{\text{inc,age}}\phi$ and $\alpha_{\text{invest}} = \mu^{\text{invest}}_{\text{inc,age}}\phi$[5]. This sampling is achieved by drawing independent variates $X_{\text{save}} \sim \text{Gamma}(\alpha_{\text{save}}, 1.0)$ and $X_{\text{invest}} \sim \text{Gamma}(\alpha_{\text{invest}}, 1.0)$, then computing $p_{\text{save}} = \frac{X_{\text{save}}}{X_{\text{save}}+X_{\text{invest}}}$ and $p_{\text{invest}} = \frac{X_{\text{invest}}}{X_{\text{save}}+X_{\text{invest}}}$.

3.1.5 Bank Savings and Investment Account Balance

This section models the monthly evolution of savings and investment accounts.

Given $\text{Balance}_{k,m}$ for $k \in \{\text{savings, investment}\}$ and $m \in \{1, 2, \ldots, 12\}$, we first adjust for monthly cash flow impacts. This includes drawing down cash savings for negative

[4] As no ready data exists for risk appetite/investment allocation by age group and income decile ($\mu^{\text{invest}}_{\text{inc,age}}$, $\mu^{\text{save}}_{\text{inc,age}}$), these values are self-supplied to respect the glide path logic. See Appendix B for details.

[5] ϕ is the concentration parameter (pseudo-count) of the Dirichlet distribution; a lower ϕ yields a more diffused distribution, while a higher ϕ results in a more peaked one. We set $\phi = 10.0$ for sufficient heterogeneity.

cash flow and allocating surplus residual cash to savings and investment accounts, resulting in Balance Pre-dynamics$_{k,m+1}$. We then model how these pre-dynamics balances evolve.

We treat the bank account as risk-free, applying a 4% annual interest rate compounded monthly. Risky investments, conversely, follow a **Geometric Brownian Motion** with an annual $\mu = 0.08$ and $\sigma = 0.20$[6].

3.1.6 Housing Asset (For Home Owner and Mortgagors)

Due to the absence of explicit house price data, a detailed imputation strategy was developed for the 100,000-row dataset to establish realistic asset valuations for home owners and mortgagors.

For **mortgagors**, current house price was inferred from their monthly mortgage payments. A standard loan amortization formula ($P = M \times \frac{(1+r)^N - 1}{r(1+r)^N}$, where r is the monthly interest rate and N is the total payments) was inverted to determine the loan principal (P), assuming a **25-year term** at an **annual interest rate of 4%**[7]. Using a **20% down payment ratio**, the original purchase price was estimated from this principal. To estimate the current house price, the number of years the mortgage had been held ('years held') was dynamically simulated. This simulation involved drawing a base age from a normal distribution (mean 32 years, standard deviation 5 years), which was then adjusted based on the customer's income decile (e.g., higher income deciles saw a reduction of 2 years from this base age, while lower deciles added 2 years). The 'years held' was calculated by subtracting this inferred purchase age from the customer's current age, with a cap at 24 years to maintain consistency with the amortization term. Finally, an **annual house price appreciation rate of 3.2%**[8] was then compounded over these 'years held' to arrive at the estimated current house price.

For **outright owners**, house prices were imputed using the **conditional mean of imputed mortgagor house prices** within identical **age group** and **income decile** cohorts. This ensured consistency with comparable market segments. Renters were assigned a house price of zero.

[6]These parameters roughly correspond to the historical time series statistics of UK interest rates and S&P 500 returns.

[7]While we recognize that interest rate for mortgage payment is reset after a set number of years in the UK (5-7 being most common), we take 4% here to simplify our modeling process

[8]Again, a simplification here is applied: 3.2% is the average annual increase in UK housing price index over the last 20 years

Chapter 4

Cluster Analysis

After modeling the monthly cash flow, spending, and saving behavior of our customers in the last chapter, we now extract useful features from these data. These features will be used in subsequent cluster analysis to better understand the typical demographic, financial health, and financial prudence of different client groups.

4.1 Feature Engineering

To generate useful metrics for tailored financial advice, we will generate some features based on our granular modeled cash-flow and asset data.

We sub-divide our features into a few classes: **cash flow**, **financial resources**, **spending**, **savings and investments**, **demographic**, **credit health and capacity**, as well as **credit discipline**. Below we provide a sample of (but not all) relevant metrics from each class.

4.1.1 Cash Flow Features

1. *Average monthly surplus*: A direct measure of financial health, as a higher surplus is likely to be channeled towards savings and investments.

2. *Percent of negative surplus months*: Calculated as the percentage of negative surplus months out of 12. A higher value suggests that a customer might be experiencing financial hardship.

3. *Housing expense ratio*: Calculated as monthly housing cost divided by monthly income, a measure of monthly fixed expense costs.

4. *Surplus-to-income ratio*: A more objective measure of financial health, computed by dividing surplus by income.

4.1.2 Financial Resources Features

1. *Monthly income*: The main financial source which funds all other spending, savings, and investments.

2. *House price*: A proxy of long-term asset which could be used as collateral for additional credit lines.

4.1.3 Spending Features

1. *Average discretionary spend as percentage of income*: Computed by dividing monthly discretionary spending by income, a measure of customer's propensity to consume over to save.

2. *Discretionary spending volatility*: Computed as the standard deviation in discretionary spending over 12 months; a higher value might suggest impulsive spending behavior.

4.1.4 Savings and Investments Features

1. *Average savings rate*: Computed by dividing monthly cash savings allocation by income, averaged over 12 months.

2. *Average investment rate*: Computed by dividing monthly investment allocation by income, averaged over 12 months.

3. *Investment/savings tilt*: Computed by taking the normalized differences of investment rate and savings rate, i.e., $\frac{p_{sav}-p_{inv}}{p_{sav}+p_{inv}}$. This can be viewed as a measure of a customer's risk aversion.

4.1.5 Demographic Features

We use **one-hot encoding** for categorical demographic data. [1]

1. *Tenure dummies*: Three binary (0/1) columns to indicate whether our customer rents, owns a home outright, or is on a housing mortgage.

2. *Age group dummies*: Five binary (0/1) columns to one-hot encode the age group of our customers.

3. *Region dummies*: Ten binary (0/1) columns to one-hot encode the region to which our customers belong.

[1] Typically, for n categories, only $n-1$ binary (0/1) columns are typically required, as the nth class can be inferred from the previous ones, but over here we include all n columns.

4. *SOC dummies*: Similarly, we perform one-hot encoding for the SOC labels of our customers.

4.1.6 Credit Health and Capacity Features

1. *Average net-debt*: Computed by averaging net-debt over 12 months. A higher value might suggest deteriorating financial health.

2. *Debt-to-income ratio*: Computed by taking the amount of total unsecured debt (in month 12) divided by income; a higher value suggests that a customer is highly indebted or leveraged.

3. *Debt volatility*: Computed by taking the standard deviation across 12-month net-debt data. A higher value suggests a highly erratic debt profile, and perhaps higher reliance on credit to finance spending.

4. *Average credit utilization*: A good measure of credit health; a low utilization suggests less reliance on debt and a higher fiscal discipline.

4.1.7 Credit Discipline and Delinquency Features

1. *Percentage of on-time payments*: Computed by taking the number of on-time payments across all credit products, divided by the total number of payments due.

2. *Average days past due*: Computed across all products and monthly payments, providing a measure on severity of delinquency.

3. *Percentage of severe delinquencies*: Fraction of months where payment is more than 30-days late.

4. *Trend in past due days*: Slope of a linear line fitted through 12-month time series for past due days. A high positive value suggests deteriorating financial condition.

5. *Delinquency and discretionary spend correlation*: Pearson correlation coefficient between number of late days and discretionary amount spending. A positive value suggests that impulsive spending behavior is likely the main cause of delinquency.

4.2 Cluster Analysis Workflow

Following feature engineering, our workflow involves scaling numerical features, pruning low-variance features, and selecting the most explanatory features for clustering. Each step is detailed below.

4.2.1 Scaling Numerical Columns

Numerical data is standardized using `sklearn`'s `StandardScaler` (demeaning and dividing by standard deviation). This is crucial for Principal Component Analysis (PCA), which is sensitive to data scale; unscaled variables with larger ranges disproportionately influence results, leading to bias.

4.2.2 Variance Filtering and Collinearity Reduction

To refine our dataset for modeling, we first implement two key feature selection steps. Columns with variance below a threshold of 0.01 are removed, as these near-uniform features are unlikely to contribute effectively to distinguishing customer clusters. Subsequently, to reduce multicollinearity and redundancy, we compute pairwise correlation coefficients. For any pair of features exhibiting an absolute correlation coefficient exceeding 0.9, one of the variables is dropped.[2]

4.2.3 PCA Analysis

PCA is performed on the combined set of binary and processed numerical columns. Principal axes are added incrementally, starting with the largest eigenvector, until a predefined cumulative explained variance threshold is met (e.g., 0.7, 0.8, or 0.9 for comparison)[3]. From each selected principal axis, the top two variables with the highest loadings are chosen to form the final list of features for clustering.

4.3 Results

4.3.1 Number of K-Means Clusters

To select the optimal number of K-means clusters, we use the **silhouette score**[4]. As shown in Figure 4.1, a value of K = 8 yields the highest silhouette score for a variance threshold of 0.7, and the second highest for thresholds of 0.8 and 0.9 (when K is restricted to $\{5, 6, 7, 8\}$). Based on this, we will cluster our customers into **8 distinct groups** using K-means[5].

[2]While this serves as an effective initial sieve, caution is advised as it relies solely on pairwise relationships, potentially overlooking more complex multivariate interactions.

[3]The choice of explained variance threshold in PCA is a common practice to balance dimensionality reduction with information retention. While there is no universally optimal value, thresholds between 70% and 95% are widely cited in data analysis and machine learning literature [13]. The aim is to capture a substantial portion of the original data's variability with a reduced number of dimensions. Testing multiple thresholds (0.7, 0.8, 0.9) allows for a comparative analysis of the resulting clusters, helping to assess the stability and interpretability across different levels of information preservation.

[4]The silhouette score assesses cluster cohesion and separation, ranging from -1 to +1. A score of +1 signifies distinct, well-clustered points, 0 indicates overlap, and -1 suggests misassignment.

[5]It's important to note the trade-off between maximizing the silhouette score and maintaining cluster explainability. Selecting a K that leads to an overly high or low score might result in either insufficient capture of data variability or the inclusion of excessive noise.

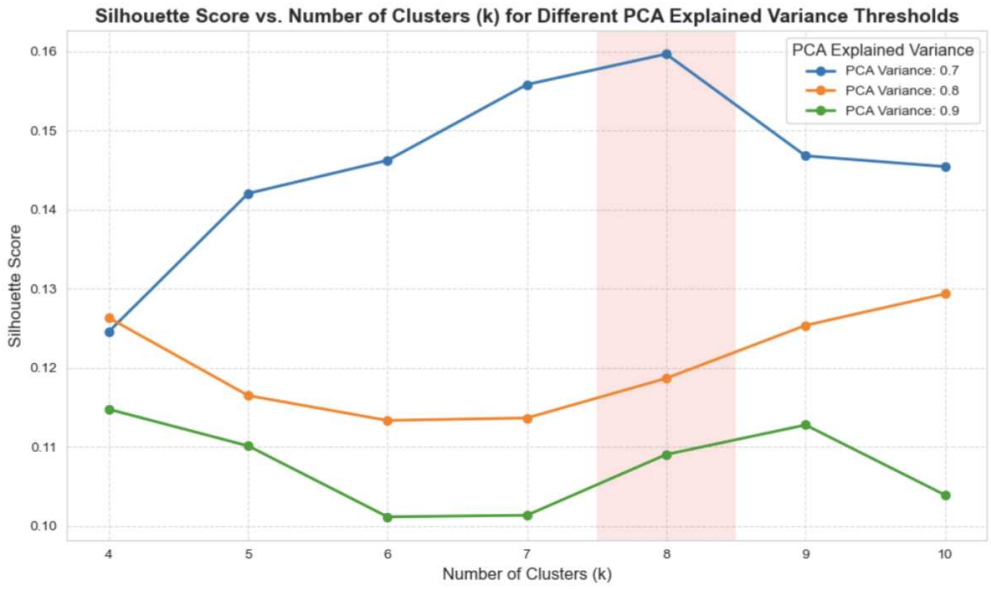

Figure 4.1: Silhouette score vs number of clusters for different variance threshold

A brief discussion on our choice of clustering method (K-means versus agglomerative, spectral, and GMM) is provided in Appendix F.

