# Canadian Households

DS 3000: Introduction to Machine Learning

cboulos4: Christopher Boulos, kpaskara: Kibi Paskaran, smajidli: Samad Majidli, sahme473: Syed Mohammed Ahmed

Word Count: 2,620

## Executive Summary

This project explores Canadian household behavior using data from DemoStats 2024 and HouseholdSpend 2024. The goal was to understand how different types of households behave and to build models that predict how much income is spent on pensions and insurance.

In the first part of the project, clustering methods like K-Means, PCA, and UMAP were used to group households based on similar characteristics. These methods helped uncover patterns without using pension and insurance data directly, allowing for an unbiased view of household differences. K-Means was used to form clusters, and PCA was applied to reduce the number of features and visualize the results. UMAP, a non-linear dimensionality reduction technique, was also used to create a more detailed view of the clusters. This helped reveal smaller or more complex groupings that PCA couldn't fully capture.

Before applying these methods, the data was cleaned by removing columns with no useful information, dealing with outliers, and handling missing values. Once cleaned, the datasets were merged and scaled to ensure everything was on the same scale. After clustering, the clusters were interpreted based on the characteristics of the households in each group, such as household size, mobility, and digital engagement.

In the second part of the project, regression models were used to predict the proportion of household income spent on pensions and insurance. Two models were tested: Elastic Net, a linear regression model with regularization, and XGBoost, a non-linear tree-based model. After preparing the target variable and cleaning the data again for modeling, both models were trained and tested. XGBoost performed better than Elastic Net, showing that it was better at handling the complex patterns in the data.

To understand why XGBoost worked better, SHAP values were used. SHAP identified the most important features in the model and showed how they influenced predictions. Age, income, and household type were key factors. The SHAP results also confirmed that household spending behavior is non-linear, which is why models like XGBoost are more effective for this kind of prediction.

## Table of Contents

## List of Figures

## 1. Clustering and Dimensionality Reduction

### 1a. Data Cleaning

First, the columns related to “pension” and “premium insurance” were dropped from the “HouseholdSpend.csv” file as they are the target variables to be predicted in the model.

After that, the columns consisting only of zeroes were dropped from both datasets, as they have no variance and provide no valuable information, which could negatively affect clustering and model results if kept.

Invalid outliers were dealt with by first analyzing the columns that contained values less than 0 in both datasets. After analyzing the total number of values less than 0 in those columns, it was found that column ***HSTT001*** only had 28 instances of values less than 0. On the other hand, the other columns each had over 15,000 instances of values less than 0, which indicates that having negative values is to be expected in those columns. Further analysis of the columns result in the same conclusion:

- Column ***HSTT001*** from the HouseholdSpend dataset: Stores information about the "Total Expenditure" of households. Since total expenditure of households cannot be negative (it being negative means that there was a refund, which should not be counted as an expense), any row with the invalid outlier should be removed.
- Column ***HSTE001ZBS*** from the HouseholdSpend dataset: Stores information about the "Total non-current consumption" of households. It involves major household purchases, like homes, vehicles, renovations, etc... Negative values could represent capital gains/losses, so do not delete the rows containing negative values.
- Columns ***HSWH040S***, ***HSWH041S***, and ***HSWH042S*** from the HouseholdSpend dataset: Store information about the "Net" purchase price of properties or residences. Since "Net" is typically a difference between total spending and refunds, sales, or returns, negative values are expected. So, keep the rows with negative values in those columns.

Null values - The DemoStats dataset is the only dataset with columns that contain null values. Those columns ("**ECYPTAMED**", "**ECYPMAMED**", "**ECYPFAMED**", "**ECYHTAMED**", "**ECYHMAMED**", "**ECYHFAMED**", and "**ECYMTNMED**") represent the median ages of various kinds of populations. Even though the number of null values of

each column is not small, replacing them with the median of their respective column is a better option than dropping them, as they seem important.

Extreme Outliers were dealt with by calculating the z-scores of each column and removing the outliers by deleting the rows they are in when their z-score is not within 3 standard deviations of the mean.

The two datasets were merged using an inner join on the first two columns, "**CODE**" and "**GEO**", since these columns are shared and help match the same geographic areas. An inner join was used to make sure that only the rows that had data in both datasets were kept. This way, only the relevant data was worked on.

The first two columns ("**CODE**", and "**GEO**") of the merged dataset were dropped since they are non numerical identifiers and should not be included. Also, since the merged dataset only contains ***int*** and ***float***-type values, they were all changed to type ***float*** for consistency. The merged data was then scaled so that all the data would be on the same scale before running the models.

