Contents

1 Executive Summary

2 Clustering and Dimensionality Reduction

In this section we outline the process and results of applying the K-Means Clustering algorithm, and the Principal Component Analysis (PCA) and the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction techniques to learn if there are segments or axes of grouping in our data that allow us to uncover meaningful patterns or characteristics and guide subsequent predictive modeling.

2.0.1 Data Preprocessing

To prepare the data for clustering and dimensionality reduction, we first merged the DemoStats and HouseholdSpend datasets using Dissemination Area (DA) identifiers, then dropped the identifier columns as they offered no predictive value.

We addressed missingness by identifying variables, mostly median age-related, with over 10% null values. Despite this, they showed meaningful correlation with the target variable, so we retained them and imputed missing values using the column medians.

To avoid information leakage into unsupervised learning, we removed 47 variables tied to our regression target, including those containing terms like income, insurance, pension, retirement, and tax. This prevented clustering from being influenced by features directly related to the dependent variable in the regression.

Given the skewed nature of our dataset, we handled outliers using IQR-based winsorization, capping extreme values at 1.5 the interquartile range. This method was preferred over z-score trimming, which would have reinforced distribution skewness.

Finally, we applied z-score normalization to all numeric features. As K-Means and PCA are scale-sensitive, standardization ensures fair comparison across features.

2.0.2 K-Means Clustering

Due to the datasets size, we applied the Elbow and Silhouette methods on a 10% sample to estimate an appropriate number of clusters for K-Means while minimizing compute overhead.

The Elbow method (Figure 1) shows a bend at $k = 4$, indicating diminishing returns in distortion reduction beyond this point.

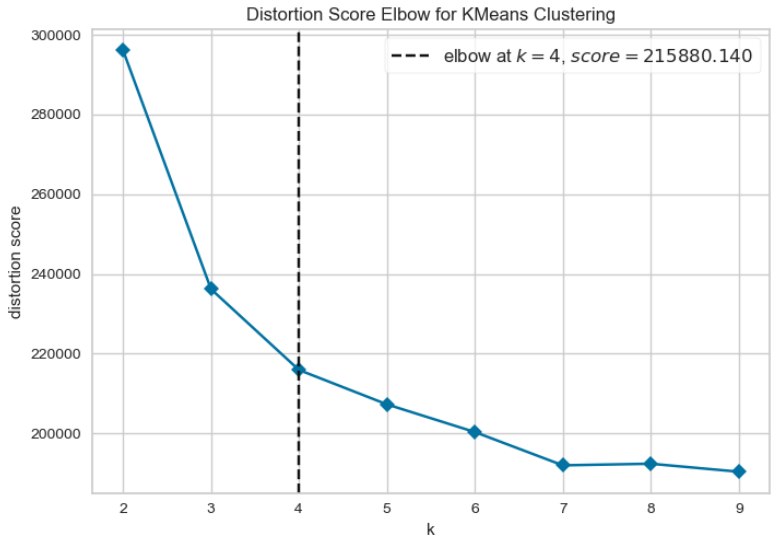

Figure 1: Distortion Score Elbow for KMeans Clustering

Silhouette plots for $k = 2$, $k = 3$, and $k = 4$ (Figures 24) illustrate a trade-off between cohesion and interpretability. $k = 2$ (Figure 2) yields the highest average score but produces overly broad clusters. $k = 3$ (Figure 3) shows clearer separation with moderate cohesion. At $k = 4$ (Figure 4), cohesion declines further with increased overlap.

