Western Science

The University of Western Ontario
Department of Statistical & Actuarial Science
DS 3000b: Introduction to Machine Learning

Coursework – Canadian Households

Submitted by:
Aarish Lakhani, 251088464,
Agasthya Nitturi, 251103633,
Azardokht Aryaei, 251015295
Zaeem Ajwad, 251283478

Instructor: Dr. Cristián Bravo

20-April-2025

Table of Contents

List of figures

List of tables

Executive Summary

This document focuses on studying Canadian households and their spending/saving habits. For this, two datasets have been provided by Environics Analytics:

- Household spending per dissemination area
- Demographic data per dissemination area

The analysis is divided into two Parts: unsupervised learning (clustering and dimensionality reduction), and supervised learning (regression).

In the first part, we cleaned the data and used K-means clustering to recognize the distinct segments within the merged household spending and demographic dataset. The optimal number of clusters has been calculated by using the Elbow method and the Silhouette method. To further explore the structure of the dataset, we implemented the Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP).

The second part focuses on the development of models for predicting a household's proportion of income spent on total personal insurance premiums and retirement/pension contributions. We employed two predictive models: elastic net linear regression and the XGBoost model. Both models were evaluated in terms of mean squared error (MSE), R-squared (R^2) and bootstrapped confidence intervals, and XGBoost significantly outperformed the elastic net linear regression. Furthermore, we employed the Shapley Additive exPlanations (SHAP) values method to identify the most important variables contributing to the household spending behavior.

1. Clustering and Dimensionality Reduction

1.a. Data Cleaning and Preprocessing

Extensive data cleaning was performed to address null values, negative values, string data types, outliers and correlated variables. Following preprocessing, the two datasets were combined into a single dataset for analysis. To ensure efficient processing, a 10% sampling was selected due to the large dataset size. Then, scalar standardization was applied to ensure all data are on the same scale, avoiding any feature dominating the clustering process.

According to the metadata files, the hierarchical format of features is clearly represented. This information helped to remove the main category columns and maintain the most granular level of information.

To address the effect of outliers, Winsorization, a technique to replace too low or too high values with a threshold value, was used. The first and third quartile values were used to calculate the Interquartile Range (IQR):

- Q1 (first quartile) = the value below which 25% of the data lies
- Q3 (third quartile) = the value below which 75% of the data lies
- IQR = Q3 - Q1

The boundary values were calculated as:

- lower bound = $Q1 - 1.5 \times IQR$
- upper bound = $Q3 + 1.5 \times IQR$

Any point below the lower bound was set to the lower bound, and any point above the upper bound was set to the upper bound. This way, the effect of outliers was minimized without deleting the data.

After preprocessing the DemoStats and Household sets individually, they were merged, and a 10% sample was selected. The data was standardized, and a correlation matrix was used to identify highly correlated features. Features with correlation coefficients above a threshold of 0.95 were marked as strongly correlated. From each highly correlated pair, only one column is removed to reduce redundancy without losing all the information.

1.b. Clustering

To determine the optimum number of clusters in the dataset, a K-Means clustering analysis was conducted and assessed with the silhouette method as well as the elbow method. The data have been scaled before clustering to satisfy the requirement that all variables weigh the same. The silhouette approach indicated that the best number of clusters is two, whereas the elbow technique suggested that four clusters are optimal.

The silhouette method evaluates how well an object is grouped within its own cluster in contrast to other clusters. The silhouette score (figure 1) ranges from -1 to +1: -1 means a data point is very poorly matched with its cluster; 0 signifies the datapoint is equally close to two clusters; +1 indicates the point is well matched to its cluster. In this instance, the highest silhouette score (0.9675) was found with two clusters, indicating that the data is organized into two distinct clusters.

Conversely, the elbow method (figure 2) assesses the within-cluster sum of squares (WCSS) as the number of clusters increases. At the "elbow" point, WCSS stops decreasing significantly, indicating that the increase of clusters after this threshold does not effectively improve clustering and adds to the complexity of the model. In this example, the elbow point was found in four clusters.