### 1b. K-Means Clustering

Using the K-Means method to perform clustering on the scaled data, setting random_state = 42, then visualizing it using the Elbow method to find the optimal number of clusters:

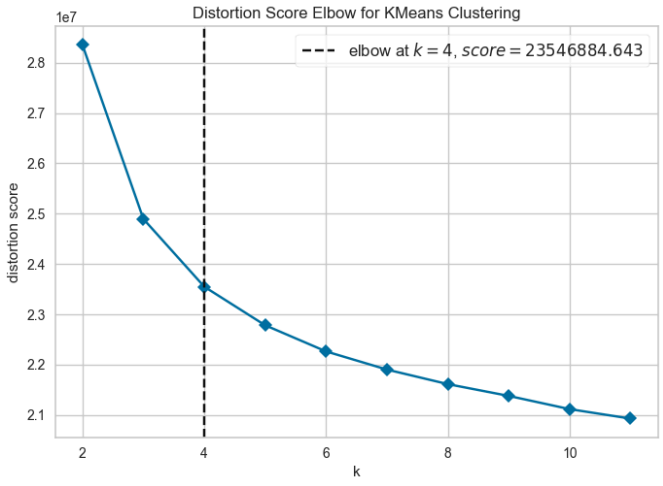


***Figure 1**: Elbow method showing distortion for different values of k*

- The Elbow method resulted in an optimal number of clusters **k = 4**.

Using the K-Means method to perform clustering on the scaled data, setting random_state = 42, then visualizing it using the silhouette method to find the optimal number of clusters resulted in:

- With 3 clusters (k = 3), the silhouette score is **0.337**

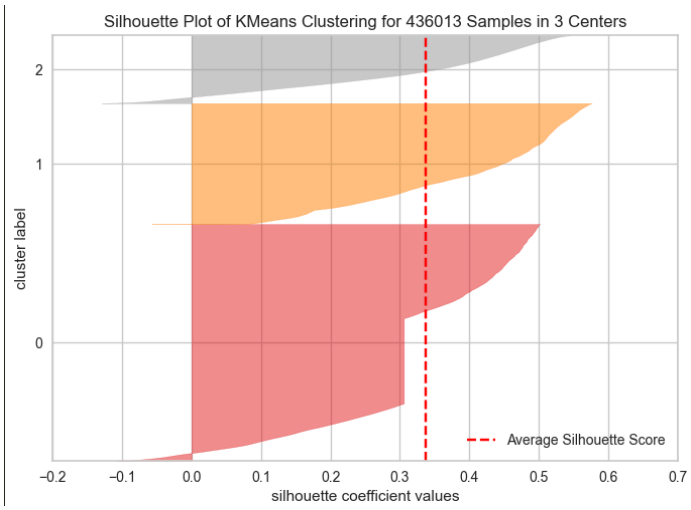


***Figure 2a:*** *Silhouette Plot for k = 3*

- With 4 clusters (k = 4), the silhouette score is **0.389**

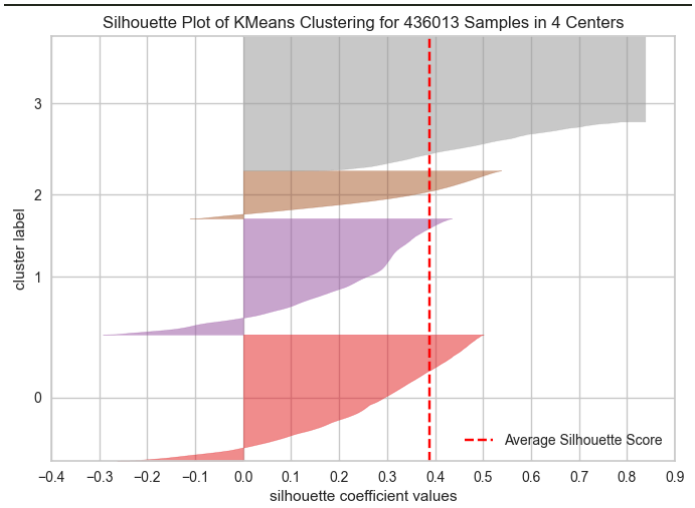


***Figure 2b:*** *Silhouette Plot for k = 4*

- With 5 clusters (k = 5), the silhouette score is **0.447**

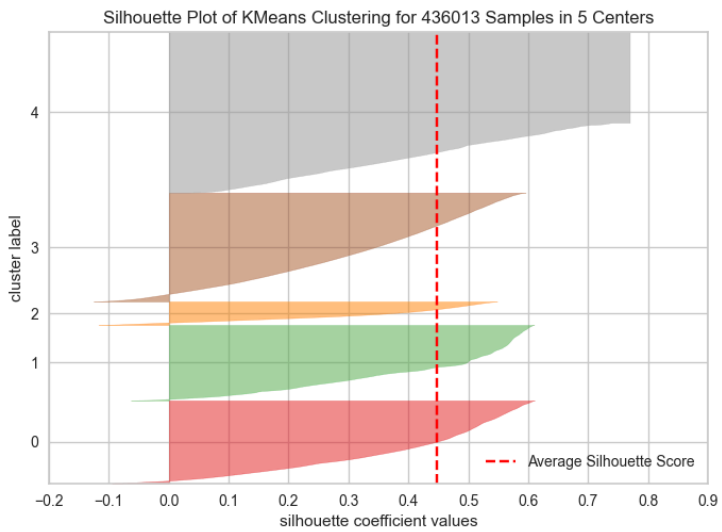


***Figure 2c:*** *Silhouette Plot for k = 5*

- With 6 clusters (k = 6), the silhouette score is **0.456**

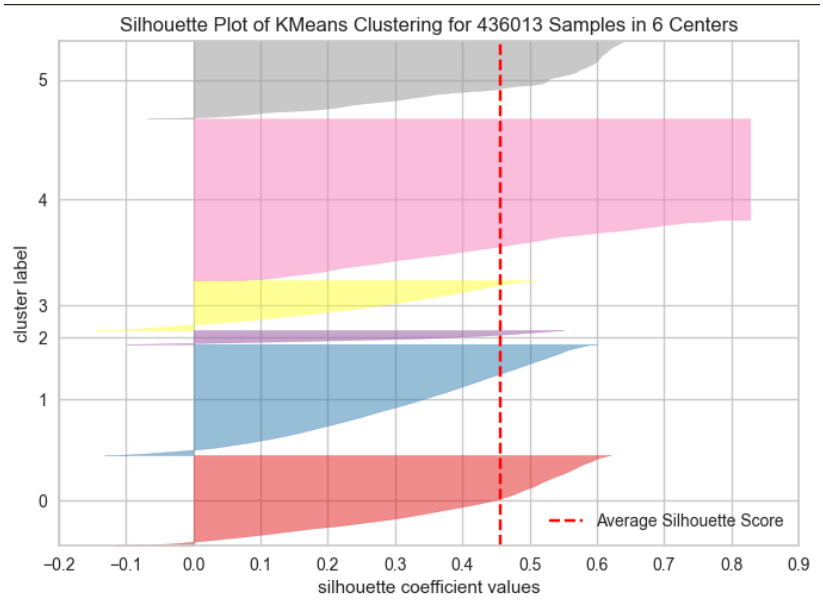


***Figure 2d:*** *Silhouette Plot for k = 6*

- With 7 clusters (k = 7), the silhouette score is **0.448**

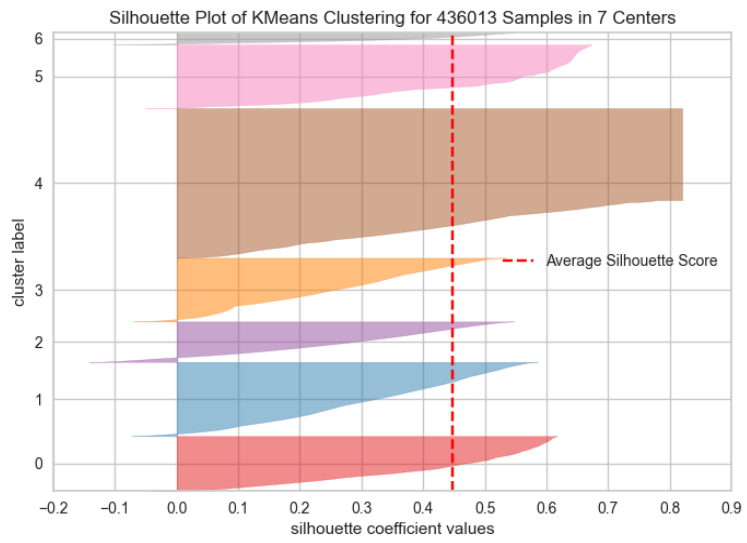


***Figure 2e:*** *Silhouette Plot for k = 7*

∴ Since the silhouette score is the highest with 6 clusters (k = 6), the clusters are therefore the most well separated and well-defined when 6 clusters are used.