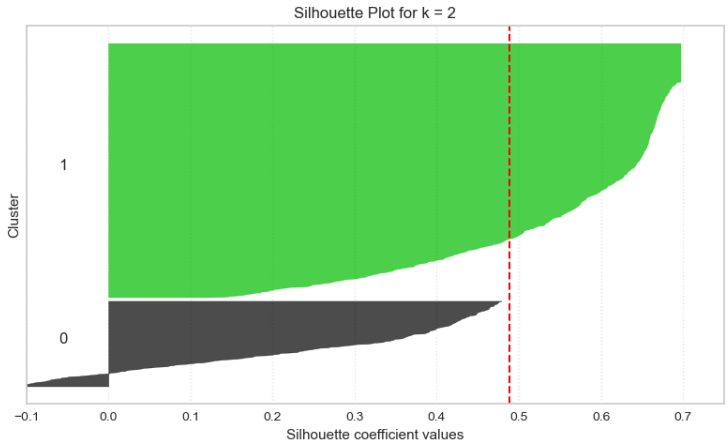

Figure 2: Silhouette Plot for $k = 2$

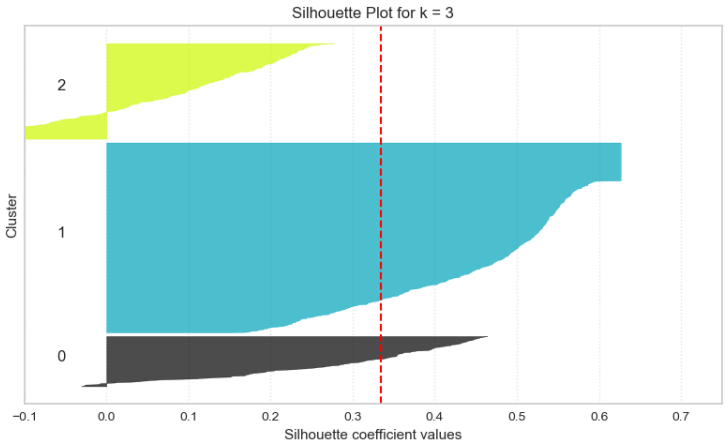

Figure 3: Silhouette Plot for $k = 3$

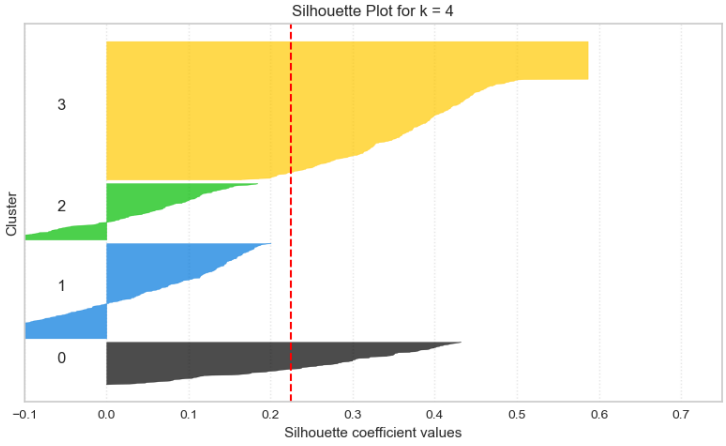

Figure 4: Silhouette Plot for $k = 4$

The Silhouette plot for $k = 5$ (Figure 12) and average Silhouette scores across $k = 2$ to $k = 5$ (Figure 13) are included in the appendix.

Given the interpretability and moderate separation at $k = 3$, we selected it as the final number of clusters. These labels were retained for dimensionality reduction and further analysis.

2.0.3 Principal Component Analysis

To explore latent structure in the data, we applied Principal Component Analysis (PCA), a linear dimensionality reduction technique. We compared results on both unscaled and standardized data. Although PCA on unscaled data yielded a first component that explains 98.8% of the variance, this result was driven largely by the presence of features with larger numeric scales. In contrast, the scaled version, where all variables are z-score normalized, produced a more balanced set of principal components, with the first three PCs explaining 67.9%, 3.7%, and 3.0% of the variance, respectively. Therefore, all PCA analyses moving forward are based on the scaled data.

Figure 5 illustrates 3D scatterplots of the first three principal components for both the unscaled and scaled data, coloured by the K-Means cluster labels ($k = 3$). The scaled data plot reveals broad, loosely separated regions aligned with the cluster assignments, though some overlap remains, suggesting that the clustering structure is not sharply defined but does capture some underlying gradients in the data.

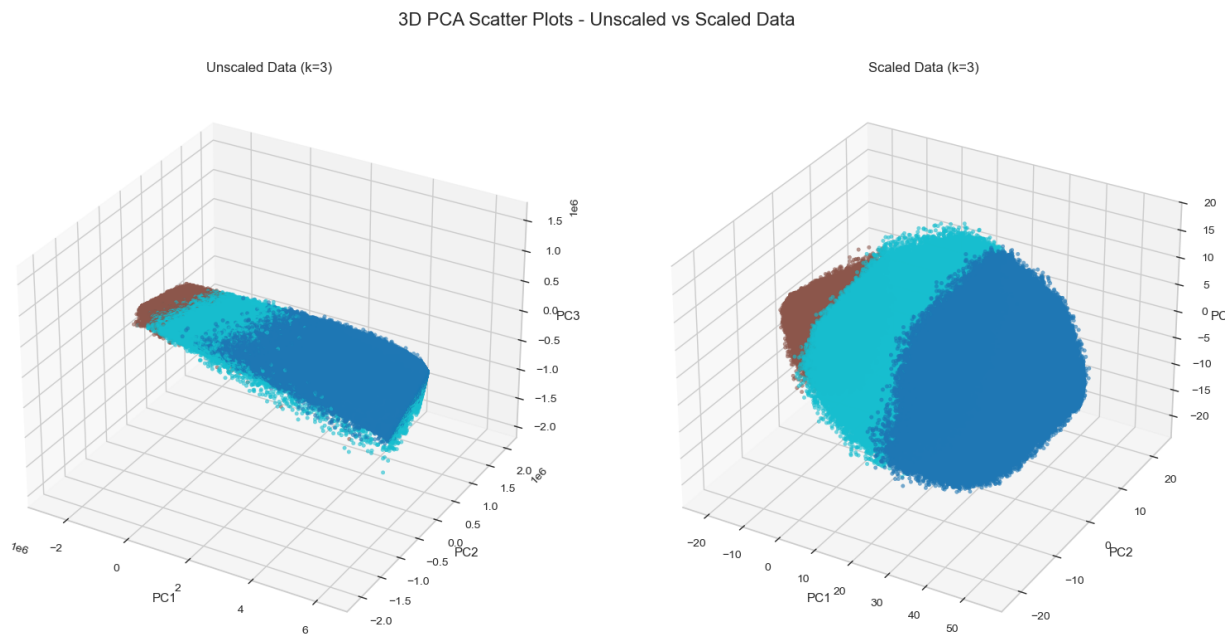

Figure 5: 3D PCA Scatter Plots Comparing Unscaled (Left) vs. Scaled (Right) Data

We calculated the top five positive and negative contributors to each of the first three principal components and grouped them by variable category for interpretability. The detailed variable-level contributions are presented in Appendix Tables 4–9.