4.3.2 Characteristics and Typical Persona of Each Cluster

Choosing K-Means clustering with K=8 yields the following clusters:

- **Cluster 0: "Emerging Professionals with Debt" (9,988 customers)**
    1. *Key Demographic Characteristics*: Primarily young adults (under 34), mostly renters. Occupations lean towards professional roles.
    2. *Financial Characteristics*: Moderate income and surplus, but low savings and investments. High total liabilities.
    3. *Fiscal Discipline*: Low with a significantly poor on-time payment rate (0.79) and high average days past due. High high-interest debt and notable financial distress.
    4. *Typical Persona*: Early-career individuals struggling with debt management, facing financial strain despite professional aspirations.
- **Cluster 1: "Disciplined Debt Navigators" (17,125 customers)**
    1. *Key Demographic Characteristics*: Diverse age range (25-64) with a mix of housing tenures. Predominantly professional occupations.

2. *Financial Characteristics*: Good income, strong surplus, and moderate savings/investments. Moderate total liabilities.
3. *Fiscal Discipline*: High with excellent on-time payment rates and very low delinquencies. Effectively manages high-interest debt, maintaining low financial distress.
4. *Typical Persona*: Responsible individuals who skillfully manage their finances and debt, leveraging credit without succumbing to financial strain.

- **Cluster 2: "Established and Secure Homeowners" (13,513 customers)**

1. *Key Demographic Characteristics*: Older demographic (35-64), mostly mortgage holders. Predominantly professional occupations.
2. *Financial Characteristics*: Second highest income (54,067 pounds), robust surplus, and substantial savings/investments. High total liabilities.
3. *Fiscal Discipline*: Very High characterized by excellent payment discipline and minimal high-interest debt. Exhibits low financial distress.
4. *Typical Persona*: Mature, financially stable professionals who are homeowners, focused on long-term wealth accumulation with sound financial practices.

- **Cluster 3: "Vulnerable and Delinquent Renters" (4,580 customers)**

1. *Key Demographic Characteristics*: Youngest age group, overwhelmingly renters. Diverse occupations including clerical and service roles.
2. *Financial Characteristics*: Second lowest income (34,357 pounds) and second lowest initial savings (5,442 pounds). Minimal surplus and investments.
3. *Fiscal Discipline*: Very Low with the absolute lowest on-time payment rate (0.54) and highest average days past due. High serious delinquencies and highest financial distress.
4. *Typical Persona*: Young, low-income renters facing acute financial distress, marked by high debt burden and very poor payment behavior.

- **Cluster 4: "Affluent and Prudent Investors" (11,835 customers)**

1. *Key Demographic Characteristics*: Mature individuals (35-64), mainly homeowners. Overwhelmingly in professional occupations.
2. *Financial Characteristics*: Highest income (68,015 pounds), largest surplus, and leading savings/investments. High total liabilities.

3. *Fiscal Discipline*: Exceptional with impeccable on-time payments and minimal delinquencies. Strategically manages debt with minimal financial distress.
4. *Typical Persona*: High-income, financially astute individuals actively accumulating significant wealth through disciplined saving and investing.

- **Cluster 5: "Struggling to Start" (15,379 customers)**

1. *Key Demographic Characteristics*: Very young adults (under 34), predominantly renters. Higher representation in clerical and elementary occupations.
2. *Financial Characteristics*: Absolute lowest income (24,917 pounds) and absolute lowest initial savings (2,646 pounds). Very low surplus and investments.
3. *Fiscal Discipline*: Maintains high on-time payment rates, but faces the highest severe financial distress due to extremely limited resources.
4. *Typical Persona*: Entry-level earners with minimal financial capacity, often debt-averse but highly vulnerable to financial shocks due to low income.

- **Cluster 6: "Steady and Responsible Managers" (15,519 customers)**

1. *Key Demographic Characteristics*: Balanced distribution across age groups and housing tenures. Predominantly professional occupations.
2. *Financial Characteristics*: Good income, strong surplus, and moderate savings/investments. Moderate total liabilities.
3. *Fiscal Discipline*: Excellent with the highest on-time payment rate and extremely low average days past due. Efficiently manages high-interest debt with low financial distress.
4. *Typical Persona*: Consistently reliable individuals demonstrating strong financial habits and expert management of their debt obligations.

- **Cluster 7: "Crisis-Prone and Underbanked" (12,061 customers)**

1. *Key Demographic Characteristics*: Younger demographic, almost exclusively renters. Higher proportions in elementary and service occupations.
2. *Financial Characteristics*: Negative monthly surplus, lowest savings and investments. Lowest total liabilities.
3. *Fiscal Discipline*: Moderately low. Chronic negative surplus indicates inability to meet obligations. However, on-time payment rate is relatively high (0.94).

4. *Typical Persona*: Individuals in severe, chronic financial distress, consistently spending more than they earn, with virtually no savings or discretionary funds. Limited access to credit (despite high payment on-time rate) could also suggest that this group is underbanked.

4.3.3 Visualization

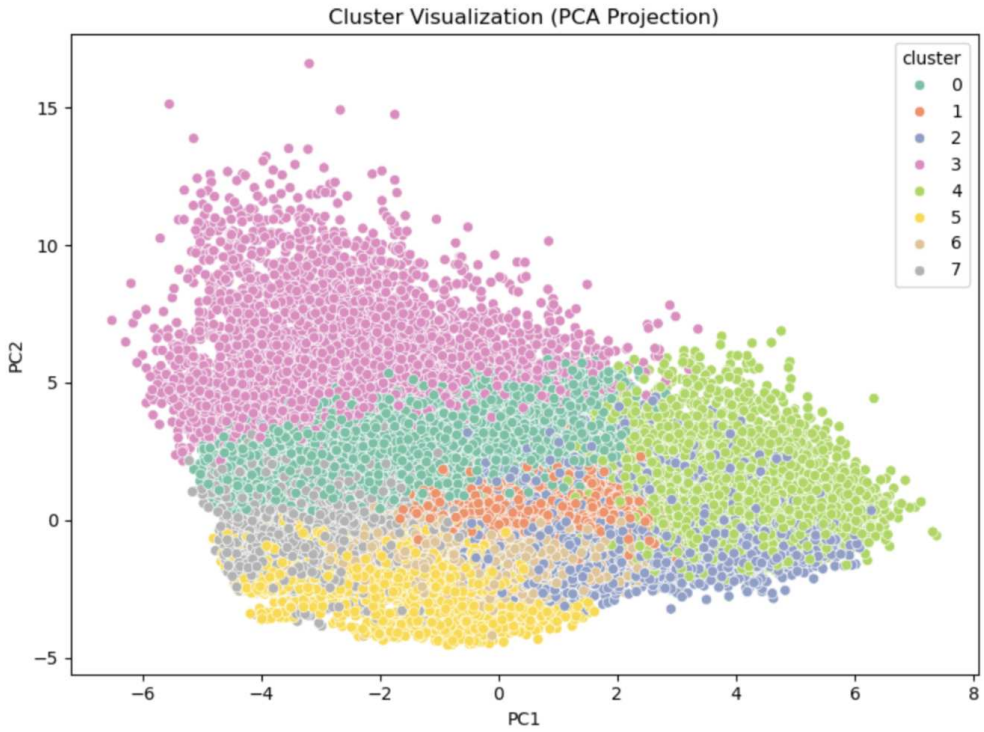

Figure 4.2: Clusters visualization along two top principal axes

Figure 4.2 shows our clusters along the top two principal component axes. We observe that our clustering produces reasonably distinct clusters, though the distance between clusters is small.

Alternatively, we can visualize the clusters along the two axes of **financial prudence** and **wealth**, as per Figure 4.3. For the wealth component, we averaged the scaled values of initial savings, total credit, income, and house price. For the financial prudence component, we included scaled scores of debt-to-income ratio, savings rate, payment on-time rate, and discretionary spending to income ratio.

The graph reveals a positive correlation: financially well-off individuals tend to be more financially prudent. This is perhaps unsurprising, as greater wealth often correlates with

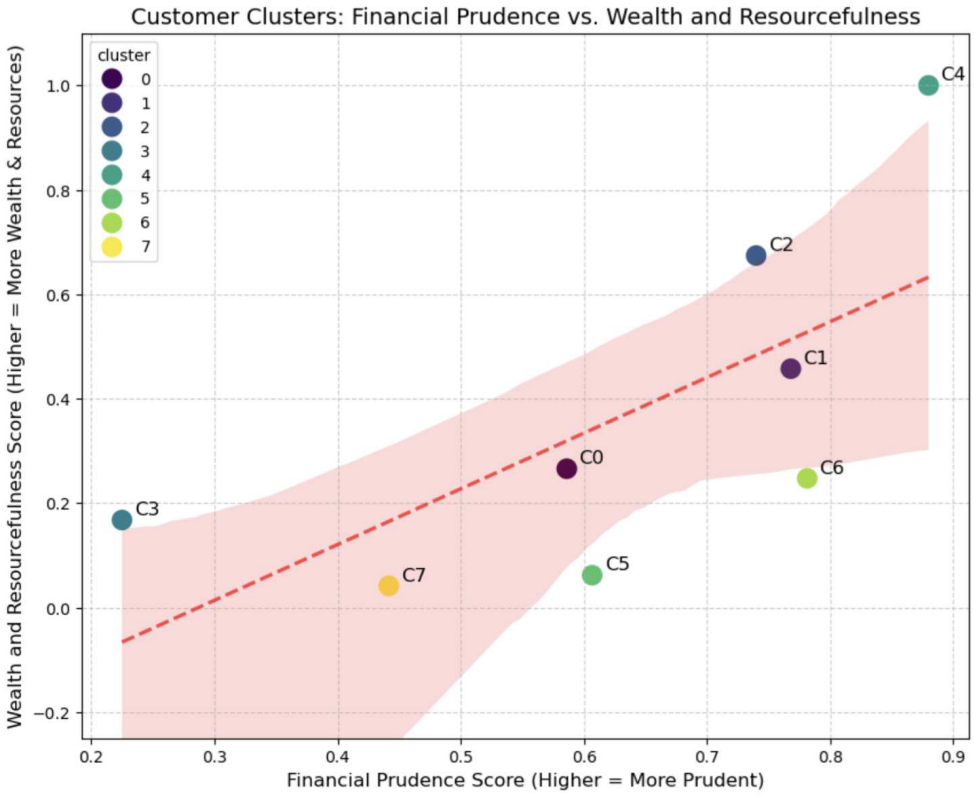

Figure 4.3: Alternative visualization of cluster along axes of financial prudence and wealth access to better financial education.

Chapter 5

Smart Personalised Finance Management

In this chapter, we set out a broad framework by which tailored financial management advice could be offered based on a customer's financial circumstances.