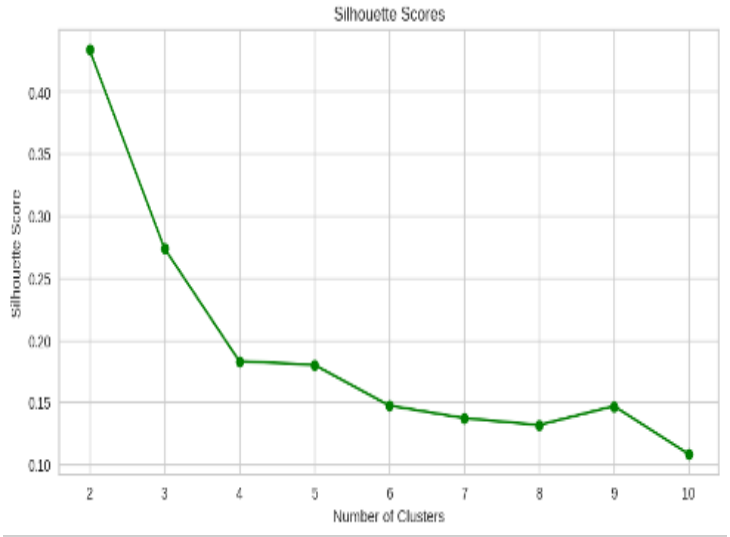

Figure 1. Optimum clusters in the Silhouette method

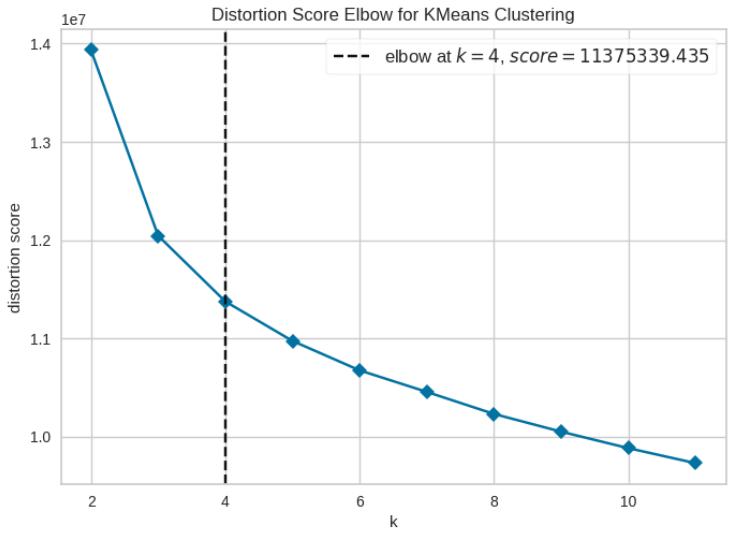

Figure 2. Optimum number of clusters in Elbow method

The two methods provided different optimum numbers of clusters, so they do not completely agree. The silhouette technique focuses on the importance of separating clusters, typically supporting a smaller number of clusters, whereas the elbow technique concentrates on reducing intra-cluster variance, usually preferring a larger number of clusters. This difference emphasizes that identifying the ideal number of clusters may be subjective and the best choice depends on whether tightly-packed clusters or well-separated clusters are important in the context.

1.c. PCA and Cluster Analysis

Our PCA Analysis reveals definitive attributes of our consumer base and their demographics. As per the explained variance ratio on Figure 3 below, we decide to focus on PC1 and PC2 (representing 56.69% and 4.96% of the total variance over the sample, respectively), and occasionally PC3 for our analysis.