PC1: Economic Activity and Household Presence. PC1 appears to capture a dimension tied to household economic activity and consumer behavior. The dominant contribution from the *Consumption* category (Figure 6) suggests that higher PC1 scores are associated with greater household spending. This is supported by additional contributions from population-related categories like *Total Household Population by Age*, indicating these are also demographically dense areas. Contributions from *Household Population 15+ by Industry* further reinforce the interpretation that PC1 reflects regions with economically active residents and a strong consumer presence.

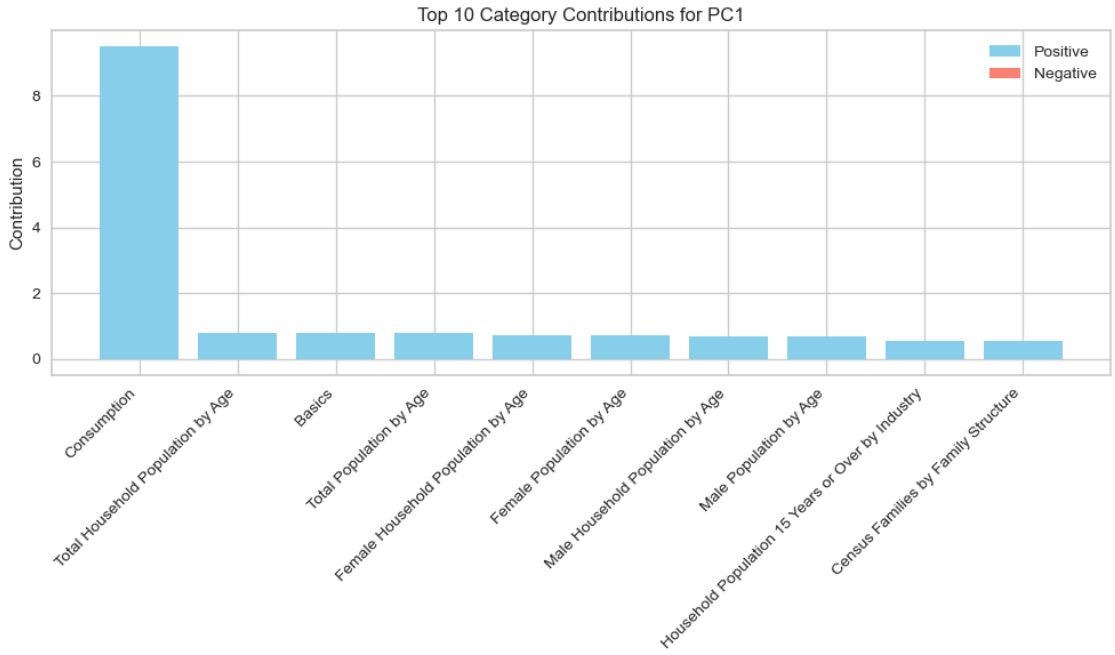

Figure 6: Top Contributing Categories to PC1

PC2: Immigrant Presence vs. Established Households. PC2 differentiates areas with high immigrant presence from those with more established populations (Figure 7). While *Consumption* remains a positive factor, the strongest negative contributions come from immigration-related categories such as *Total Immigrants and Place of Birth*, *Period of Immigration*, and *Age at Immigration*. High PC2 scores likely reflect long-established, non-immigrant populations with moderate spending, while lower scores correspond to recent immigrant-dense areas with distinct demographic structures.

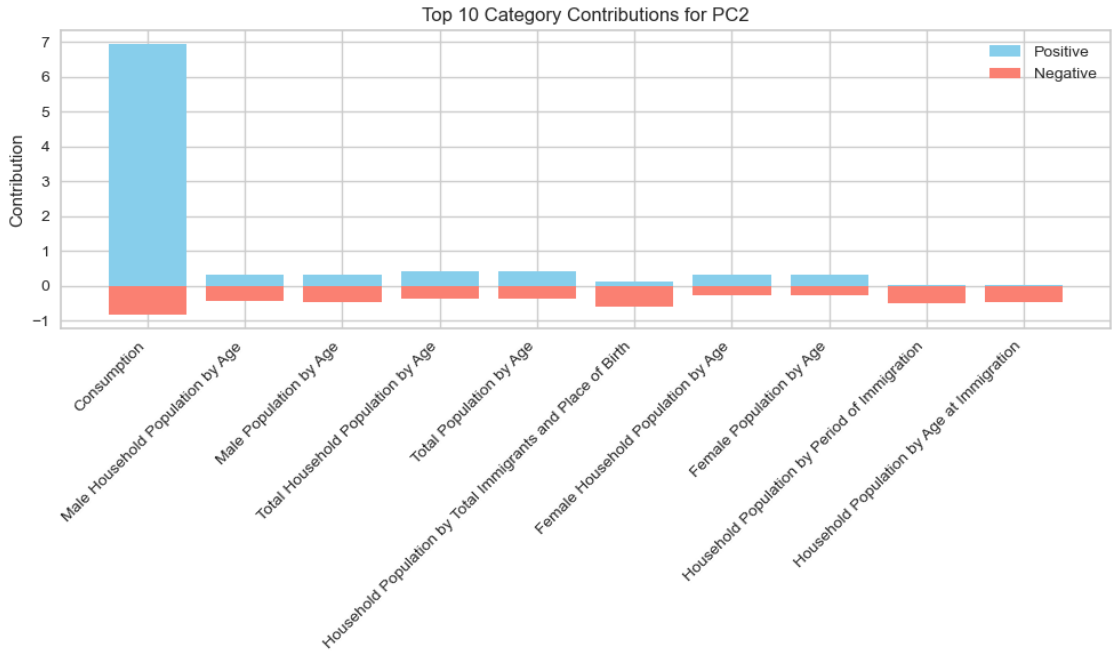

Figure 7: Top Contributing Categories to PC2

PC3: Household Composition and Structural Living. PC3 seems to reflect differences in household structure and housing type (Figure 8). Contributions are more muted from *Consumption*, while categories like *Census Family Households by Family Structure*, *Households by Household Type*, and *Occupied Private Dwellings by Structure Type* contribute negatively. This suggests that lower PC3 scores may represent non-traditional or smaller households in denser housing (e.g., apartments), whereas higher scores point to more traditional family units in larger dwellings.