The UK Financial Conduct Authority (FCA) published finalized guidance (FG15/1) in 2015 [10], which aimed to clarify the level of financial recommendation that doesn't constitute 'regulated advice'. The relevant clarification, quoted verbatim, is as follows:

> For advice to be regulated at all, it must relate to a specific investment and must be given to the person in their capacity as an investor or potential investor, or in their capacity as agent for an investor or potential investor, and relate to the merits of them buying, selling, subscribing for or underwriting (or exercising rights to acquire, dispose of or underwrite) the investment. If it does not have all of these characteristics then it is generic advice and is not regulated.

In our dissertation, to adhere to FCA's guidelines, we will be mindful not to offer product-specific investment advice.

5.1 A Framework for Personalized Financial Advice

The earlier chapter suggests a few crucial observations, from which we shall base our personalized financial advice:

1. There is a significant portion of customers struggling with negative monthly cash flow (cluster 7) and little to no savings. For them, the priority would be to start generating positive cash flow and to build up a sufficient savings buffer to weather temporary income shocks.

2. Our clustering results suggest that young renters (clusters 3, 5, 7) are spending a disproportionate amount of their income on housing costs. They might benefit from moving back in with parents, thereby reducing their housing costs.

3. Some customers (clusters 0, 3) are struggling with debt arrears, as well as high unsecured debt-to-income ratio. For them, after establishing a positively monthly cash flow, the priority would be to bring down the arrears balance of high-interest debt.

4. For customers who are already financially secured (positive cash flow with 3 months of savings buffer), our advice will take into account their overall financial health, financial objectives and risk appetite.

   (a) For customers who do not have any unpaid arrears, but is carrying significant revolving debt balance from month-to-month (unsecured debt payment over income ratio of $> 40\%$), we should advise them to gradually reduce the carrying amount as interest would start to immediately accrue if they miss even a single payment. Moreover, a lower DTI helps with building good credit score, which is important for future credit applications.

   (b) Subsequently, if a customer has a healthy DTI and is saving for the down payment of a home, we shall provide a potential trajectory of timeline and an estimate of when the customer is likely to achieve his or her goal.

   (c) Finally, if a customer is not saving for a home, we should provide a fan chart of potential trajectories of their wealth (savings + investments) account over the next 5 years. We will base the savings/investment split on their initial preference (p_{inv} and p_{sav}). [1]

5.2 Decision Tree Sequential Optimization Engine

To implement the suggested framework above, we will employ a sequential optimization approach to guide individuals through a structured financial improvement journey. This process ensures that foundational financial health is established before progressing to more advanced goals. It operates in three distinct layers:

[1] Adhering to a customer's outlined risk preference is the best way to align with FCA's guidance. It is important to recognize that however reasonable our investment advice is, (Eg recommending a portfolio with higher risk to financially well-off customers), it is the customer's own preferences which matter in the end.

Layer 1: Cash Flow Stability Optimization

This initial layer focuses on establishing immediate financial stability and building an emergency savings buffer.

1. **Objective:** To ensure the customer has a positive cash flow and sufficient savings to cover unforeseen expenses.

2. **Process:**

   (a) The system first calculates the customer's overall cash flow after accounting for discretionary spending.

   (b) If cash flow is negative, it identifies potential areas for reduction, such as discretionary spending cuts (up to 100 percent) or rebudgeting essential housing expenses (up to 15 percent). For young renters below 24 years old, it also recommends that they move back with family if possible to cut down on housing costs.

   (c) If, after all possible cuts, cash flow remains negative, the customer is advised to seek professional financial advice, and the optimization process halts for this individual.

   (d) If cash flow turns positive, the system determines the required emergency savings buffer and calculates any shortfall from current savings.

   (e) It then assesses if the customer's potential monthly surplus is sufficient to build this buffer within a 6-month or 12-month target.

3. **Outcome:** If the required buffer is already met, or if a viable plan to build it within the target timeframe exists, Layer 1 is deemed stable. Otherwise, further action or professional advice regarding cash flow is recommended. The available monthly surplus, after ensuring cash flow stability, is then designated as "Monthly Repayment Capacity" for the next layer.

Layer 2: Debt Management & Arrears Optimization

This layer is only activated if Layer 1 (Cash Flow Stability) is deemed stable. It addresses immediate debt obligations and aims to manage revolving debt effectively.

1. **Objective:** To clear any outstanding arrears and ensure a healthy Debt Service Ratio (DSR) by managing revolving credit.

2. **Process:**

(a) The system first identifies any arrears (overdue payments) and calculates the total amount due.

(b) It utilizes the "Monthly Repayment Capacity" and any available savings to pay down these arrears.

(c) If arrears cannot be cleared even with these allocations, professional debt advice is recommended, and the process halts for this individual.

(d) Once arrears are cleared, the system calculates the customer's Debt Service Ratio (the proportion of income used for debt payments).

(e) If the DSR exceeds a defined threshold (e.g., 40%) or significant revolving debt exists, the system calculates the amount needed to improve the DSR.

(f) It then uses the remaining "Monthly Repayment Capacity" and available savings to simulate a paydown strategy over several months (e.g., up to 12 months).

3. **Outcome:** If all arrears are cleared and the DSR is brought within the acceptable threshold, or revolving debt is prudently managed, Layer 2 is stable. Otherwise, further debt management actions or professional advice are advised. The remaining "Monthly Repayment Capacity" is then transferred to the next layer as "Monthly Surplus for Wealth."

Layer 3: Wealth Maximization

This final layer is only activated if both Layer 1 (Cash Flow Stability) and Layer 2 (Debt Management) are deemed stable. It focuses on long-term wealth growth and investment.

1. **Objective:** To help the customer grow their assets and work towards long-term financial goals.

2. **Process:**

(a) The system first checks if the customer has specific savings goals, such as saving for a home.

(b) If such a goal exists and has not been met, the "Monthly Surplus for Wealth" is primarily allocated towards this goal.

(c) If no specific goal exists, or if the goal has been met, and there is a positive "Monthly Surplus for Wealth," the system advises on allocating these funds towards general investments and savings strategies.

(d) It may also suggest further optimizing discretionary spending to accelerate wealth accumulation.

3. **Outcome:** If there is a surplus available and allocated for wealth building, Layer 3 is considered stable, providing recommendations for continued financial growth.

This sequential, layered approach ensures that critical financial issues are addressed in a logical order, building a strong financial foundation before moving on to less urgent, but equally important, wealth-building objectives.

5.3 Program Output and Robustness Checks

5.3.1 Program Output

Our program generates a report for each customer, detailing measures which they could take for better financial health. It does not explicitly recommend an investment portfolio, but merely simulates asset evolution based on a customer's already established investment risk preference. A sample report with associated output diagrams is available in Appendix A.

5.3.2 Robustness Checks

In this section, we briefly checks the impact of our layer 1 (cash flow stability) advice across different K-means clusters.

Figure 5.1: Recession buffer gap (initial versus post savings advice) across clusters

Figure 5.1 reveals general improvement in the savings buffer (where positive values indicate a shortfall and negative values indicate a surplus). The most significant reduction is

evident in Cluster 5 (struggling to start young renters), with its median buffer gap reaching zero post-advice. Likewise, Cluster 6 (steady and responsible managers) achieved a zero buffer gap at its 75th percentile. This strong improvement in Clusters 5 (315 pounds) and 6 (296 pounds) aligns with their higher financial prudence scores (Figure 4.3), suggesting their inherent financial discipline enabled more effective adaptation of the advice and accelerated gap reduction.

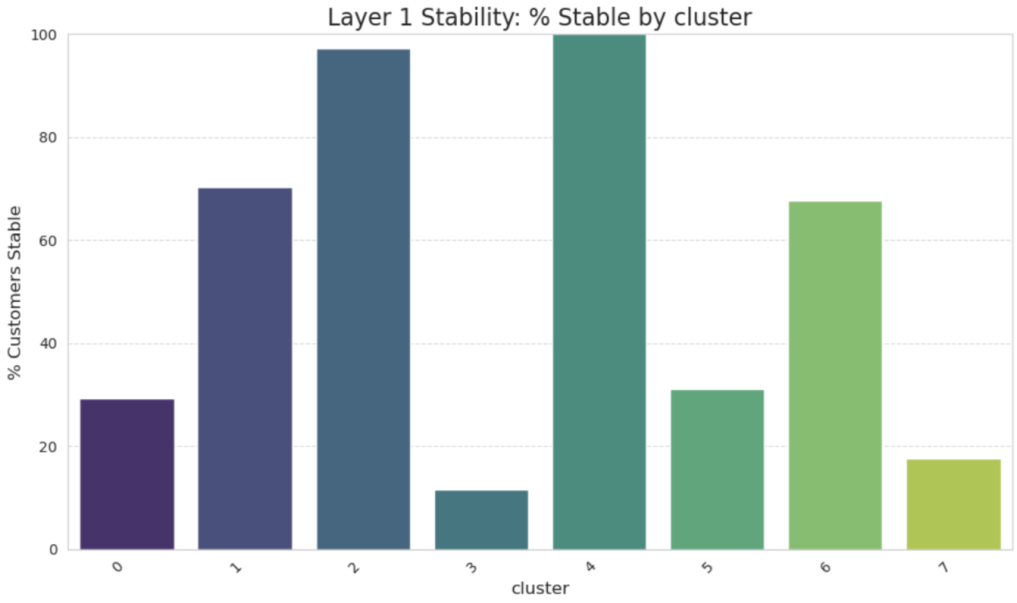

Figure 5.2: Layer 1 stability by K-mean clusters

Figure 5.2 illustrates the varying success of each cluster in achieving cash flow stability. Predictably, wealthier clusters (1, 2, 4, 6) exhibit a higher propensity for stable monthly cash flow and adequate post-advice savings buffers. Notably, among clusters 0, 3, 5, and 7—which possess comparable wealth and financial resources—financial prudence emerges as a critical determinant, with Clusters 0 and 5, having higher prudence scores, demonstrating superior stability in cash flow and savings buffers.

This similar trend could be observed for layer 2 and 3 stability, analysis of which we omit here.

Chapter 6

Conclusions

6.1 Recap of Key Findings and Contributions

This dissertation initiated a thorough analysis of customer financial data, commencing with the enrichment of our dataset through the imputation of supplementary labels. Following this, we developed a preliminary month-to-month cash flow model for all 100,000 customers. Notably, the imputation of these multiple labels was achieved using Bayesian inference methods, which, while not a novel technique, effectively demonstrates how real-world distribution data can be leveraged for robust data enrichment.

Another contribution of this work was to provide a proof-of-concept demonstrating the feasibility of delivering personalized financial management advice under the open finance framework. This advice framework was tailored to individual customer financial circumstances, encompassing their risk preferences, discretionary spending habits, debt situations, and financial objectives.

6.2 Scope for Future Work

The current sequential optimization engine primarily operates on static data, derived from averaging 12 months of time series information, largely due to limitations in the temporal data available. A significant avenue for future work involves leveraging richer datasets that span multiple time periods. Such data would enable the application of more advanced time series models, including ARMA, (G)ARCH, and t-GAN, facilitating the provision of real-time monthly financial advice, rather than mere snapshots.

The current optimization framework operates under a predetermined hierarchy, prioritizing the achievement of positive cash flow before addressing debt reduction and subsequently wealth maximization. This hierarchical approach generally aligns with established financial

recommendation models and principles akin to Maslow's Hierarchy of Needs[1]. However, a potential enhancement for future iterations could involve allowing users to customize their financial priorities (e.g., opting for wealth maximization even with a high debt-to-income ratio), provided they are fully informed of the associated financial consequences.

Finally, the extensive availability of transaction-level data facilitated by the open finance infrastructure removes the traditional constraints on financial advisory objectives, allowing us to move beyond conventional goals like debt management and wealth maximization. This rich data environment is particularly well-suited to tackle more sophisticated objectives, such as comprehensive estate and tax planning.