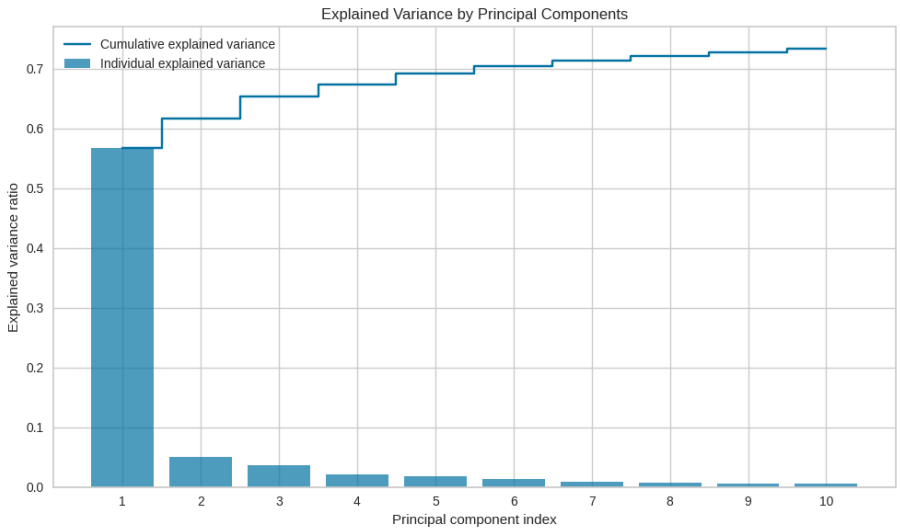

Figure 3. Explained variance ratio by PCs

To start, our scatterplot (see Figure 4 below) depicts a non-linear relationship with well-defined clusters that indicate certain consumer spending patterns. For example, the leftmost cluster (cluster 0) converges around the origin, hinting at households with average spending patterns, with little variance. To the right of it, cluster 3 is slightly more spread out, hinting at a more diverse spending pattern with a bit more variance among them. To the right of that, cluster 2 has the most variance indicating a non-stable spend pattern with some of the highest and lowest traits. Lastly, the rightmost cluster (cluster 1) has higher minimum spend habits than the neighboring cluster 2, and also a higher max: hinting at just as much variance but better spend patterns. With this we draw that PC1 likely represents total household spending. Higher values correspond to high spending households, while lower values indicate more frugal/low income households. PC2 might be showing types of spending (essential vs. non) or urban vs. rural spending habits. Some clusters show high separation along this axis.

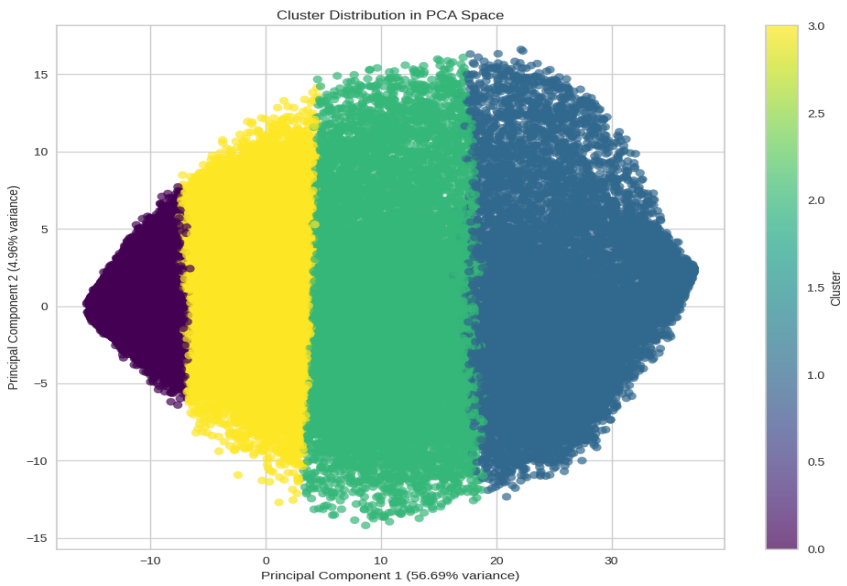

Figure 4. Cluster scatter plot for distribution among PCAs.

Lastly, Figure 5 below gives us hints about the cluster attributes:

- Cluster 0: "Consistent Non-Spenders or lower-earners (Minimal Engagement)”
    - Significantly negative PC1 values (-11.57)
    - Near-zero PC2 values, suggesting average spend
    - Slightly negative PC3 values
- Cluster 1: "High Spenders or Power Users" (High PC1 Dominant Group)
    - Extremely high positive PC1 values (25.64), indicating highest position on primary component
    - Slightly positive PC2 values
    - Most negative PC3 values (-0.59), suggesting distinctive spending
- Cluster 2: "Middle-High Income Urban Households”
    - Moderately high PC1 values (10.25), suggesting above-average habits
    - Most negative PC2 values (-0.37) indicating distinctive least spend

- Positive PC3 values (0.43)

- Cluster 3: "Middle-Income Diverse Households”
  - Slightly negative PC1 values (-2.31), placing it between Clusters 0 and 2
  - Most positive PC2 values (0.21)
  - Similar PC3 values to Cluster 2 ((0.41)