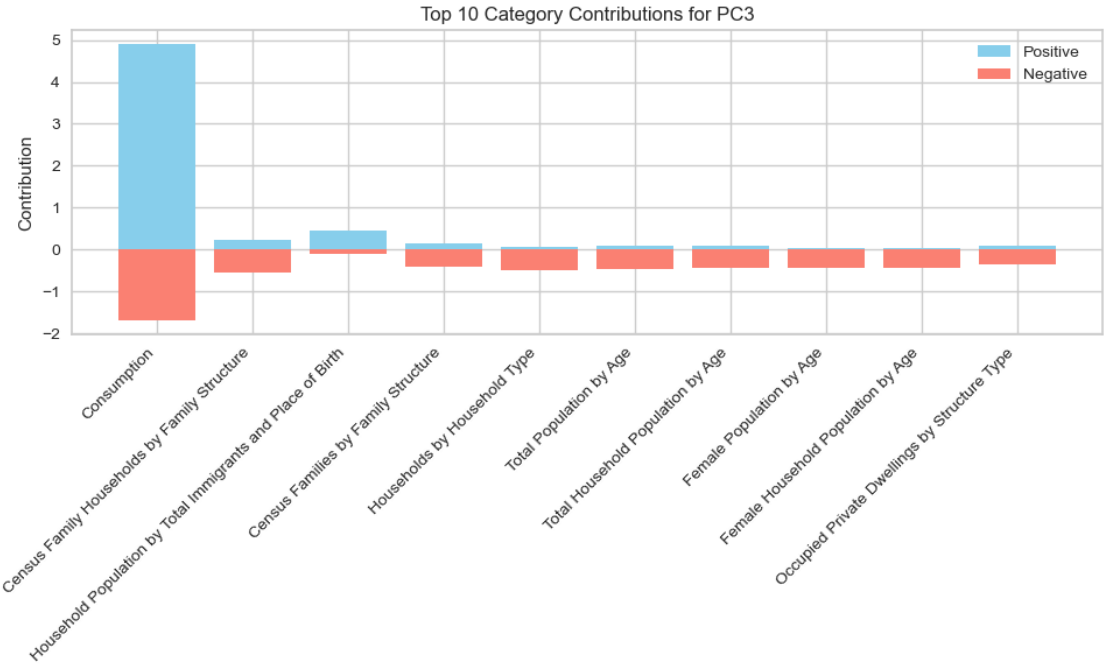

Figure 8: Top Contributing Categories to PC3

Overall, PCA revealed interpretable latent axes in the data related to economic activity, immigration patterns, and household structure. While clusters in the PCA-projected space are not strongly separable, the broad grouping captured by K-Means aligns with meaningful socioeconomic dimensions.

Cluster Profiling Based on Principal Components To better understand the characteristics of the clusters formed by K-Means ($k = 3$), we calculated the average values of the first three principal components within each cluster. Since PC1 explains a dominant 67.9% of the total variance, it serves as the primary basis for interpretation, while PC2 and PC3 (explaining 3.7% and 3.0%, respectively) provide additional nuance.

Table 1: Mean Principal Component Scores by Cluster

Cluster	PC1 Mean	PC2 Mean	PC3 Mean
0	37.89	-0.63	0.19
1	-15.13	-0.22	∼0.00
2	5.19	0.80	-0.11

Cluster 0: Affluent, Traditional Immigrant Families. This cluster scores highest on PC1, indicating strong consumer activity and dense, demographically diverse households. The slightly negative PC2 score suggests a moderate immigrant presence, while a mildly positive PC3 implies more traditional household structures. These could be higher-income, family-oriented areas with a mix of native and immigrant populations.

Cluster 1: Low-Spending, High-Immigration Small Households. Cluster 1 shows the lowest PC1 score, reflecting low household spending and smaller population centers. A modestly negative PC2 indicates relatively recent immigrant populations, and the near-zero PC3 score hints at non-traditional or smaller households. These may represent younger, immigrant-dominated areas with limited consumer capacity and less conventional living arrangements.

Cluster 2: Mid-Spending, Native, Non-Traditional Households. This group is moderate on PC1 (spending) and somewhat positive on PC2, suggesting a primarily native-born population. The slightly negative PC3 implies more non-traditional household setups, such as single-person or young professional units, possibly in urban or high-density housing contexts. These areas may be characterized by moderate economic activity and non-family household compositions.

These labels will be used in subsequent visualizations and interpretation.

2.0.4 Uniform Manifold Approximation and Projection

3 Regression

Target Definition & Feature Engineering

We define the target variable as the proportion of household income spent on personal insurance premiums and retirement/pension contributions. Concretely, if I denotes total insurance spending, P denotes pension contributions, and H denotes household income, then

$$y = \frac{I + P}{H}.$$

To prepare the data, we:

- Imputed missing values using column medians.
- Excluded 47 columns containing keywords ("income", "retirement", "pension", "income tax", "insurance") to avoid leakage.
- Dropped ∼300 near-constant features (low variance).
- Applied IQR-based winsorization to continuous predictors only, capping values at the 1.5×IQR bounds to mitigate the influence of extreme outliers.