∴ **Thus**, the elbow method and the silhouette method do not agree since the optimal number of clusters found using the elbow method is 4, whereas the optimal number of clusters found from the silhouette method is 6.

### 1c. Principal Component Analysis (PCA)

**c.i.** PCA was applied to reduce the dataset's dimensionality. The first two principal components were PC1 and PC2. They accounted for approximately 46.7% of the total variance. 43.24% from PC1 and 3.45% from PC2. A scatterplot of these components was made. They were colored by K-Means cluster labels which showed several groupings. PC1 influenced by population-based household features. For example: ECYMOBHPOP ECYTIMHPOP. This created horizontal separation. PC2, which was shaped by technology and media engagement features. For example: ECYTIMA, ECYMOTNOFF. This created some vertical distinction. This scatterplot highlights underlying trends and it also shows areas of overlap. This suggests the need for non-linear techniques for finer resolution.

**c.ii.** The clusters defined by K-Means were created using the first two PCs (in part c.i.). Cluster 1 is positioned positively along PC1. This indicates high values in mobility and household population features. This suggests active or larger households. Cluster 0 lies negatively along PC1. This shows lower engagement in those same variables as Cluster 1. This probably shows quieter or less mobile populations. Cluster 2 is centered compared to the other 2. This shows a more balanced behavior. Cluster 3 is defined more by PC2 which means more digitally active users regardless of household size. Although there is overlap, PCA allows for a more general interpretation of user types. Its linear nature limits full separation.

**c.iii.** When examining the average values of the first three PCs across clusters. We assigned meaningful labels. Cluster 1 which has high PC1 values which are household population metrics, is labeled **Family Households**. Cluster 0 which has low PC1 and near-zero values on other components. It is Low Engagement Households. Cluster 2 which is centered around average PC1 and slightly negative PC2 values. It is **Balanced Households** which shows moderate activity across variables. Cluster 3, which has high PC2 and PC3 values. It is driven by digital and rental/housing features. It is named **New Gen Households** which suggests younger or digitally inclined groups in rental housing. These names provide identities rooted in PCA contributions.

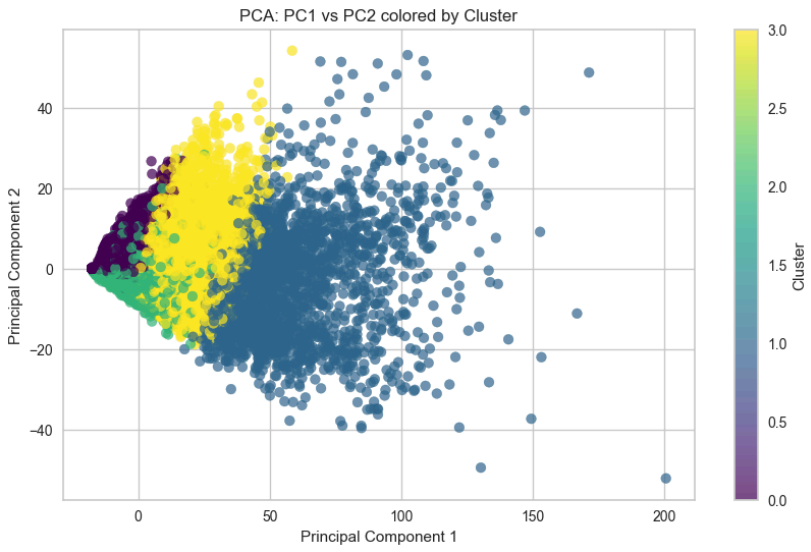


***Figure 3:*** *PCA scatterplot of first two components colored by K-Means cluster labels*

**1d. UMAP Analysis**

To apply UMAP for dimensionality reduction. Experimented with several parameter combinations to optimize cluster separation. After testing many values. We selected n_neighbors=15 and min_dist=0.1, with metric='euclidean'. This configuration balanced the local and global structure and it also maintained cluster compactness. A lower min_dist allows tighter groupings. It enhances visual separation. The n_neighbors value of 15 made the algorithm get sufficient neighborhood context without over-smoothing. This was ideal for datasets with moderate noise and overlapping traits. The same cluster labels were used to derive from K-Means, the UMAP projection in 2D. The result was a visualization that showed more distinct and compact clusters than PCA. It was especially helpful in identifying smaller or nuanced groups. UMAP's nonlinear capabilities helped it to uncover latent structures. Which was not captured by linear methods like PCA. Which makes it a powerful complement for exploratory data analysis in multi-feature datasets.