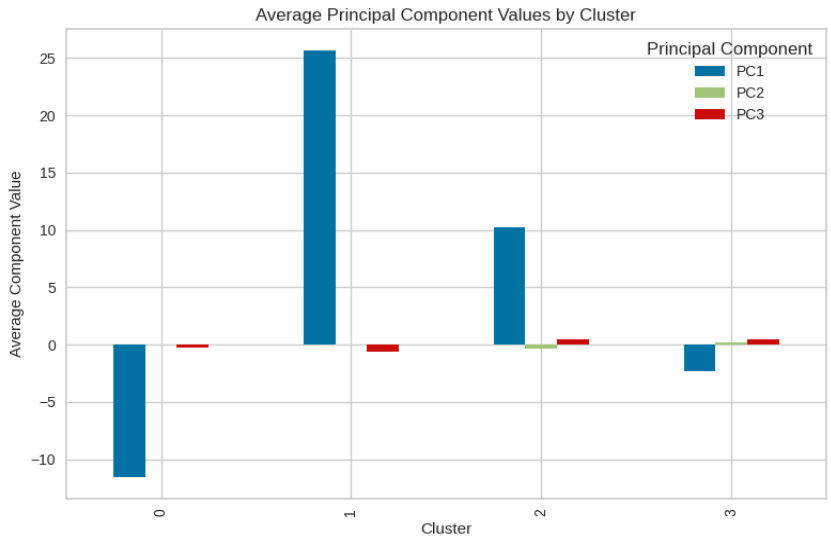

Figure 5. Average PC values by cluster

1.d. Nonlinear Dimensionality Reduction

To find the best UMAP parameters, we ran a quick silhouette-score grid search on a 10 000-point subsample. We tested four (n_neighbors, min_dist) pairs and got the following values.

Table 1. Silhouette-score grid search on a 10 000 point subsample

n_neighbors	min_dist	silhouette
5	0.00	0.1578
15	0.10	0.1920
30	0.05	0.2112
50	0.30	0.2377

Since (50, 0.30) achieved the highest silhouette (0.2377), we used those settings on the full dataset as seen in the figure 6.

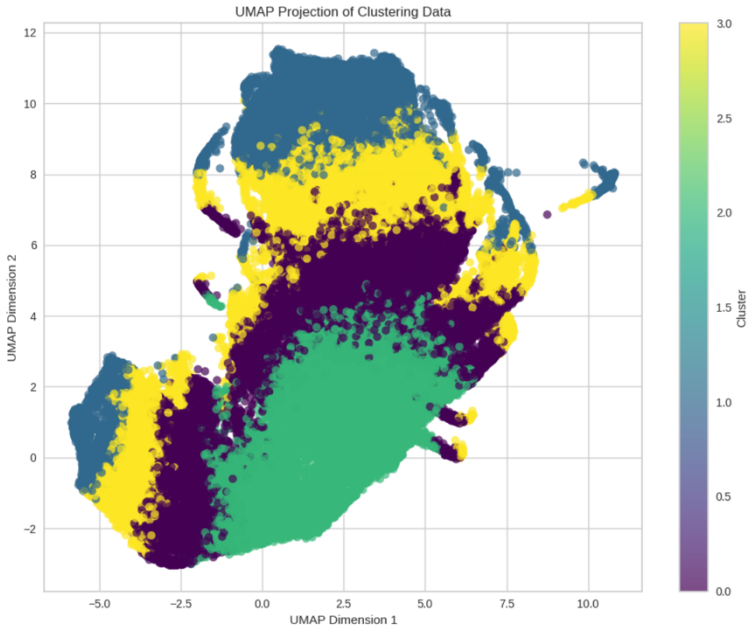

Figure 6. UMAP Plot for n_neighbors=50 and min_dist=0.30

In the UMAP plot above , each of the four K-Means clusters appears as its own color:

- Purple (Cluster 0): a narrow “stripe” cutting across the center, which is small but a well-defined group.
- Green (Cluster 2): the largest blob in the lower-right, which describes many households sharing core spending characteristics.
- Blue (Cluster 1): the top and left bottom region, households with more varied patterns.
- Yellow (Cluster 3): wraps around and between purple and blue, an intermediate segment bridging extremes.

This configuration highlights smooth transitions. For instance, purple and yellow touch on both sides of green, revealing the range of spending behavior.

Key drivers behind these shapes include furniture spending (HSHF003), dining out (HSFD994), adult population share (ECYTCA_18P), plus income brackets (ECYHNI2040, ECYHRI2040). Compared to PCA’s straight-line separations, UMAP with (50,0.30) uncovers a

more organic manifold. PCA is great for interpreting component loadings, whereas UMAP shows how clusters blend and separate in a way that better reflects the underlying non-linear relationships in Canadian household data.