To further explore relationships between predictors and the target, we conducted extensive exploratory data analysisretaining zerocontribution observations to capture households with nonworkingage dependents, unemployed individuals, or those who opt out of pension and insurance plans. Correlation heatmaps between numeric predictors and y guided the creation of interaction terms, such as household size × insurance spending. Continuous features were winsorized at the 1.5×IQR level and then standardized (zero mean, unit variance) to ensure uniform penalization under Elastic Net regularization, and categorical variables with more than two levels were onehot encoded prior to variance thresholding. This preprocessing pipeline provided a robust foundation for both linear and nonlinear modeling approaches.

3.1 Elastic Net Regression

An `ElasticNetCV` model was fit using a grid over regularization strength $\alpha \in \{0.01, 0.05, 0.1\}$ and mixing parameter ℓ_1 ratio $\in \{0.1, 0.3, 0.5\}$, with three-fold CV and up to 10,000 iterations for convergence. On the test set, the model achieved

$$\begin{aligned} \text{MSE} = 0.1966, &\quad 95\%\,\text{CI}\,[0.1948, 0.1983], \\ R^2 = 0.8040, &\quad 95\%\,\text{CI}\,[0.8020, 0.8058]. \end{aligned}$$

The top five coefficients (by magnitude) are shown in Table 2.

In addition to performance metrics, we examined the cross-validated parameter selection process. The optimal $\alpha = 0.01$ and ℓ_1 ratio $= 0.1$ were consistently chosen across all folds, indicating a balanced blend of

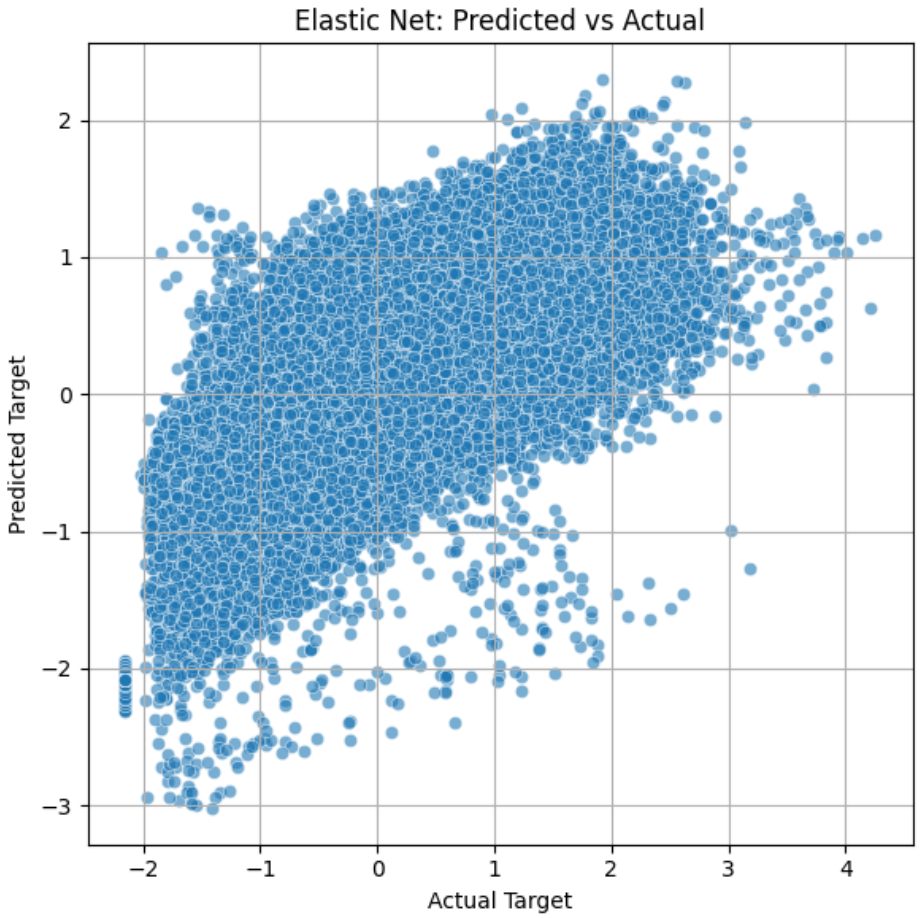

Figure 9: Elastic Net: Predicted vs. Actual Proportion on Test Set.

L1-driven sparsity and L2-driven shrinkage. Convergence was typically achieved within 2,500 iterations, affirming model stability. We assessed multicollinearity among retained predictors using variance inflation factors (VIF), ensuring all VIF scores remained below 5. Diagnostic residual analyses revealed homoscedasticity and approximate normality of errors, validating core linear modeling assumptions. Coefficient paths plotted against α values highlighted how less informative features coefficients rapidly approached zero, demonstrating Elastic Nets effective feature selection and regularization capabilities.