When comparing UMAP to PCA, UMAP provided a more refined visual clustering of the data compared to PCA. PCA revealed general trends using linear projections. Which were mainly separating clusters along axes dominated by household population and digital engagement. UMAP captured more nuanced group boundaries and better showed local density variations. PCA was helpful for understanding variable influence, it still struggled with cluster compactness due to its linear nature. UMAP, on the other hand, did very well at preserving both local and global structures. Which helped reveal subclusters and finer distinctions. Therefore, UMAP serves as a more intuitive and interpretable visualization tool for this dataset. Mostly when it comes to validating clustering performance it is the better tool. Overall, UMAP complements PCA; it does this by offering a clearer picture of group separation and internal cohesion. It also helps by reinforcing our interpretations from earlier clustering and component analyses.

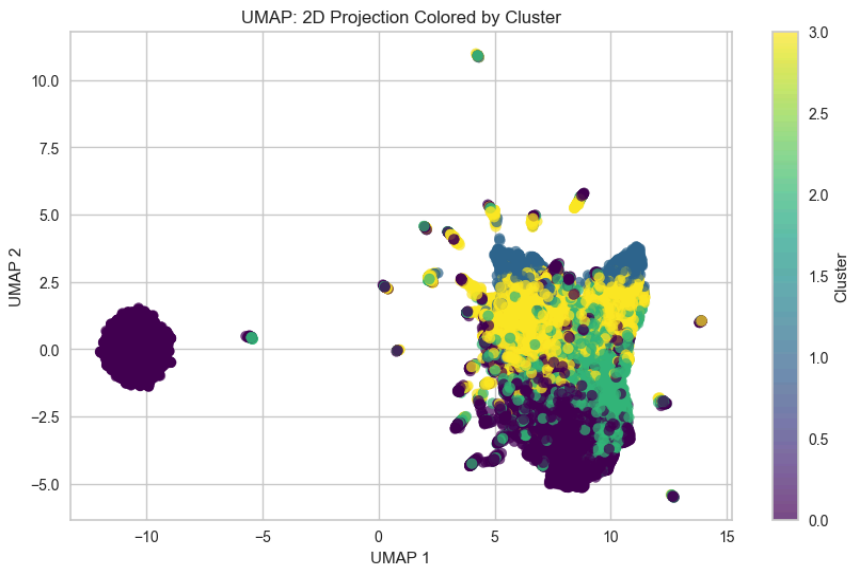


***Figure 4**: UMAP 2D projection using K-Means cluster labels*

## 2. Regression Modeling

### 2a. Elastic Net Regression

This section of the analysis focused on predicting household expenditure on pensions and insurance as a proportion of total household income. To achieve this, two regression methods were implemented: Elastic Net, a regularized linear regression method, and XGBoost, a gradient-boosted tree algorithm. The performance of these models was then compared.

**a.i.** First, the target variable was constructed as the ratio between total household spending on pensions and insurance, and the total household income. Since pension and insurance spending columns had previously been removed during the data-cleaning step due to their exclusion from clustering and PCA analyses, the original household spending dataset was re-loaded briefly to calculate this target variable. After computing the target, these original columns were dropped immediately to avoid their unintended use as predictors. The calculated target was then merged onto the cleaned demographic and spending dataset, resulting in a clean and finalized dataframe for regression analysis.

**a.ii.** The data was split into training and testing sets (80% train, 20% test) to evaluate how well each regression model generalized to unseen observations. Prior to model training, a pipeline was used to handle missing data (median imputation) and standardize all features, ensuring that feature scales did not disproportionately affect the model outcomes.

**a.iii.** For the Elastic Net model, hyperparameter optimization was performed using randomized search cross-validation to identify the most effective combination of regularization strength (alpha) and the balance between L1 and L2 penalties (l1_ratio). The best hyperparameters found were:

- alpha ≈ 0.014
- l1_ratio = 0.9

**a.iv.** Elastic Net yielded limited performance, with an $R^2$ score of **-0.00002** and a mean squared error (MSE) of approximately **0.00032**. This result indicates that the linear model was unable to effectively capture the complex, nonlinear relationships within the dataset.

**2b. XGBoost Regression**

**b.i.** XGBoost, known for effectively modeling complex and nonlinear patterns, was applied next. To minimize overfitting and improve predictive generalization, early stopping was employed, which automatically halted training once validation accuracy stopped improving. XGBoost produced substantially improved results compared to Elastic Net:

- MSE ≈ 0.00009
- $R^2$ ≈ 0.72

This $R^2$ score indicates that XGBoost successfully explained approximately 72% of the variability in pension and insurance expenditures relative to household income, demonstrating significant predictive capability.