[1]Further details on this alignment can be found in Appendix E

Appendix A

Sample Advice Report

In this section we provide a sample report which is generated by our program.

```
================================================================================
FINANCIAL ADVICE REPORT FOR CUSTOMER ID: 30959
================================================================================

--- Customer Profile ---
Monthly Income: £6180.13
Housing Cost: £798.92
Essential Cost: £2613.11
Age Group: Aged 50 to 64 years
Housing Tenure: Mortgage
SOC Group: Professional occupations
Region: North West
Income Quintile: 5
Initial Savings Account Balance: £60512.89
Initial Investment Account Balance: £8003.91

--- Financial Optimization Recommendations ---

Layer 1: Cash Flow Stability
--------------------------------------------------------------------------------
Initial overall cash flow is stable and positive.
Emergency savings buffer is already sufficient.
Estimated Savings Post-Advice (Layer 1): £58203.88

Layer 2: Debt Management & Arrears Optimization
--------------------------------------------------------------------------------

 No current debts in arrears.
 - Your average monthly debt service ratio is 93.86% (Target: 40%).
 You have outstanding revolving debt totaling £4525.65 across credit cards and overdra
fts.

**❗ Important Note on Revolving Debt:**While your current debt balances may not incu
r explicit interest charges if payments due are covered by new debt, it's crucial to u
nderstand the true financial implications.If you were to miss even a single payment, s
ignificant interest charges would immediately kick in. Continuously carrying high revo
lving debt, even if paid off promptly through new borrowings, can negatively impact yo
ur credit score and demonstrates an unsustainable financial cycle. Reducing the princi
pal balance is the ultimate goal for improving your financial health.

**🎯 Action: Targeted Revolving Debt Paydown.**
 To bring your Debt Service Ratio closer to 40%, we recommend clearing approximately £
3328.70 of your outstanding revolving debt balance.
 - Used £1019.69 from this month's remaining surplus capacity to reduce this target am
ount.
 - Used £2309.01 from available savings (above buffer) to reduce this target amount.
```

Figure A.1: Generated personalised financial advice report (part 1)

```
**✅ Outcome:** The targeted revolving debt amount could be cleared within 0 months!
Your Debt Service Ratio would then be approximately 40.00% (down from 93.86%).

Layer 3: Wealth Maximization
--------------------------------------------------------------------------------
Your current total wealth (cash + investments) is approximately £66207.78.
You have a monthly surplus of £1019.69 available for wealth building.

**📈 General Wealth Growth Strategy:**
 **⛏ Action:** Allocate your monthly surplus of £1019.69 as follows:
 - **£556.05 to Investments** (approximately 55%)
 - **£463.64 to Savings** (approximately 45%)

**Smart Investment Choices:**
 - Prioritize Tax-Advantaged Accounts: Maximize contributions to vehicles like ISAs (U
K) or 401(k)/IRAs (US) to grow your wealth tax-efficiently.
 - Diversify: Invest in broad market index funds or ETFs to spread risk across many co
mpanies/assets.
 - Long-Term Perspective: Wealth building is a marathon, not a sprint. Stay invested t
hrough market fluctuations.
 - Review Regularly: Rebalance your portfolio periodically and adjust your strategy as
your goals or risk tolerance change.

---
**🚀 Accelerate Your Goals!**
You currently have approximately £1607.39 of additional monthly discretionary spending
available (beyond cuts already recommended for cash flow stability).
By voluntarily reducing this amount, you could significantly speed up your debt reduct
ion or wealth accumulation journey.
Consider what discretionary expenses you could further optimize (e.g., cutting non-ess
ential subscriptions, reducing dining out, etc.) to free up more funds for your financ
ial goals.

--- Recession Impact Analysis (at 99% Confidence) ---
(Based on 1000 simulations for a general unemployment rate of 10.0%)

Mean Total Shortfall over 3 months: £245.67
Maximum Total Shortfall over 3 months: £10236.11
Required Savings for 99% Sustainability: £10236.11

Initial Savings Conclusion:
✅ Initial savings are likely sufficient to withstand income drop (at 99% confidenc
e).

Savings After Advice Conclusion:
✅ Savings after advice are likely sufficient to withstand income drop (at 99% confid
ence).
```

Figure A.2: Generated personalised financial advice report (part 2)

Figure A.3 shows a savings trajectory which is generated by our program for a first-time home buyer putting aside monthly surplus towards his housing deposit. The interest is compounded monthly at an annual interest rate of 4 %.

Figure A.4 shows a fan chart, where 1000 Monte-Carlo simulations were performed over a time horizon of 5 years. The purple region represents the 25-75 percentile range of potential wealth trajectories over 5 years. The chart is generated for customers who are not saving up for a housing deposit but had instead chosen to maximize their risk-adjusted wealth.

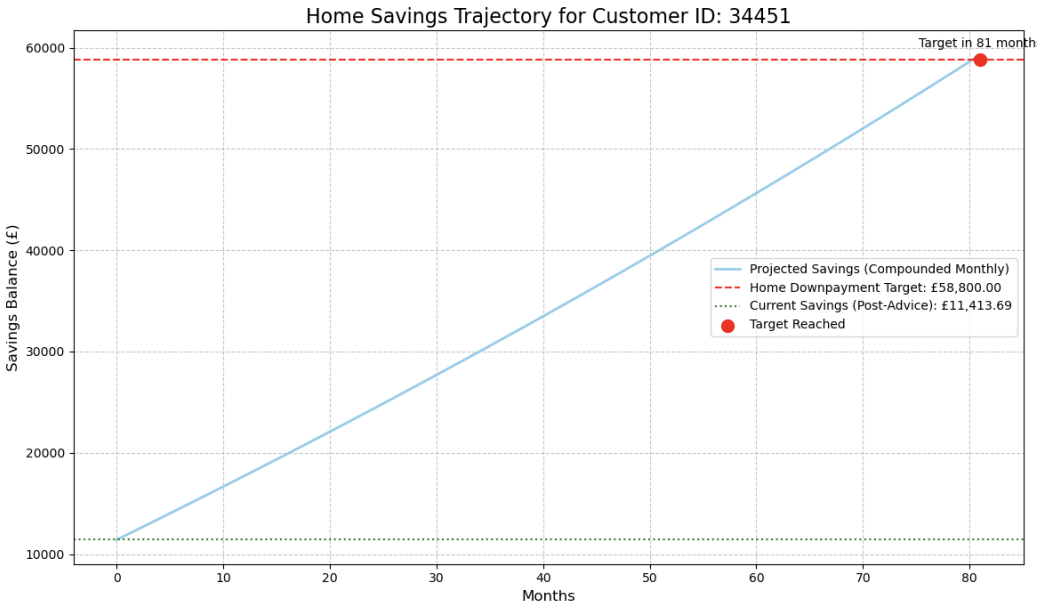

Figure A.3: Sample savings trajectory for housing deposit

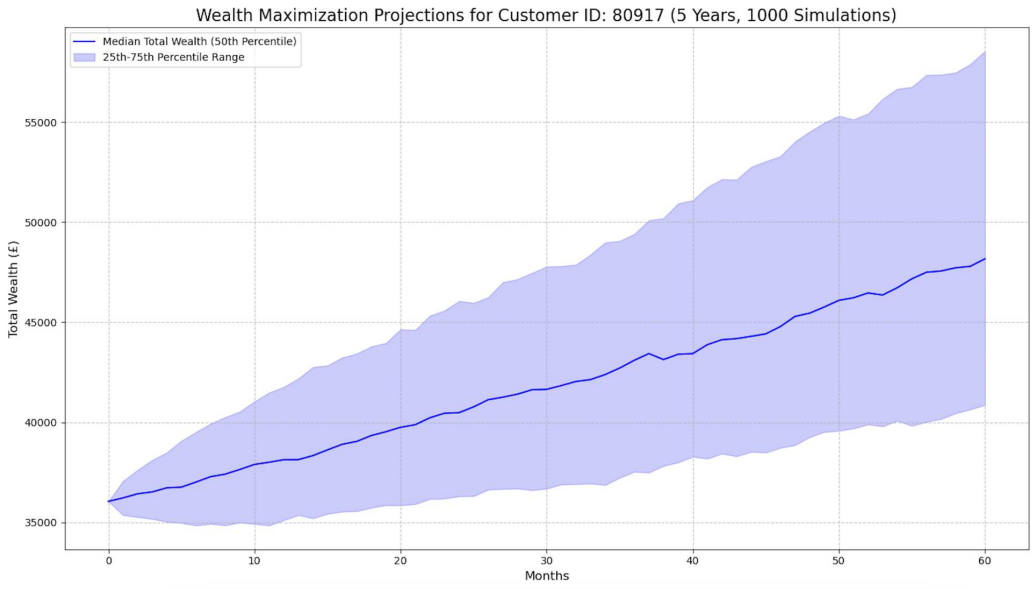

Figure A.4: Sample wealth maximisation projection fan chart

Appendix B

Income and Age-based Risk Appetite Glide Path

As mentioned in the main text, we do not have ready access to investment preferences broken down along the axes of age groups and income deciles.

One potential inference we can make is that middle-aged individuals often show a higher propensity for equity investment, benefiting from long investment horizons that allow recovery from market downturns and potential capital gains tax benefits. In contrast, those nearing retirement typically favor conservative cash holdings for capital preservation, while the youngest cohort may lack market access or experience.

It also stands to reason that people from higher income deciles have a higher risk appetite, and so could invest more of their wealth in risky assets.

Given all the information above, we can furnish a potential glide path for investment allocation towards risk-free cash deposits, broken along the axes of age group and income decile.

	1	2	3	4	5
24 and under	0.85	0.80	0.75	0.70	0.65
25 to 34	0.80	0.75	0.70	0.65	0.60
35 to 49	0.75	0.70	0.65	0.60	0.55
50 to 64	0.70	0.65	0.60	0.55	0.50
65 and above	0.80	0.75	0.70	0.65	0.60

Table B.1: Income and age-based risk appetite glide path for investment allocation towards risk-free cash deposits

Appendix C

Maslow's Hierarchy of Financial Needs

When applied to personal finance, Maslow's Hierarchy of Needs [1]illustrates a natural progression of financial priorities. This framework guides individuals from ensuring basic financial survival towards achieving wealth accumulation, financial independence, and ultimately, financial abundance. It effectively highlights how financial objectives evolve as one attains greater security and well-being.

Below is a breakdown of how each level of Maslow's Hierarchy translates to personal finance:

1. Physiological Needs: Financial Survival

At the 'Physiological Needs' level, the primary goal is to achieve **Financial Survival**. This involves covering basic living expenses such as housing, food, and transportation. Key financial actions include effective budgeting to ensure positive cash flow for meeting day-to-day financial obligations.

2. Safety Needs: Financial Safety

Ascending to 'Safety Needs' involves establishing **Financial Safety**. It focuses on building a robust financial safety net, primarily by creating an emergency fund and paying down high-interest debt. Adequate insurance coverage (e.g., health, home, auto) is also vital at this stage in safeguarding against unforeseen financial shocks.

[1]A lot of the material here (including the illustration) is directly inspired from the 'Hierarchy on Financial Needs' published by New York Life Investments [24]

Figure C.1: Hierarchy of Financial Needs

3. Love and Belonging Needs: Financial Independence

The 'Love and Belonging Needs' level, in a financial context can be interpreted as **Financial Independence**. This stage is about accumulating assets which generate sufficient income to cover living expenses, thereby eventually reducing reliance on active earnings. This stage includes strategic investment planning for both short-term goals and long-term objectives such as retirement planning.