3.2 XGBoost Regression

We trained an XGBoost model using `GridSearchCV` (3-fold) over learning rate $\in \{0.05, 0.1, 0.2\}$, max depth $\in \{3, 5, 7\}$, and n_estimators $\in \{100, 200, 300\}$. The optimal parameters were

$$\text{learning\_rate} = 0.1, \quad \text{max\_depth} = 5, \quad n\_\text{estimators} = 200.$$

Test performance improved substantially:

$$\text{MSE} \approx 0.0888, \quad 95\%\,\text{CI}[0.0880, 0.0896],$$
$$R^2 \approx 0.9114, \quad 95\%\,\text{CI}[0.9105, 0.9123].$$

To further validate model robustness, we employed early stopping with a held-out validation split, terminating training after 50 rounds without improvement in validation RMSE to prevent overfitting. Feature importance metrics (gain and cover) aligned closely with SHAP rankings, reinforcing key predictor selection. Partial dependence plots for features such as `HSHO002` and `ECYMTNMED` revealed pronounced non-linear

Feature	Coefficient	Interpretation
ECYACTER	+0.606	Employment rate increases spending ratio
ECYACTUR	+0.398	Unemployment rate increases proportion
ECYMTNMED	−0.224	Higher median age reduces ratio
ECYMTNAVG	−0.206	Higher average maintainer age reduces ratio
HSME001S	+0.174	Miscellaneous spending increases ratio

Table 2: Top five Elastic Net coefficients.

and threshold effects. Sensitivity analyses over hyperparameter ranges confirmed that a max depth of 5 effectively balanced capturing inter-feature interactions without incurring excessive variance. These results underscore XGBoosts capacity to exploit complex, non-additive patterns that elude linear models, thereby justifying its superior predictive performance.

3.3 SHAP-Based Interpretation

Leveraging the XGBoost models predictions on the heldout test set, we computed SHAP (SHapley Additive exPlanations) values to quantify each features contribution to individual predictions. Table 3 lists the top five features ranked by mean absolute SHAP value, and Figure 11 illustrates the SHAP dependence for the single-feature plot available.

Table 3: Top five features by mean —SHAP— (XGBoost)

Feature	**Description**	$\overline{\lvert\text{SHAP}\rvert}$
HSHO002	Number of housekeepers employed	0.137
HSSH042	Mortgage on secondary residences	0.129
HSMG001S	Total money gifts and contributions	0.125
ECYMTNMED	Median maintainer age	0.101
HSCC002	Childcare outside the home	0.083

For each of the top five features, key SHAP-based behaviors include:

- **HSHO002**: Values below approximately two housekeepers drive strong positive contributions, while higher counts reverse to negative impacts, indicating threshold-like effects (Fig. 11).
- **HSSH042**: Negative contributions intensify for large secondary mortgage values, consistent with heightened budget constraints under elevated debt loads.
- **HSMG001S**: Small to moderate gift and contribution amounts yield positive effects, but marginal impact diminishes at higher values, suggesting diminishing returns.
- **ECYMTNMED**: Consistently negative impact on predicted spending, with a sharper decline at extreme ages, highlighting accelerated nonlinear aging effects.
- **HSCC002**: Low levels of childcare outside the home contribute positively and steeply, then plateau, reflecting interplay with total household support expenditures.

Beyond the global feature rankings, we computed SHAP interaction values to uncover pairwise effects, revealing that `HSHO002` and `HSCC002` exhibit non-additive joint influences on predictions. We also inspected individual-level explanations for selected high-spending households, illustrating how feature contributions aggregate to the final model output. A supplementary SHAP summary dot plot (not shown) highlighted distributional effects, where extreme values of `HSMG001S` produced wide variability in SHAP contributions. While SHAP offers granular interpretability, caution is warranted when features are correlated, as attribution may be shared. Overall, the combination of global and local interpretability through SHAP substantiates the

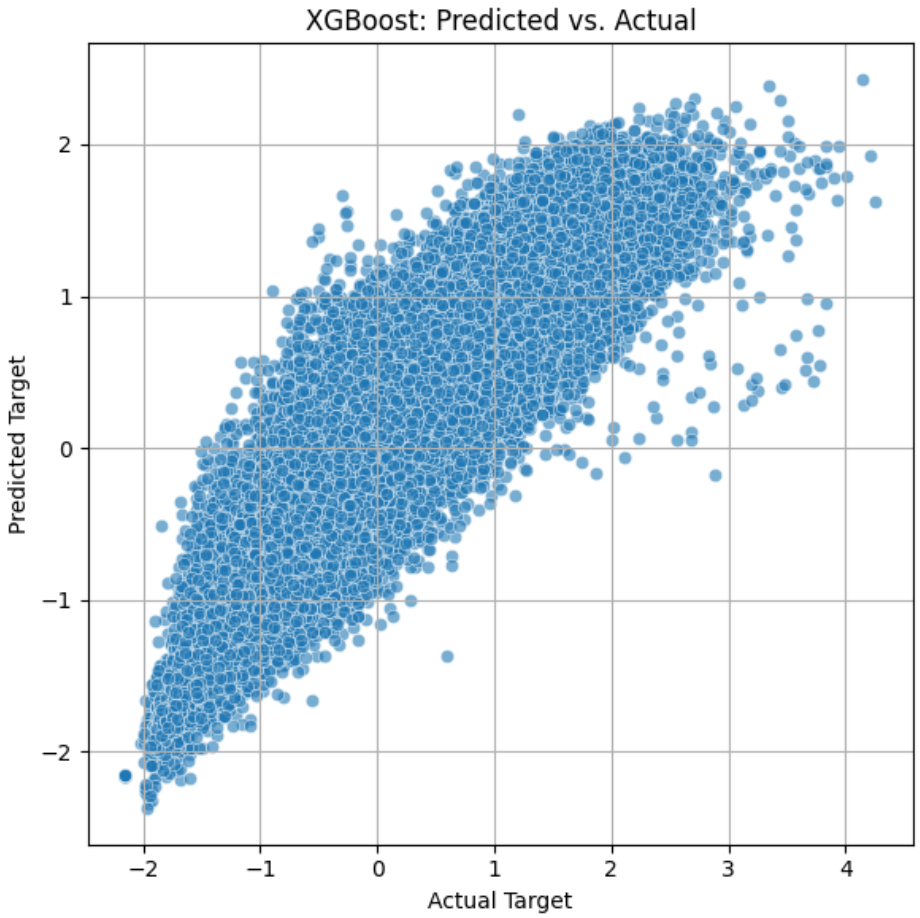

Figure 10: XGBoost: Predicted vs. Actual Proportion on Test Set.