**b.ii.** The analysis clearly indicated that tree-based, nonlinear methods such as XGBoost significantly outperform linear methods such as Elastic Net for predicting pension and insurance expenditure ratios. Given this substantial improvement in accuracy and generalization, gradient-boosted models are recommended for future modeling efforts

on similar datasets. Further model tuning and optimization could provide additional predictive performance gains.

**b.iii.** The scatterplot comparing actual versus predicted values for the XGBoost model revealed a strong alignment along the diagonal line, indicating that the model made accurate predictions for a wide range of data points. The XGBoost model achieved a Mean Squared Error (MSE) of **0.000102** and an $R^2$ score of **0.689**, showing it captured approximately 69% of the variance in household pension and insurance expenditures.

In contrast, the Elastic Net model produced a significantly higher MSE of **0.00031689** and an $R^2$ score of **-0.00002**, indicating it failed to model the underlying relationships effectively.

**∴** These results confirm that XGBoost provided more accurate and reliable predictions compared to the linear Elastic Net approach.

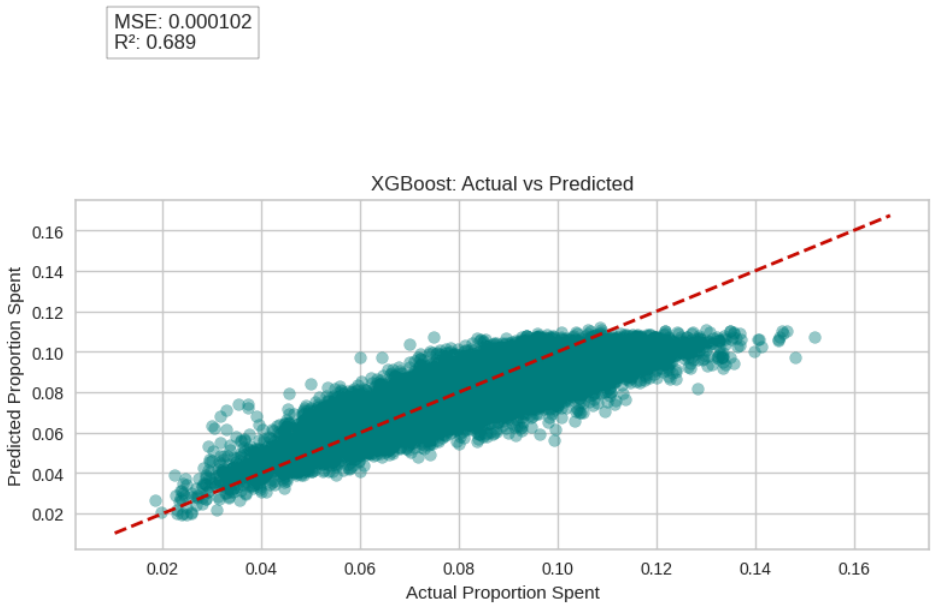


***Figure 6:*** *XGBoost actual vs predicted pension and insurance spending*

### 2c. SHAP (SHapley Additive exPlanations)

SHAP values were used to interpret the XGBoost model's predictions. The top 5 features based on the SHAP summary plot are:

1. **ECYHNIX150** (household income between $125K–$149K)
2. **ECYMTN85P** (maintainers aged 85+)
3. **ECYHOMSING** (single responses)
4. **ECYMTN7584** (maintainers aged 75–84)
5. **HSHE021** (rental of heating equipment).