4. Esteem Needs: Financial Freedom

Reaching the 'Esteem Needs' stage corresponds to achieving **Financial Freedom**. This implies having enough financial resources to live life doing what one loves (pursuing passions,

hobbies) and giving back to the community, all while maintaining the lifestyle one desires. Comprehensive legacy planning (estate and tax) also comes into play at this stage, to ensure individuals could leave a positive legacy by bequeathing part of their wealth to charitable organizations.

5. Self-Actualization: Financial Abundance

The pinnacle of Maslow's hierarchy, 'Self-Actualization', aligns with **Financial Abundance**. This signifies a state where accumulated assets significantly surpass expenses, affording a high quality of life and the liberty to pursue personal fulfillment. At this level, financial resources are actively leveraged to pursue creative endeavors, align financial goals with personal values, and to make substantial, purposeful contributions to society.

Appendix D

More Robustness Checks on Synthetic Imputations

Here we include more robustness checks which were omitted in the main text.

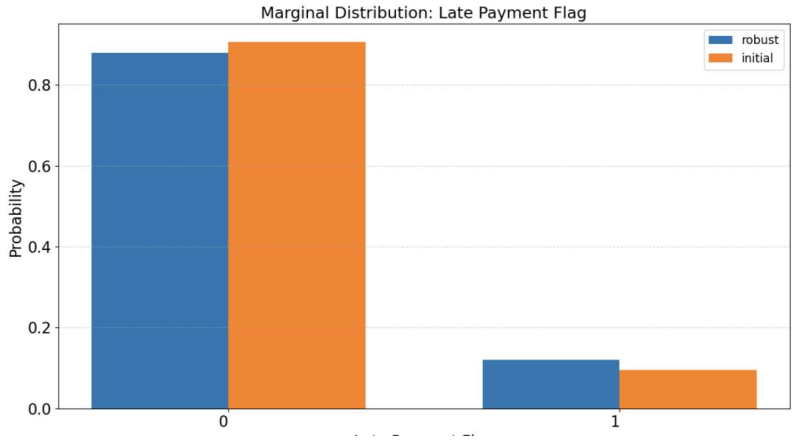

Figure D.1: Late payment flag: UK population (FCA) versus sample

Figures D.1, D.2 and D.3 provide justification for selecting the threshold of **3 late monthly payments across any credit products (cc1, cc2, overdraft, short-term loan)** as the criteria for satisfying the late payment flag. Under this heuristic, Figure D.1 shows that the marginal distribution for the late payment flag is roughly the same as what we could expect from FCA's distribution.[1]. Figure D.2 and D.3 shows that both the conditional marginals for late payment flag given age group and tenure label of our data are also roughly in line with that of FCA's distribution.

To inspect the modeling quality of the average proportion of income spent on housing

[1]One could argue that we should not just match FCA's marginal blindly since our data is biased (higher average income and lower debt-to-income ratio) anyway. However, in the absence of additional info, trying to match FCA's marginal is a good starting point

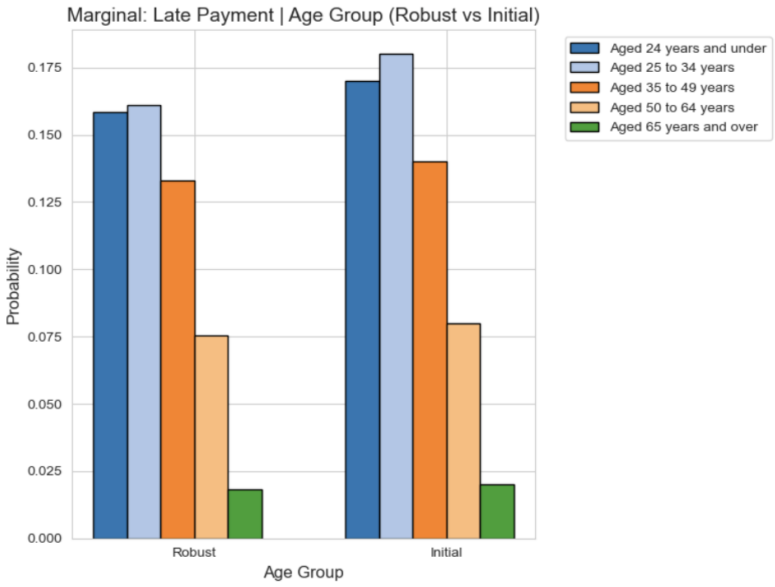

Figure D.2: Late payment flag by age group: UK population (FCA) versus sample

costs, we compare our sample data against figures from the Office for National Statistics (ONS). As shown in Figure D.4, a scatter diagram plots the average proportion of income spent on housing costs for our sample against ONS's average, segmented by region and housing tenure. While the fit is not perfectly aligned, a high Pearson correlation coefficient of 0.89 demonstrates a strong positive linear relationship between our sample's data and the ONS's. This indicates a robust consistency in the modeled housing cost proportion across various regional and tenure groups.

Similarly, Figure D.5 further reinforces this validation by comparing the average proportion of income spent on housing costs, segmented by income quintile and housing tenure, between our sample and ONS data. This comparison also reveals a very strong linear relationship, evidenced by a **Pearson correlation coefficient of 0.95**.

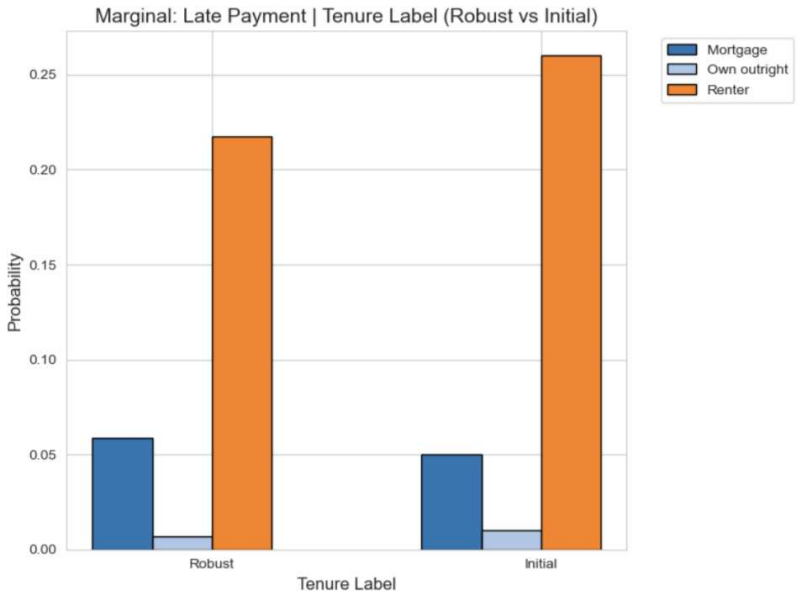

Figure D.3: Late payment flag by housing tenure: UK population (FCA) versus sample

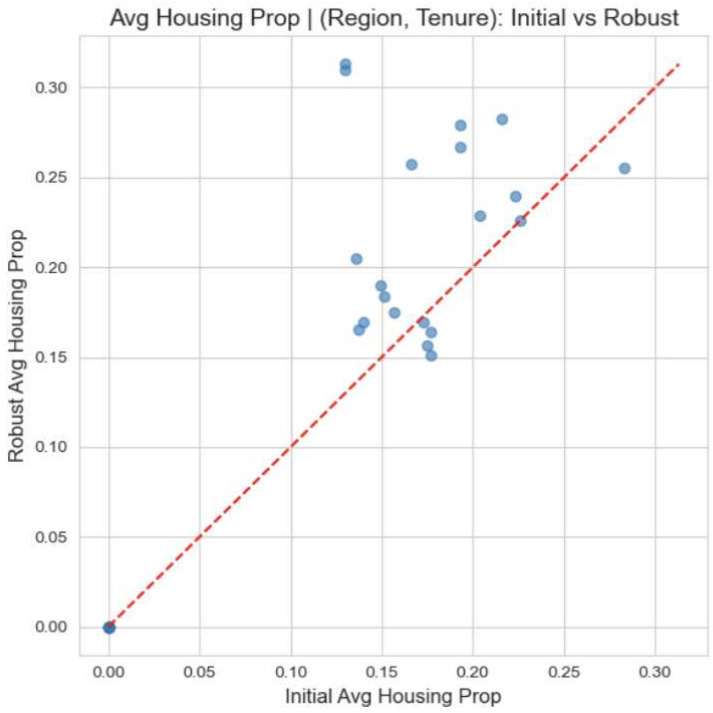

Figure D.4: Average proportion of income spent on housing by region and housing tenure: UK population (ONS) versus sample

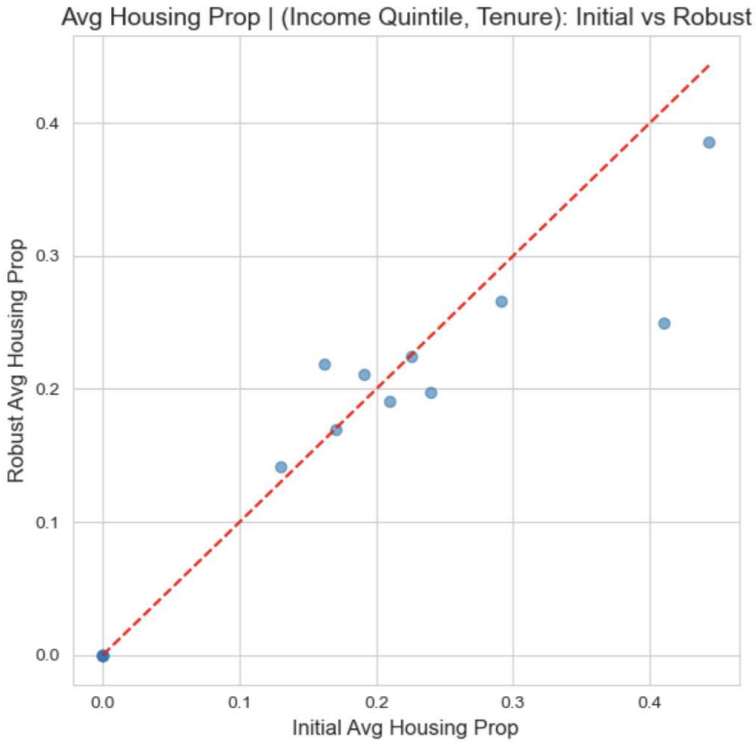

Figure D.5: Average proportion of income spent on housing by income quintile and housing tenure: UK population (ONS) versus sample

Appendix E

The IPF Approach

E.1 Introduction

The **Iterative Proportional Fitting (IPF)** method is a widely used statistical technique for adjusting a multi-dimensional table (e.g., a joint distribution or a contingency table) to match known marginal totals. It operates by iteratively scaling the entries of an initial table to satisfy the given marginal constraints, converging to a unique solution under certain conditions. [1]

This technique is particularly valuable when we possess two distinct marginal distributions or expectations for a set of variables and aim to construct a coherent joint distribution that is consistent with both. The method allows for flexibility in its application: if the input marginals represent probability distributions, IPF will generate a joint probability distribution that respects these, preserving the probabilistic interpretation. Alternatively, if the marginals are given as expectations or means, IPF can be adapted to produce a joint distribution whose marginal expectations align with the given constraints. This adaptability makes IPF a powerful tool for integrating information from various sources to infer a consistent underlying joint structure.

The core IPF algorithm operates iteratively:

1. Initialize a base joint distribution or matrix (e.g., uniform or from prior knowledge).
2. Adjust the current table entries proportionally to match the first set of marginal totals (e.g., row sums).

[1] Convergence for IPF is generally guaranteed when all marginal totals are positive and mutually consistent, and a solution consistent with these marginals exists. Issues can arise with zero marginals or highly sparse initial tables.