2. Regression

2.a. Elastic Net Regression Model

To determine how much of the household's proportion of income spent on total personal insurance premiums and retirement/pension contributions, we set out by dividing total insurance + pension spending (HSEP001S) by total income (HSHNIAGG). We used all 588 standardized spending and demographic variables as inputs. We then split our data into 70% training and 30% testing, and built a two-step pipeline: first scaling the features, then fitting an Elastic Net regression . To find the best balance between the two regularization penalties, we tried five values of α (from 0.001 to 10) and four ratios (from 0.2 to 1.0) using cross-validation. The winning model used α = 0.001 and l1_ratio = 0.2.

When we applied it to our test data, we saw a very low mean squared error (MSE = 0.0001) and an R^2 of 0.3724. Bootstrapping gave a tight 95% confidence interval for MSE ([0.0001, 0.0001]) and R^2 ([0.3641, 0.3803]), showing these results are consistent. As shown in the scatter plot, most points lie close to the red diagonal line—especially in the mid-range of 3%–7% spending—so our predictions are pretty accurate there, though they spread out more at very low and very high values.

Looking at the five largest coefficients, we found that variables such as HSH0002, ECYMTN7584, ECYMTN6574, and ECYHNI200P are associated with a slight decrease in insurance and pension spending, while ECYMTN4554 has a positive effect. The model effectively eliminates weaker predictors by shrinking their coefficients to zero, highlighting only the most influential factors. Elastic Net (figure 7) provides a clear and interpretable model that captures the key drivers of how much households allocate to insurance and pension contributions.

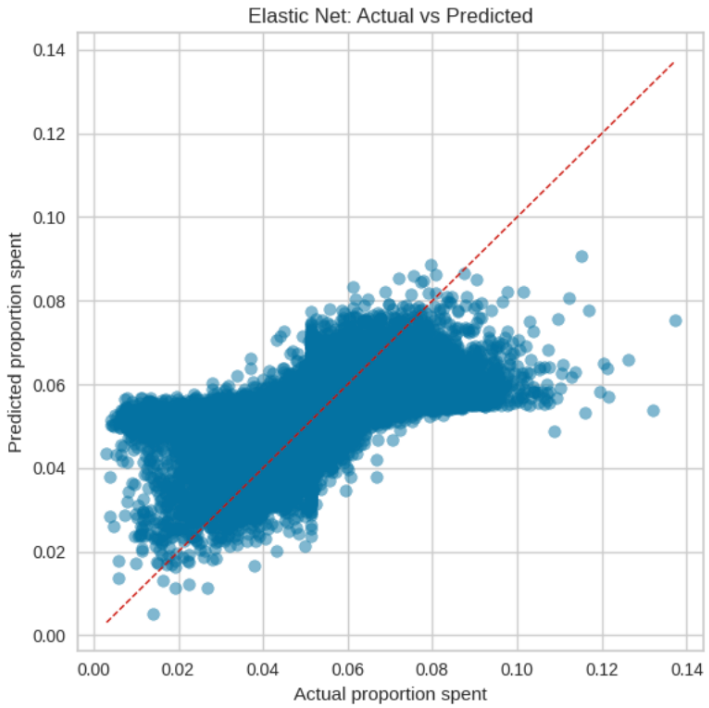

Figure 7. Actual proportion spent v/s predicted proportion spent elastic net

2.b. XGBoost Model Training and Evaluation

A grid search was conducted utilizing GridSearchCVwith 2-fold cross-validation. The size of the hyperparameters set was deliberately set, a small size, to prevent excessive memory consumption while still examining model complexity (max_depth), learning rate (learning_rate), and the number of estimators (n_estimators). A total of eight combinations were assessed, and the optimal configuration was chosen based on the highest negative mean squared error.. This method strikes a balance between computational efficiency and performance optimization.

Comparing XGBoost (figure 8) and linear regression models indicates XGBoost explains 73.5% of the variance in the target variable, whereas the linear model explains 36%-38%, indicating XGBoost model significantly has a better performance. The XGBoost model has a lower RMSE (0.0078 vs. 0.01), showing its predictions are closer to the real values.