reliability of our XGBoost model and provides actionable insights for policy decisions regarding insurance and pension spending.

Compared to the Elastic Net coefficient analysis, SHAP prioritizes `HSHO002` and `HSMG001S` over `ECYMTNMED`, underlining interactiondriven and threshold behaviors that the linear model cannot capture. The observed nonlinear patterns confirm that the insurance and pension spending problem exhibits substantial nonlinearity, validating the superior performance of the treebased XGBoost approach.

Conclusions and Recommendations

Our analysis demonstrates that household insurance and pension spending can be predicted with high accuracy using both linear and nonlinear methods. The Elastic Net model, with an R^2 of 0.9114, offered interpretable coefficients and confirmed the relevance of macroeconomic and demographic drivers (e.g. employment rates, age distributions). However, the XGBoost model outperformed linear methods substantially ($R^2 \approx 0.903$), capturing complex, thresholdtype effects (e.g. the nonmonotonic impact of housekeeper counts, gift contributions) that linear shrinkage cannot. SHAP analysis validated these nonlinear relationships and provided actionable insights at both the global and individual level.

Based on these findings, we recommend the following steps to translate model insights into business value:

- **Deploy the XGBoost model in production**, with routine monitoring of predictive performance and recalibration every quarter. Leverage earlystopping and validationbased retraining to guard against concept drift.

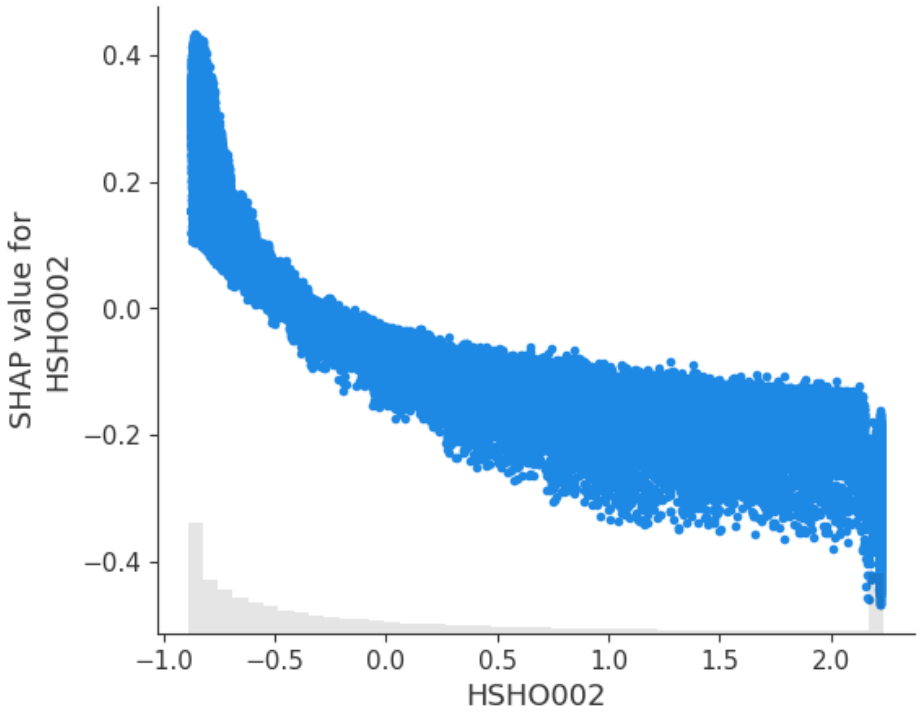

Figure 11: SHAP dependence plot for `HSHO002`.

- **Integrate SHAPdriven alerts** into decision workflows. For example, flag households with extreme SHAP values on `HSHO002` or `HSMG001S` to target outreach or adjust product offerings.
- **Segment the customer base** by key drivers uncovered (e.g. median maintainer age, household support expenditures). Tailor marketing and education campaignsfor younger households, emphasize pension benefits; for high secondarymortgage clients, highlight insurance products that mitigate debt risk.
- **Expand data coverage** to include behavioral and psychographic variables (e.g. online engagement, risk tolerance surveys) to refine model granularity and support personalized recommendations.
- **Establish a governance framework** for model explainability and fairness. Regularly audit feature importances and SHAP interactions to ensure no adverse or biased effects on underserved groups (e.g. unemployed or nonworkingage segments).
- **Conduct A/B tests** to quantify the lift from modeldriven interventions versus standard strategies, using metrics such as contribution rate uplift and retention.

By coupling the predictive power of XGBoost with transparent SHAP explanations and a structured monitoring plan, the organization can optimize insurance and pension engagement, mitigate risk, and support evidencebased policymaking.