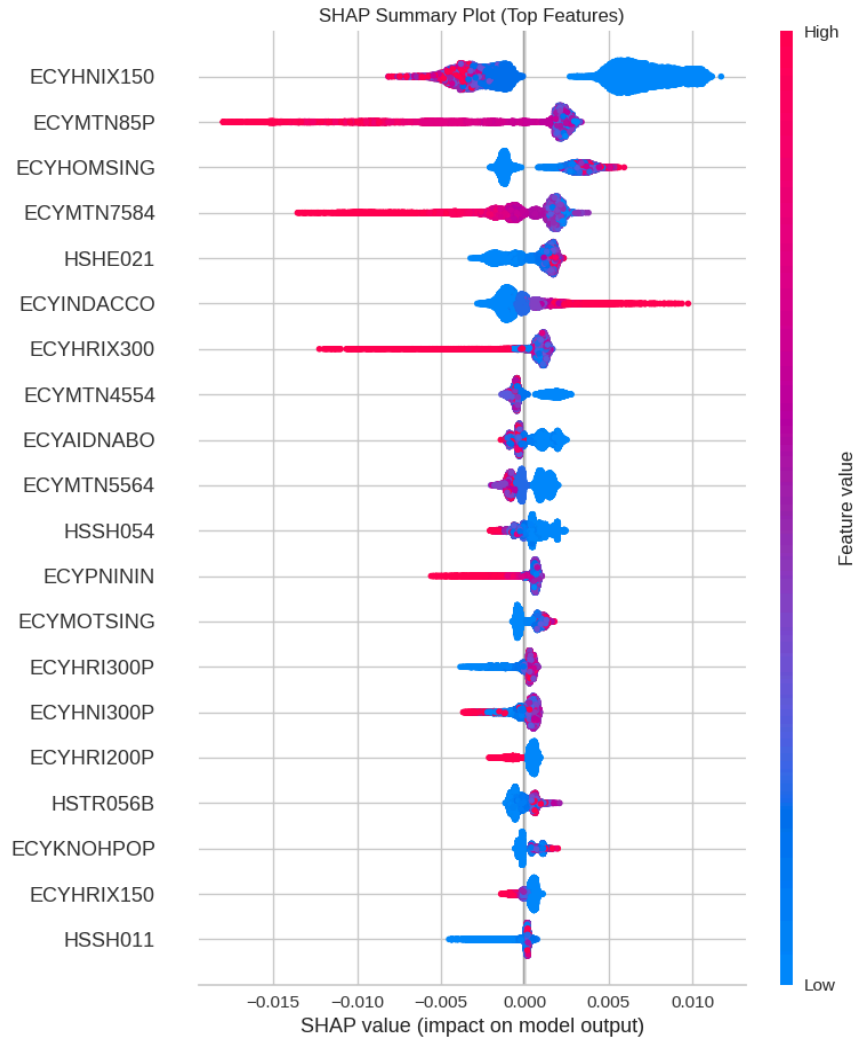


***Figure 7**: SHAP summary plot showing feature importance*

- **ECYHNIX150** shows a clear inverse relationship with the target. As income in this range increases, the SHAP values decrease, suggesting that higher-income households tend to spend a smaller share of their income on pensions and insurance.

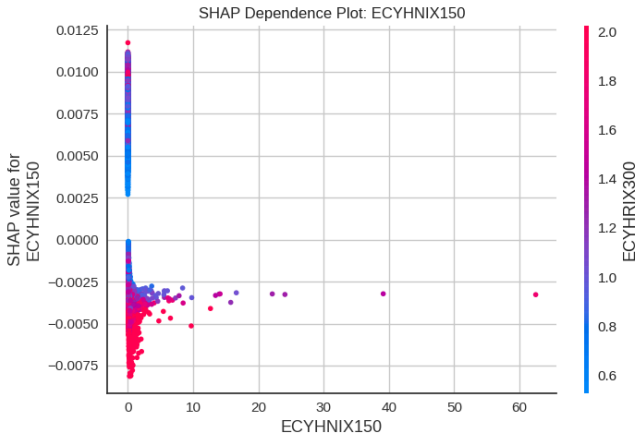


***Figure 8a:*** *ECYHNIX150 – Household Income $125,000–$149,999*

- **ECYMTN85P** has a strong negative impact at higher values, meaning areas with more seniors aged 85+ tend to spend less proportionally on pensions and insurance.

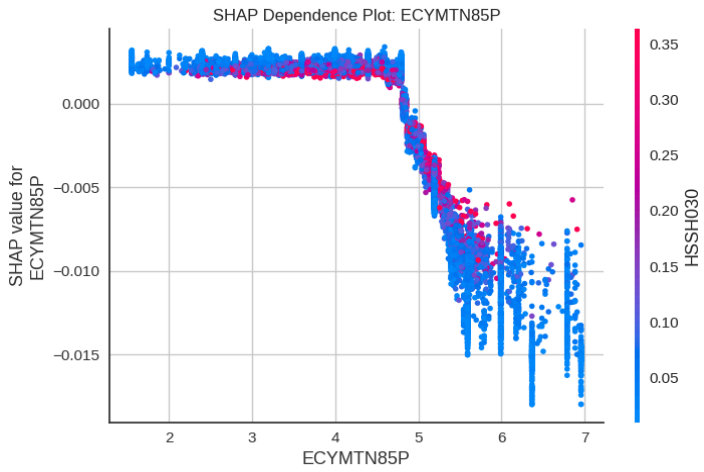


***Figure 8b:*** *ECYMTN85P – Maintainers 85 or Older*

- **ECYHOMSING** is positively correlated, indicating that areas with more single-person households see higher proportional spending.