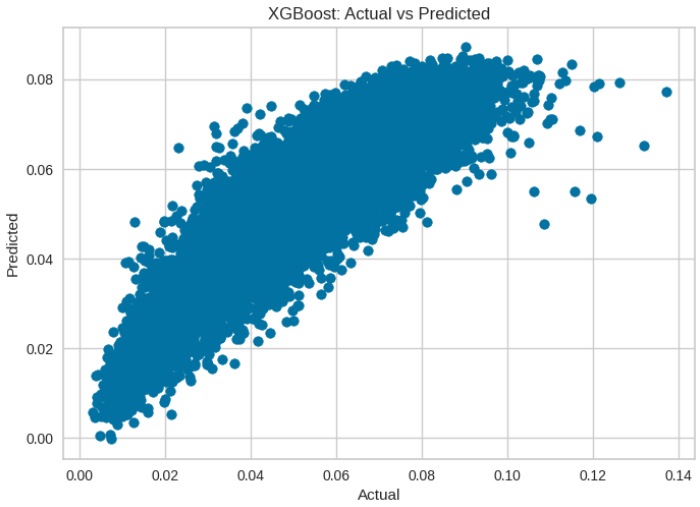

Figure 8. Predicted values vs. actual values in the XGBoost model

2.c. SHAP-Based Interpretation of XGBoost Results

The Figure 9 below allows us to determine the most impactful variables using the Shap Method, with a basis on the variables with the most SHAP impact and the highest (or lowest) Feature Value. From the graph, the 5 most important variables (positive or negative) influencing the model's prediction of proportion spent are:

- ECYHNI200P
- HSHO002
- ECYMTN7584
- ECYMTN6574
- HSCC003

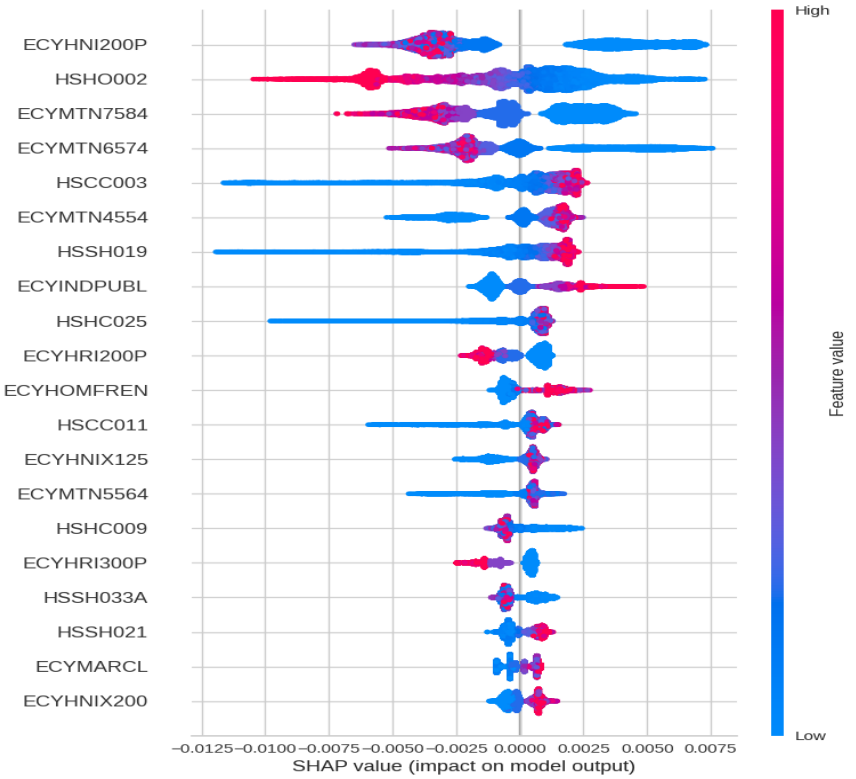

Figure 9. Most impactful variables on spend

To begin analysing these top 5 variables, we may start with ECYHNI200P, which has the strongest average effect on the model output. As per the Figure 10 below, low values (blue) mostly lead to positive SHAP values (higher prediction), while high values (red) tend to push predictions down. As ECYHNI200P increases, its SHAP value tends to decrease, i.e higher values reduce the predicted spending proportion. This variable might represent a metric where higher scores mean less need-based spending, reducing the predicted proportion.