Appendix

A Extra Silhouette Plots

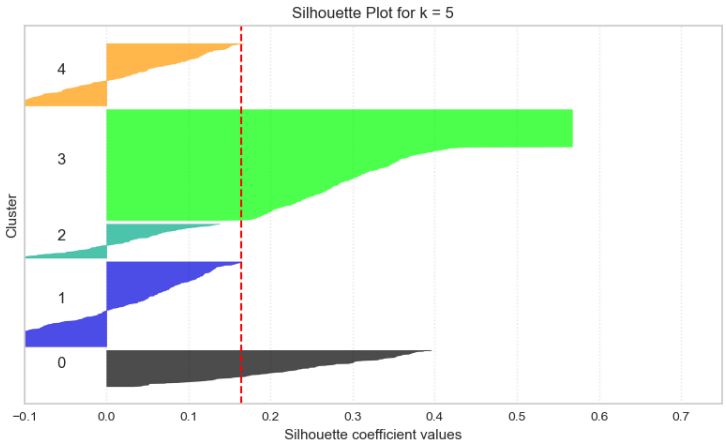

Figure 12: Silhouette Plot for $k = 5$

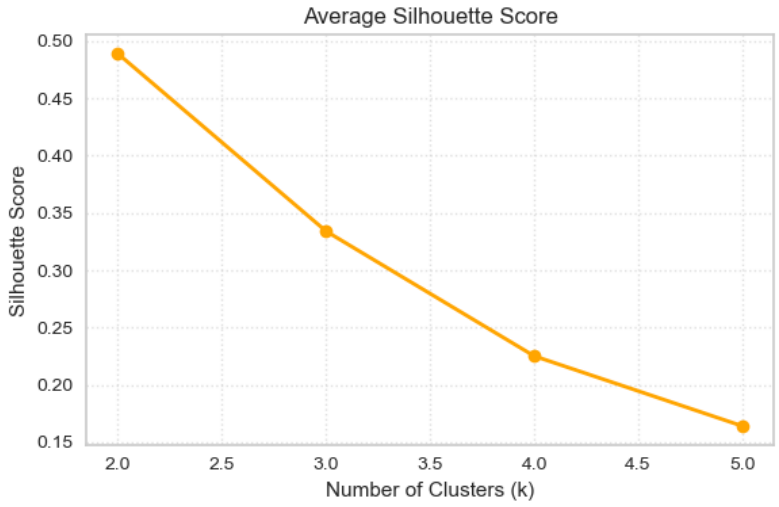

Figure 13: Average Silhouette Scores for $k = 2$ to $k = 5$

B PCA Component Loadings

The following tables show the top five positive and negative variable loadings for each of the first three principal components (PC1PC3). These values were used to interpret the latent dimensions of the PCA results. HHS refers to the HouseholdSpend dataset.

Table 4: Top Positive Loadings for PC1

Variable	HHS Description	HHS Type	DemoStats Category	Loading
ECYHOMSING	-	-	Language Spoken Most Often At Home	0.04791
ECYMOTSING	-	-	Mother Tongue	0.04789
ECYMOBHPOP	-	-	5-Year Mobility	0.04789
ECYAIDHPOP	-	-	Indigenous Identity	0.04788
ECYRIMHPOP	-	-	Recent Immigrants (2017Present)	0.04788

Table 5: Top Negative Loadings for PC1

Variable	HHS Description	HHS Type	DemoStats Category	Loading
ECYMTNMED	-	-	Maintainer Age	-0.00437
ECYPFAMED	-	-	Female Population by Age	-0.00413
ECYHFAMED	-	-	Female Household Population by Age	-0.00348
ECYPMAMED	-	-	Male Population by Age	-0.00348
ECYHMAMED	-	-	Male Household Population by Age	-0.00282

Table 6: Top Positive Loadings for PC2

Variable	HHS Description	HHS Type	DemoStats Category	Loading
HSSH037A	Wood/Fuel for Heating	Consumption	-	0.11518
ECYHTAMED	-	-	Household Pop. by Age	0.11505
ECYPTAMED	-	-	Total Population by Age	0.11242
ECYHMAMED	-	-	Male Household Population by Age	0.10444
HSSH034	Other Fuel	Consumption	-	0.10254

Table 7: Top Negative Loadings for PC2

Variable	HHS Description	HHS Type	DemoStats Category	Loading
ECYNCA_18P	-	-	Citizenship	-0.09457
ECYVISVM	-	-	Visible Minority Status	-0.09315
ECYHOMNOFF	-	-	Language Spoken Most Often At Home	-0.09130
ECYNCANCIT	-	-	Citizenship	-0.09076
ECYPIM1621	-	-	Period of Immigration	-0.08743

Table 8: Top Positive Loadings for PC3

Variable	HHS Description	HHS Type	DemoStats Category	Loading
HSWH040S	Net Purchase Price of Residences	Consumption	-	0.13164
HSTE001ZBS	Non-current Consumption	Consumption	-	0.12020
HSSH033A	Natural Gas (Owned Residence)	Consumption	-	0.11436
HSSH033	Natural Gas	Consumption	-	0.10837
ECYCHAHHCH	-	-	Children at Home by Age	0.10319

Table 9: Top Negative Loadings for PC3

Variable	HHS Description	HHS Type	DemoStats Category	Loading
ECYSTYAPU5	-	-	Structure Type	-0.13074
ECYMOTFREN	-	-	Mother Tongue	-0.12838
ECYSTYAPT	-	-	Structure Type	-0.12515
ECYHOMFREN	-	-	Language Spoken Most Often At Home	-0.12314