3. Adjust the resulting table entries proportionally to match the second set of marginal totals (e.g., column sums).

4. Repeat steps 2 and 3 until the table converges, meaning its marginals sufficiently match the target marginals.

Despite its utility, IPF has certain limitations. For very large or sparse tables, computational cost and convergence speed can be a concern. More critically, IPF implicitly assumes a log-linear relationship between the variables. This implies it generates the "least informative" or "maximum entropy" distribution consistent with the given marginals, often assuming conditional independence not explicitly captured by the marginals. This assumption may not always reflect the true underlying dependencies within the data.

E.2 Application of IPF in This Dissertation

1. In Chapter 3.1.2, given $\mathbb{E}(p|\text{income decile})$ and $\mathbb{E}(p|\text{region})$, we used IPF to generate $\mathbb{E}(p|\text{region, income decile})$, where p represents proportion of income spent on exhousing essentials.

2. In Chapter 3.1.4, given $\mathbb{E}(\text{savings rate}|\text{income decile})$ and $\mathbb{E}(\text{savings rate}|\text{age group})$, we used IPF to generate $\mathbb{E}(\text{savings rate}|\text{age group, income decile})$

3. In Chapter 2.3.3, as an additional robustness check, we use IPF to impute the housing tenure label and verify that both Bayesian inference and IPF generate imputed labels with similar distributional properties. See the next section for more details.

E.3 Alternative Robustness Check: Using IPF to Impute Housing Tenure Label

Following the imputation of region, SOC code, and age group labels via Bayesian inference (as detailed in Chapter 2.2), the housing tenure label was subsequently imputed. This imputation relied on a Naive Bayes independence assumption, expressed as:

$$P(\text{tenure}|\text{age group, income quintile, region, DTI, late flag})$$

$$
\begin{aligned}
&\propto P(\text{tenure}) \times P(\text{age group}|\text{tenure}) \\
&\times P(\text{income quintile}|\text{tenure}) \\
&\times P(\text{region}|\text{tenure}) \times P(\text{DTI}|\text{tenure}) \\
&\times P(\text{late flag}|\text{tenure})
\end{aligned}
$$

As an additional robustness check, we now employ a hybrid probabilistic approach instead. First, Iterative Proportional Fitting (IPF) was utilized to construct a four-way joint probability distribution of (tenure, age group, income quintile, region). This IPF step was constrained by three bivariate marginal distributions: $P(\text{tenure}, \text{age group})$, $P(\text{tenure}, \text{income quintile})$, and $P(\text{tenure}, \text{region})$. The resulting four-way joint distribution served as a base probability. Subsequently, this base probability was multiplied by the likelihoods of Debt-to-Income (DTI) and late payment flag ($P(\text{DTI}|\text{tenure})$ and $P(\text{late flag}|\text{tenure})$) to compute the final posterior probability for each tenure label. The ultimate tenure label for each customer was then sampled from this calculated posterior distribution.

This constructed joint distribution inherently respects all these specified bivariate marginals, offering a different perspective on the relationships compared to the Naive Bayes assumption.

As can be seen, figures E.1 and E.2 indeed demonstrate the remarkable similarity of the conditional distributions derived from both the pure Bayesian and hybrid IPF approaches, highlighting the robustness of our methods.

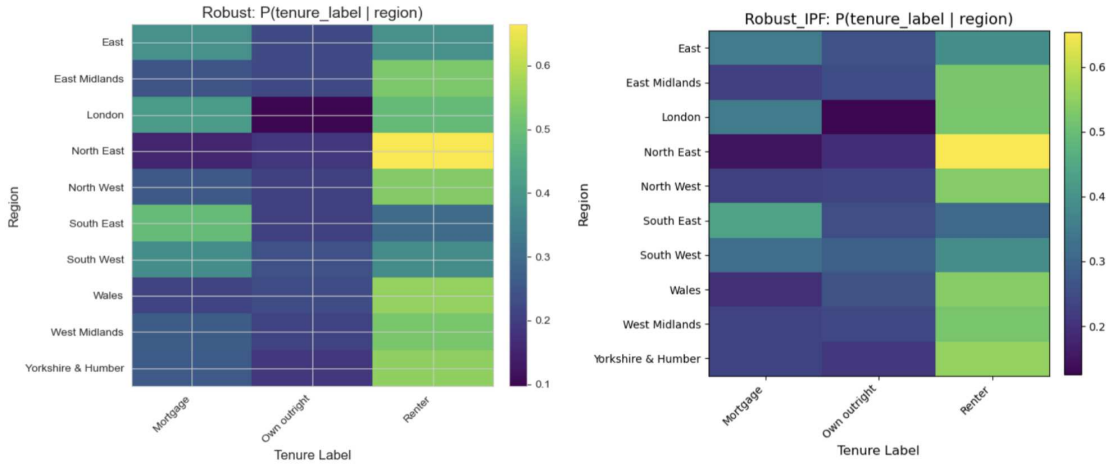

Figure E.1: Conditional distribution of housing tenure given region: Bayesian versus IPF

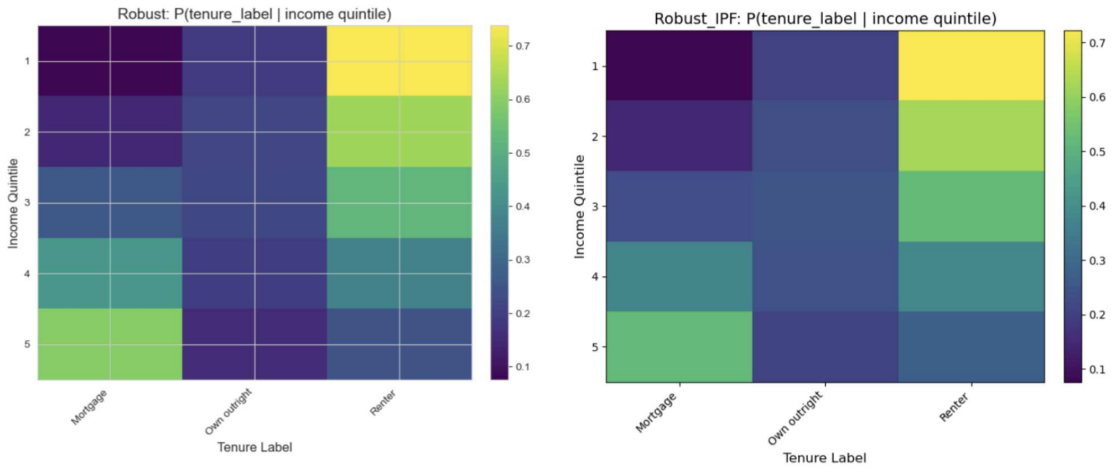

Figure E.2: Conditional distribution of housing tenure given income quintile: Bayesian versus IPF

Appendix F

Why K-Means?

F.1 Overview of Clustering Methods and Computational Considerations

This appendix details the rationale behind selecting K-means for our customer segmentation, following a comparison with other common clustering algorithms. When choosing a clustering method for a dataset of 100,000 data points with over 20 features, computational cost and scalability are significant concerns.

- **K-Means Clustering**: A centroid-based, partitional clustering algorithm that partitions n observations into k clusters, where each observation belongs to the cluster with the nearest mean (centroid).
  - *Computational Cost*: Relatively efficient, typically $O(nkdI)$ where n is the number of data points, k is the number of clusters, d is the number of features, and I is the number of iterations. Its speed makes it a strong contender for large datasets.
  - *Downsides*: Requires pre-specification of k, sensitive to initial centroid placement, and assumes spherical, equally sized clusters, struggling with non-convex or irregularly shaped clusters.
- **Agglomerative Clustering**: A hierarchical clustering method that builds a hierarchy of clusters from individual data points. It starts with each point as a single cluster and iteratively merges the closest clusters until a single cluster containing all points is formed or a stopping criterion is met.
  - *Computational Cost*: Computationally intensive, especially for calculating the distance matrix, typically $O(n^2 d)$ or $O(n^2 \log n)$ depending on the linkage method. This quadratic complexity makes it very slow for datasets with 100,000 points.

  - *Downsides*: High memory usage, challenging for large datasets, and once a merge is made, it cannot be undone, potentially leading to suboptimal groupings.

- **Spectral Clustering**: A graph-based method that transforms the clustering problem into a graph partitioning problem. It constructs a similarity graph from the data points and then performs dimensionality reduction using eigenvalue decomposition on the graph Laplacian before applying k-means or another simple clustering method on the reduced dimensions.

  - *Computational Cost*: The most significant cost comes from constructing the similarity graph (up to $O(n^2)$) and performing eigenvalue decomposition (typically $O(n^3)$). This cubic complexity renders it computationally infeasible for 100,000 data points.
  - *Downsides*: High computational expense for large n, sensitive to the choice of similarity function and the number of components.

- **Gaussian Mixture Models (GMM)**: A probabilistic model that assumes data points are generated from a mixture of several Gaussian distributions with unknown parameters. It uses an Expectation-Maximization (EM) algorithm to iteratively estimate the parameters.

  - *Computational Cost*: Involves iterative optimization, with complexity roughly $O(nkd^2I)$ or $O(nkd^3I)$ depending on the covariance matrix type. For high-dimensional data and many components, this can be computationally demanding.
  - *Downsides*: Can be slow to converge, sensitive to initialization, and assumes data follows Gaussian distributions.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: A density-based algorithm that groups together data points that are closely packed together, marking as outliers points that lie alone in low-density regions.

  - *Computational Cost*: Generally efficient for spatial data, typically $O(n \log n)$ or $O(n^2)$ in the worst case, depending on the spatial indexing structure used.
  - *Downsides*: Highly sensitive to its 'epsilon' ϵ and 'min samples' parameters, and struggles with varying densities in clusters. It also identifies a significant number of points as noise, which might not be desirable for a comprehensive customer segmentation.

F.2 Comparative Analysis and Justification for K-Means

To evaluate the performance and feasibility of these clustering methods on our large dataset, we conducted a comparative analysis. Given the prohibitive computational cost of some methods on the full 100,000 data points, we randomly sampled 10,000 data points and performed each clustering algorithm 10 times. The average silhouette score was then calculated for each method to assess cluster quality.

The average silhouette scores across 10 runs are summarized below in Table F.1:

As evident from the results, K-means consistently yielded higher silhouette scores across various cluster numbers, significantly outperforming Spectral and GMM, and being highly competitive with Agglomerative clustering (especially for higher number of clusters) while offering a far superior computational efficiency on larger datasets.

A specific note on DBSCAN: The initial run showed "N/A", indicating DBSCAN struggled to form meaningful clusters with default parameters. With suitable tuning of parameters (e.g. dbscan params = (1.4, 5)), DBSCAN can achieve a high silhouette score (e.g., 0.576). However, this high score primarily reflects the quality of the *identified clusters*, but a substantial portion of the dataset (86,459 out of 100,000 data points) was classified as noise and assigned a silhouette score of -1. For the purpose of comprehensive customer segmentation where most customers need to be assigned to a group, this high proportion of unclustered "noise" points makes DBSCAN unsuitable, despite its ability to find dense regions.

Therefore, considering both performance (silhouette score) and computational scalability for a dataset of our size, **K-means clustering was selected** as the most appropriate method for this dissertation.