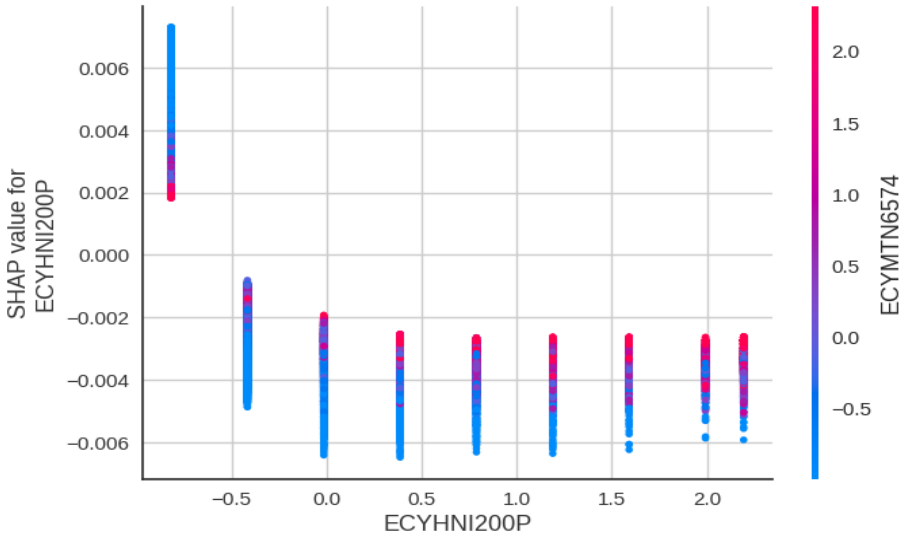

Figure 10. SHAP value for variable ECYHNI200P

Next, HSHO002 is the second most impactful feature. See Figure 11. The higher values (red) clearly reduce the predicted outcome, and lower values (blue) increase it. There's a strong non-linear downward trend, and as HSHO002 increases, the proportion of spending decreases sharply, but this levels off at the high values.

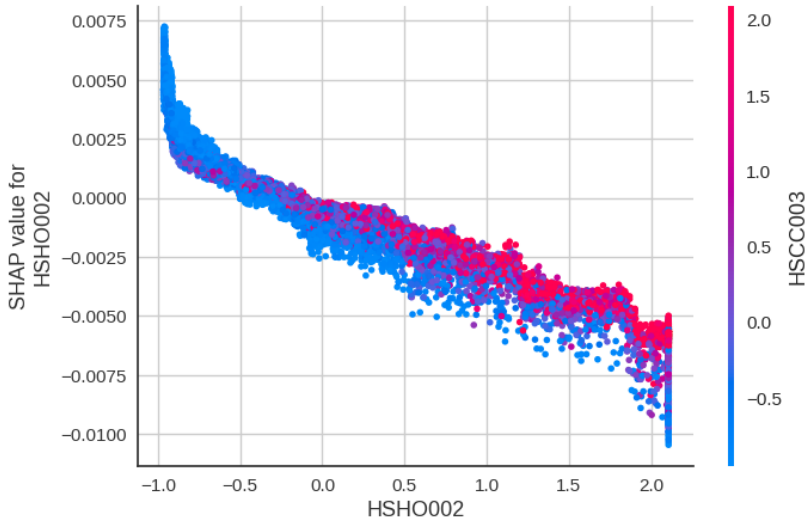

Figure 11. SHAP value for the variable HSHO002

Next we show ECYMTN7584 in figure 12, which contributes negatively to the prediction when its values are pretty high. And as it increases, the SHAP values become consistently

negative, i.e higher values lower the predicted spending. This strong inverse relationship shows that the model is confident that higher values mean reduced spending.

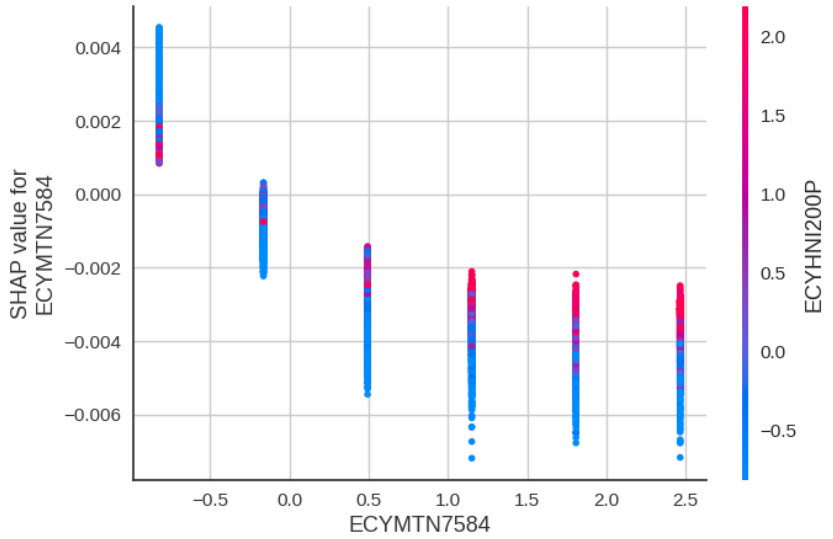

Figure 12. SHAP value for variable ECYMTN7584

Now we show ECYMTN6574 in figure 13, which follows a similar pattern where higher values on this variable yield lower predictions. There's also a big drop in SHAP values with increasing input value. Now, since this variable also contributes negatively, it likely represents a factor associated with reduced economic need/lower spending.