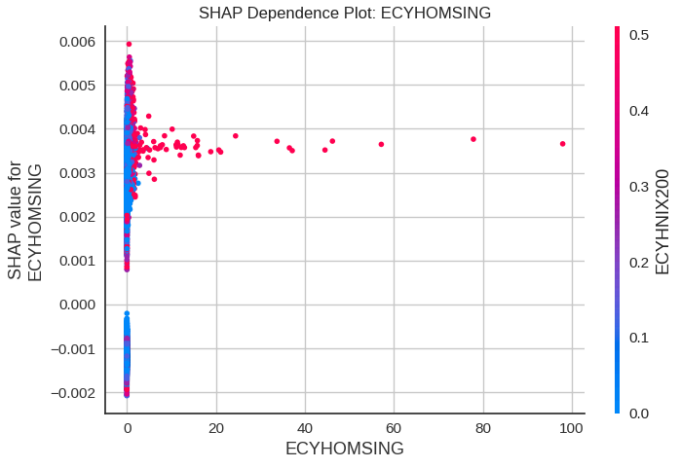


***Figure 8c:*** *ECYHOMSING – Single Responses*

- **ECYMTN7584** mirrors the trend of ECYMTN85P, with SHAP values dropping as the proportion of elderly maintainers increases, further supporting age-based variation in spending behavior.

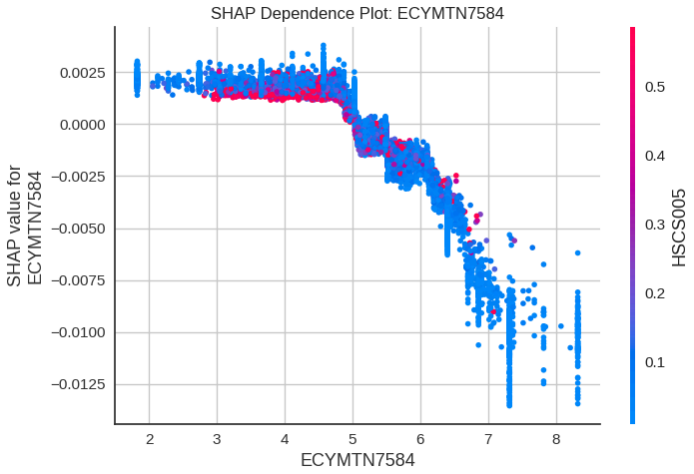


***Figure 8d:*** *ECYMTN7584 – Maintainers Aged 75 to 84*

- **HSHE021**, related to rental of heating equipment, shows a positive association, suggesting that households incurring these specific utility-related expenses may also allocate more income toward financial services like insurance and pensions.

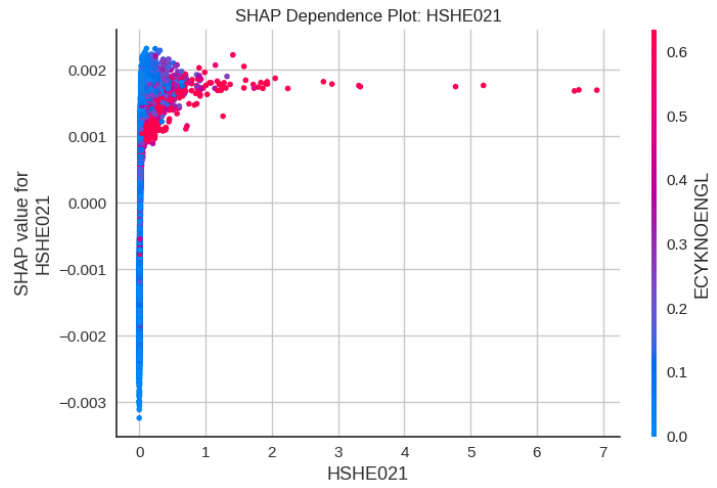


***Figure 8e:*** *HSHE021 – Rental of Heating Equipment*

∴ Compared to the Elastic Net model's coefficients, SHAP provides clearer, feature-level interpretations with nonlinear interactions. While the linear model struggled to capture these dynamics, SHAP shows that the model behavior changes at different ranges of the same feature. These nonlinear patterns support the conclusion that pension and insurance spending is influenced by complex, nonlinear relationships in the data. Tree-based methods like XGBoost are better suited for capturing these effects.