Table F.1: Average Silhouette Scores Across 10 Runs

Method	k	**Silhouette Score**
Agglomerative	3	0.138656
Agglomerative	2	0.130510
K-Means	2	0.119445
K-Means	**8**	**0.113536**
K-Means	3	0.112691
K-Means	7	0.112266
K-Means	9	0.110510
K-Means	5	0.110286
K-Means	10	0.109831
Agglomerative	4	0.100437
Spectral	2	0.093496
GMM	3	0.091931
Spectral	3	0.091851
Agglomerative	6	0.091838
Agglomerative	9	0.090095
Agglomerative	5	0.088824
Agglomerative	10	0.088701
Agglomerative	7	0.088175
Agglomerative	8	0.087319
Spectral	4	0.086610
GMM	4	0.078621
Spectral	8	0.071310
Spectral	9	0.067314
Spectral	6	0.066264
Spectral	5	0.065079
Spectral	10	0.063910
Spectral	7	0.060284
GMM	6	0.043439
GMM	5	0.024314
GMM	9	0.019503
GMM	7	0.018852
GMM	10	0.012412
GMM	8	0.008595
DBSCAN	N/A	N/A

Appendix G

Selected Code Samples

[1] Selected code snippets, illustrating the generation of plots, results, and analyses, are provided in this section; a comprehensive inclusion was omitted to avoid excessive appendix length.

G.1 Imputing SOC Label through Bayesian Sampling

```
from collections import Counter

# 1. Load the priors and means, setting SOC as the index
means_df = pd.read_csv('Downloads/occupation x region.csv', index_col=0,
    thousands=',')
prior_df = pd.read_csv('Downloads/prior2.csv', index_col=0, thousands=',')

# 2. Compute a global income std-dev for the Gaussian likelihood
sigma = salary_df['income'].std()

# 3. Extract the list of SOC labels from the index
soc_labels = prior_df.index.tolist()

# 4. Compute posterior vector for one row:
def compute_posterior(row):
    region = row['region']
    inc    = row['income']
    prior  = prior_df[region].astype(float)
    prior /= prior.sum()
    means      = means_df[region].astype(float)
    likelihood = np.exp(-0.5 * ((inc - means) / sigma) ** 2)
    post       = prior * likelihood
    return post / post.sum()

# 5. Draw 10 samples from that posterior, take the most common label:
def sample_modal_soc(row, n_samples=10):
    post = compute_posterior(row).values
    draws = np.random.choice(soc_labels, size=n_samples, p=post)
    # Counter.most_common(1) returns [(label, count)]
```

```
    return Counter(draws).most_common(1)[0][0]

# 6. Reproducibility
np.random.seed(42)

# 7. Apply to each row
salary_df['SOC_label'] = salary_df.apply(sample_modal_soc, axis=1)
```

G.2 PCA Analysis for Feature Selection prior to K-Means Clustering

```
import numpy as np
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import VarianceThreshold
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from tqdm import tqdm

# tqdm setup
tqdm.pandas(desc="Progress")

# --- Step 1: Infer binary vs. numeric ---
binary_cols = [
    col for col in features_df.columns
    if set(features_df[col].dropna().unique()).issubset({0, 1})
]
numeric_cols = [col for col in features_df.columns if col not in binary_cols]

# --- Step 2: Scale numeric features ---
scaler = StandardScaler()
scaled_num = pd.DataFrame(
    scaler.fit_transform(features_df[numeric_cols]),
    index=features_df.index,
    columns=numeric_cols
)
scaled = pd.concat([scaled_num, features_df[binary_cols]], axis=1)

# --- Step 3: Variance threshold on numeric cols---
vt = VarianceThreshold(threshold=0.01)
vt.fit(scaled[numeric_cols])
keep_num = [c for c, keep in zip(numeric_cols, vt.get_support()) if keep]
reduced = pd.concat([scaled[keep_num], scaled[binary_cols]], axis=1)

# --- Step 4: Drop highly correlated numeric features ---
corr = reduced[keep_num].corr().abs()
upper = corr.where(np.triu(np.ones(corr.shape), k=1).astype(bool))
drop_corr = [col for col in upper.columns if (upper[col] > 0.9).any()]
pruned = reduced.drop(columns=drop_corr)
```

```
# --- Step 5: Fill NaNs in numeric columns ---
numeric_pruned = [c for c in keep_num if c not in drop_corr]
pruned[numeric_pruned] = pruned[numeric_pruned].fillna(0)

# --- Step 6: PCA-driven selection of top features ---
pca = PCA(n_components=0.8, svd_solver='full', random_state=42)
pca.fit(pruned[numeric_pruned])
loadings = pd.DataFrame(
    pca.components_.T,
    index=numeric_pruned,
    columns=[f'PC{i+1}' for i in range(pca.n_components_)]
)

top_feats = set()
for pc in tqdm(loadings.columns, desc="Selecting top PCA features"):
    top_feats.update(loadings[pc].abs().nlargest(2).index)

# --- Step 7: Final feature list ---
final_features = list(top_feats) + binary_cols
features_final = pruned[final_features]

# --- Step 8: Evaluate KMeans from k=2 to k=10 ---
results = []
for k in tqdm(range(2, 11), desc="Evaluating KMeans"):
    km = KMeans(n_clusters=k, random_state=42).fit(features_final)
    sil = silhouette_score(features_final, km.labels_)
    results.append({'k': k, 'inertia': km.inertia_, 'silhouette': sil})

results_df = pd.DataFrame(results)
print(results_df)
```

G.3 Personalized Advice for Cash Flow Stability

```
import pandas as pd
import numpy as np

#--- Define Thresholds and Parameters
SAVINGS_BUFFER_MONTHS = 3 # Number of months of essential + housing costs for
    buffer
DTI_THRESHOLD = 0.40 # NEW: 40% for Debt Service Ratio (Debt Payments / Income)
BUFFER_BUILD_MONTHS_TARGET_6M = 6 # Target timeframe (in months) to build savings
    buffer (first attempt)
BUFFER_BUILD_MONTHS_TARGET_12M = 12 # Target timeframe (in months) to build
    savings buffer (second attempt)
ESSENTIAL_HOUSING_REBUDGET_PCT = 0.15 # 15% initial reduction percentage for
    rebudgeting suggestion
ARREARS_IMPACT_INTEREST_RATE = 0.20 # Example yearly interest rate for arrears
    impact advisory (20%) # Not used in final code

# New: APRs for specific debt products
DEBT_APRS = {
```

```
    'credit_card_type1': 0.2007, # 20.07%
    'credit_card_type2': 0.25, # 25%
    'short_term_loan': 0.0687, # 6.87% (Note: Fixed loan, not optimized for
        paydown here)
    'overdraft': 0.3827 # 38.27%
}

# Sequential Optimization Pipeline
# Dictionary to store recommendations for all customers
all_customer_recommendations = {}

print("--- Starting Financial Optimization Engine ---")

for customer_id, customer_data in sampled_df.iterrows():
    print(f"\n--- Optimizing for Customer ID: {customer_id} ---")

    # Initialize recommendations dictionary for the current customer
    recommendations = {
        'cash_flow_stability': 'Not applicable yet.',
        'debt_management': 'Not applicable yet.',
        'wealth_maximization': 'Not applicable yet.'
    }

    # Flags to track if a layer's objectives are met, determining progression to
        the next layer
    is_layer1_stable = False
    is_layer2_stable = False
    is_layer3_stable = False # Added for Layer 3 tracking

    # Variables to carry over from Layer 1 for Layer 2 calculations
    # avg_current_discretionary_spend is needed early for Layer 1 re-logic
    # Ensure disc_spend_X columns exist
    disc_spend_cols = [f'disc_spend_{m}' for m in range(1, 13) if
        f'disc_spend_{m}' in customer_data.index]
    avg_current_discretionary_spend = np.mean(customer_data[disc_spend_cols]) if
        disc_spend_cols else 0

    # LAYER 1: Cash Flow Stability Optimization
    # Objective: Ensure consistent positive cash flow and a sufficient emergency
        savings buffer.
    # print("\n [Layer 1: Cash Flow Stability]") # Commented out to reduce output

    # 1.1 Initial Cash Flow Assessment (before any adjustments)
    # Ensure cash flow X columns exist
    cash_flow_cols = [f'cash_flow_{m}' for m in range(1, 13) if f'cash_flow_{m}'
        in customer_data.index]
    monthly_cash_flow_before_disc = [customer_data[col] for col in
        cash_flow_cols] if cash_flow_cols else [0]
    initial_avg_cf_before_disc = np.mean(monthly_cash_flow_before_disc)

    # Initialize variables for hypothetical cuts
    hypothetical_disc_cut_for_overall_positive_cf = 0 # Amount of discretionary
        cut applied
```

```
hypothetical_cf_increase_from_essential_housing = 0 # Amount from
    essential/housing cuts

# Calculate initial overall cash flow (before any cuts)
overall_cash_flow_initial = initial_avg_cf_before_disc -
    avg_current_discretionary_spend

# Start building recommendations for Layer 1 based on initial cash flow
if overall_cash_flow_initial >= 0:
    recommendations['cash_flow_stability'] = "Initial overall cash flow is
        stable and positive."
    # If overall cash flow is already positive, then final_cf_after_all_cuts
        is just this initial positive
    final_cf_after_all_cuts = overall_cash_flow_initial
    potential_surplus_after_initial_cuts = overall_cash_flow_initial # If
        already positive, this is the surplus
else: # overall_cash_flow_initial is negative, so cash flow stability issues
    exist
    recommendations['cash_flow_stability'] = " Cash flow stability issues
        found. Recommendations:"

    # Add details about negative cash flow before discretionary, if applicable
    if initial_avg_cf_before_disc < 0:
        recommendations['cash_flow_stability'] += f"\n Detected negative cash
            flow (before discretionary spending) in {sum(1 for cf in
            monthly_cash_flow_before_disc if cf < 0)} month(s) totaling
            {abs(sum(cf for cf in monthly_cash_flow_before_disc if cf <
            0)):.2f} pounds over the last year."

    # Step 1: Prioritize cutting discretionary spending to achieve positive
        overall cash flow
    disc_cut_needed_to_zero = abs(overall_cash_flow_initial) # How much cut is
        needed to bring to zero
    hypothetical_disc_cut_for_overall_positive_cf =
        min(disc_cut_needed_to_zero, avg_current_discretionary_spend)
    recommendations['cash_flow_stability'] += f"\n\n** Action 1: Optimize
        Discretionary Spending.** To achieve a consistent positive overall
        cash flow, reduce your average monthly discretionary spending by at
        least {hypothetical_disc_cut_for_overall_positive_cf:.2f} pounds."
    recommendations['cash_flow_stability'] += f" (Current average
        discretionary spend: {avg_current_discretionary_spend:.2f}) pounds."

    # Calculate overall cash flow after this discretionary cut
    overall_cash_flow_after_disc_cut = overall_cash_flow_initial +
        hypothetical_disc_cut_for_overall_positive_cf

    # Step 2: If cash flow is STILL negative after discretionary cuts, apply
        essential/housing cuts
    if overall_cash_flow_after_disc_cut < 0: # This means even after cutting
        all possible discretionary, CF is negative
        is_young_renter = (customer_data.get('age_group') == 'Aged 24 years
            and under' and
                           customer_data.get('housing_tenure') == 'Renter')
```

```
        customer_housing_cost = customer_data.get('housing_cost', 0) # Ensure
            column exists
        customer_essential_cost = customer_data.get('essential_cost', 0) #
            Ensure column exists

        if is_young_renter:
            recommendations['cash_flow_stability'] += "\n\n** Action 2 (Young
                Renter):** Given your young age and renting status, consider
                the significant financial benefit of temporarily returning to
                live with parents/family to substantially reduce housing costs.
                This could quickly turn your cash flow positive."
            hypothetical_cf_increase_from_essential_housing =
                customer_housing_cost # Assume full housing cost reduction
        else:
            avg_essential_housing_cost = customer_essential_cost +
                customer_housing_cost
            hypothetical_cf_increase_from_essential_housing =
                avg_essential_housing_cost * ESSENTIAL_HOUSING_REBUDGET_PCT
            recommendations['cash_flow_stability'] += f"\n\n** Action 2:**
                Re-evaluate your essential and housing costs. Aim to rebudget
                and reduce these expenses by approximately
                {ESSENTIAL_HOUSING_REBUDGET_PCT:.0%} (e.g.,
                {hypothetical_cf_increase_from_essential_housing:.2f} per
                month) pounds."
            recommendations['cash_flow_stability'] += "This might involve
                seeking cheaper accommodation, optimizing utility usage, or
                finding more cost-effective essentials."

        # Update overall cash flow after essential/housing cuts
        final_cf_after_all_cuts = overall_cash_flow_after_disc_cut +
            hypothetical_cf_increase_from_essential_housing
    else: # If cash flow is positive after only discretionary cuts, no
        essential/housing cuts for CF stability
        final_cf_after_all_cuts = overall_cash_flow_after_disc_cut
        # hypothetical_cf_increase_from_essential_housing remains 0

    # Step 3: Critical Check - If cash flow remains deeply negative after BOTH
        types of cuts
    if final_cf_after_all_cuts < 0:
        recommendations['cash_flow_stability'] += "\n\n** Critical:** Even
            after considering significant discretionary and essential/housing
            expense reductions, your cash flow remains deeply negative. It is
            highly recommended to **seek professional financial advice**
            immediately to explore all options, including income generation,
            debt restructuring, or welfare support."

        # Set final_savings to current_savings_balance as advice is deemed
            impossible to implement
        recommendations['final_savings_post_advice'] =
            customer_data.get('savings_account_12', 0)
        all_customer_recommendations[customer_id] = recommendations
        # print(recommendations['cash_flow_stability']) # For debugging
        # print("ln"+"="*50) # For debugging
        continue # Stop optimization for this customer
```

```
    else:
        # Only add this summary if cuts were actually recommended and applied
        if hypothetical_disc_cut_for_overall_positive_cf > 0 or
            hypothetical_cf_increase_from_essential_housing > 0:
            recommendations['cash_flow_stability'] += f"\nThese actions
                combined could lead to a positive average overall cash flow of
                ~ {final_cf_after_all_cuts:.2f}pounds ."

        # This is the actual surplus available after all *necessary cuts to
            achieve positive CF
        potential_surplus_after_initial_cuts = final_cf_after_all_cuts

# Calculate max potential monthly surplus if ALL discretionary and potential
    essential/housing cuts are made
# This value represents the highest possible surplus by maximizing all
    possible expense reductions.
max_potential_monthly_surplus = (initial_avg_cf_before_disc +
    hypothetical_cf_increase_from_essential_housing) +
    avg_current_discretionary_spend

# Step 4: Savings Buffer Plan (Aggressive Pursuit)
current_savings_balance = customer_data.get('savings_account_12', 0)

# Ensure required_buffer is calculated or available from data
required_buffer_amount = (customer_data.get('essential_cost', 0) +
    customer_data.get('housing_cost', 0)) * SAVINGS_BUFFER_MONTHS
shortfall_to_buffer = required_buffer_amount - current_savings_balance

if shortfall_to_buffer <= 0:
    recommendations['cash_flow_stability'] += "\nEmergency savings buffer is
        already sufficient."
    is_layer1_stable = True # Set to True if buffer is sufficient
    final_savings_post_advice = current_savings_balance # No change needed to
        existing savings
else:
    recommendations['cash_flow_stability'] += f"\n\n** Action:** Your current
        savings ({current_savings_balance:.2f} pounds) are below the target
        buffer of {required_buffer_amount:.2f} pounds."
    monthly_contribution_6m = shortfall_to_buffer /
        BUFFER_BUILD_MONTHS_TARGET_6M
    monthly_contribution_12m = shortfall_to_buffer /
        BUFFER_BUILD_MONTHS_TARGET_12M

    # Reset is_layer1_stable before re-evaluating based on buffer achievability
    is_layer1_stable = False

    # Calculate remaining discretionary spend after initial cuts
    remaining_discretionary_spend_after_initial_cuts =
        avg_current_discretionary_spend -
        hypothetical_disc_cut_for_overall_positive_cf

    if max_potential_monthly_surplus >= monthly_contribution_6m:
        # Calculate the additional cut needed
```

```
            additional_cut_for_buffer_raw = monthly_contribution_6m -
                potential_surplus_after_initial_cuts
            # Cap the additional cut to what's remaining of discretionary spending
                after initial cuts
            required_additional_disc_cut_for_buffer = max(0,
                min(additional_cut_for_buffer_raw,
                remaining_discretionary_spend_after_initial_cuts))

            recommendations['cash_flow_stability'] += f"\nTo achieve your
                emergency fund within {BUFFER_BUILD_MONTHS_TARGET_6M} months, save
                at least {monthly_contribution_6m:.2f} pounds per month."
            if required_additional_disc_cut_for_buffer > 0:
                recommendations['cash_flow_stability'] += f" This will require an
                    additional reduction of
                    {required_additional_disc_cut_for_buffer:.2f} pounds from your
                    discretionary spending (beyond cuts needed for overall positive
                    CF)."
            else:
                recommendations['cash_flow_stability'] += " This can be achieved
                    with your current potential surplus."

            is_layer1_stable = True # Achievable with 6-month plan
            final_savings_post_advice = required_buffer_amount

        elif max_potential_monthly_surplus >= monthly_contribution_12m:
            # Calculate the additional cut needed
            additional_cut_for_buffer_raw = monthly_contribution_12m -
                potential_surplus_after_initial_cuts
            # Cap the additional cut to what's remaining of discretionary spending
                after initial cuts
            required_additional_disc_cut_for_buffer = max(0,
                min(additional_cut_for_buffer_raw,
                remaining_discretionary_spend_after_initial_cuts))

            recommendations['cash_flow_stability'] += f"\nIf a
                {BUFFER_BUILD_MONTHS_TARGET_6M}-month plan is too aggressive, you
                can aim to build your emergency fund within
                {BUFFER_BUILD_MONTHS_TARGET_12M} months by saving at least
                {monthly_contribution_12m:.2f} pounds per month."

            if required_additional_disc_cut_for_buffer > 0:
                recommendations['cash_flow_stability'] += f" This will require an
                    additional reduction of
                    {required_additional_disc_cut_for_buffer:.2f} pounds from your
                    discretionary spending (beyond cuts needed for overall positive
                    CF)."
            else:
                recommendations['cash_flow_stability'] += " This can be achieved
                    with your current potential surplus."
            is_layer1_stable = True # Achievable with 12-month plan
            final_savings_post_advice = required_buffer_amount
        else:
            recommendations['cash_flow_stability'] += f"\nBased on your potential
                monthly surplus (even after maximizing all possible cuts), it
```

```
                might be challenging to build the full buffer within
                {BUFFER_BUILD_MONTHS_TARGET_12M} months."
            recommendations['cash_flow_stability'] += "\n** Recommendation:** Keep
                this emergency fund objective in mind as your income grows, and
                prioritize increasing your monthly surplus through additional
                income streams or further long-term expense optimization."
            is_layer1_stable = False # Not achievable within target timeframes
            final_savings_post_advice = current_savings_balance # Savings balance
                itself doesn't change in Layer 1, only future surplus allocation

    # Store final_savings_post_advice (after Layer 1 considerations)
    recommendations['final_savings_post_advice'] = final_savings_post_advice
```

References

[1] In line with departmental policy use of ai in msc dissertation, the author declares his use of ai for the following purposes: formatting bibliographies (bibtex), improving grammar and sentence structures, generating copies of modified code (from earlier toy or iterative versions) as well as non-substantive alterations to plots and figures which improve clarity and readability.

[2] Samuel Assefa, Danial Dervovic, Mahmoud Mahfouz, Tucker Balch, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. Ai research discussion paper, JPMorgan Chase & Co., 2020.

[3] John O. Awoyemi, Adebayo O. Adetunmbi, and Samuel A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pages 1–9, 2017.

[4] Kal Bukovski, Aurelija Cepkauskaite, Sarfarfaz Shaikh, John Finch, and Chuks Otioma. Consumers at the Heart of Innovation: Financial Health Evaluation in the UK Regulatory Landscape. Whitepaper, Sopra Steria UK and University of Glasgow, Adam Smith Business School, May 2025.

[5] Ryan Butler, Taha Hussain, and Hakan Kapucu. Customer Behaviour Classification Using Simulated Transactional Data. ResearchGate, 2022. Accessed: 19 June 2025.

[6] Cambridge Centre for Alternative Finance (CCAF). The Global State of Open Banking and Open Finance. Research report, Cambridge Centre for Alternative Finance (CCAF), University of Cambridge Judge Business School, November 2024.

[7] Centre for Finance, Innovation and Technology (CFIT). CFIT Open Finance Blueprint. Technical report, Centre for Finance, Innovation and Technology (CFIT), February 2024. Accessed: 19 June 2025.

[8] Charlotte Crosswell. TRUSTEE END OF IMPLEMENTATION ROADMAP REPORT: Recommendations for the Future of Open Banking. Roadmap report, Open Banking Implementation Entity (OBIE), January 2023.

[9] European Central Bank. Revised Payment Services Directive (PSD2): An important step towards an integrated European retail payments market, March 2018.

[10] Financial Conduct Authority. Retail investment advice: Clarifying the boundaries and exploring the barriers to market development. Finalised Guidance FG15/1, Financial Conduct Authority, Jan 2015. Accessed: 16 June 2025.

[11] Financial Conduct Authority. Financial lives 2020 survey: the impact of coronavirus, 2020. Accessed: 5 June 2025.

[12] Financial Conduct Authority. Financial lives 2020 survey: Jan 2024 re-contact survey summary, 2024. Accessed: 9 June 2025.

[13] I. T. Jolliffe. *Principal Component Analysis*. Springer, New York, NY, 2nd edition, 2002.

[14] Ministry of Housing, Communities and Local Government. English housing survey, 2024. Accessed: 9 June 2025.

[15] Max A. Mosley. Box c: Households savings amid the cost-of-living crisis, Aug 2022. Accessed: 13 June 2025; Published as part of National Institute UK Economic Outlook, Issue 7, pp. 34-36.

[16] Office for National Statistics. Ashe: Table 3 - region by occupation (2-digit soc), 2022. Accessed: 9 June 2025.

[17] Office for National Statistics. Small area model based income estimates, financial year ending 2020, 2022. Accessed: 5 June 2025.

[18] Office for National Statistics. Distribution of uk household disposable income, financial year ending 2022, Aug 2023. Accessed: 5 June 2025.

[19] Office for National Statistics. Family spending workbook 2: expenditure by income, 2024. Accessed: 10 June 2025.

[20] Office for National Statistics. Family spending workbook 3: expenditure by region, 2024. Accessed: 10 June 2025.

[21] Office for National Statistics. Custom census 2021 data, breakdown by age, soc and region, 2025. Accessed: 9 June 2025.

[22] Statista Research Department. Average monthly savings as percentage of average income in great britain in summer and autumn 2014, by age group, Apr 2015. Accessed: 13 June 2025; Data for survey period: November 2014.

[23] Thistle Initiatives. Challenges with Open Banking, 2024.

[24] Wealth Connexion. The hierarchy of financial needs: Your path to financial freedom, October 2020. Accessed: 17 June 2025.

[25] Jinyong Yang. Study of an Adaptive Financial Recommendation Algorithm Using Big Data Analysis and User Interest Pattern with Fuzzy K-Means Algorithm. *International Journal of Computational Intelligence Systems*, 17(1):310, 2024.

[26] Xu Zhu, Qingyong Chu, Xinchang Song, Ping Hu, and Lu Peng. Explainable prediction of loan default based on machine learning models. *Data Science and Management*, 6(1):123–133, 2023.