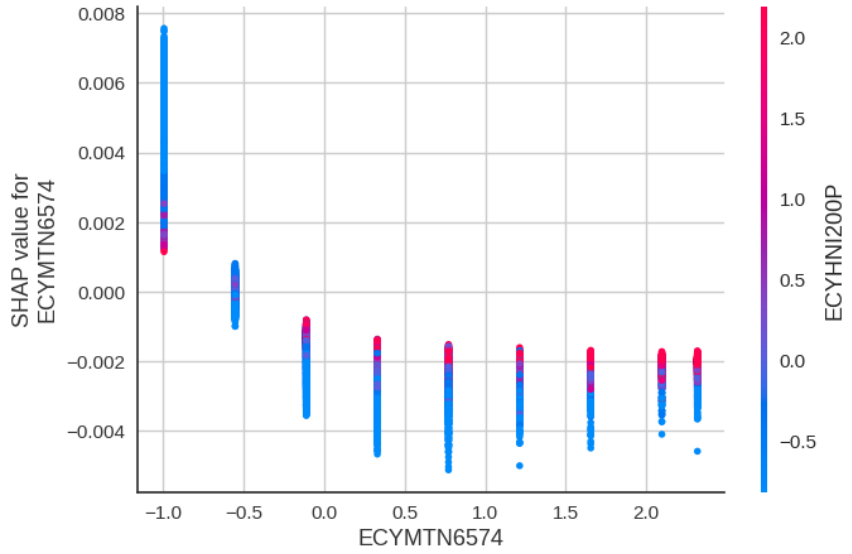

Figure 13. SHAP value for the variable ECYMTN6574

Lastly, we analyse HSCC003 in figure 14, depicting non-linear behavior again. At the lower values, the SHAP contribution is very negative, but as HSCC003 goes up, the SHAP value rises sharply, approaching 0 or positive. This hints that low values of HSCC003 vastly decrease spending, while higher values weaken or rather reverse that effect, giving a very non-linear, saturated pattern.

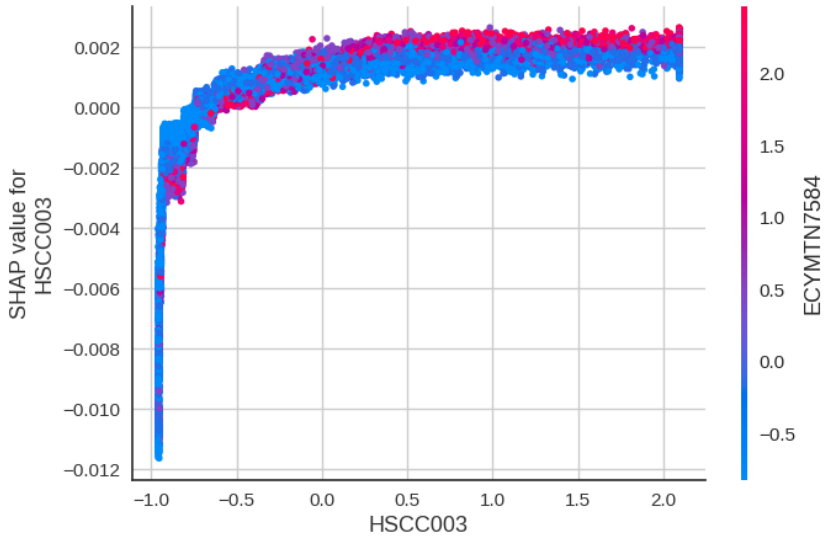

Figure 14. SHAP value for the variable HSCC003

Now, in our linear Elastic Net model, the top 5 features were:

1) HSHO002 (-0.003281)
2) ECYMTN7584 (-0.002302)
3) ECYMTN4554 (+0.002155)
4) ECYMTN6574 (-0.001978)
5) ECYHNI200P (-0.001853)

Which means there was overlap of 4 of these features (namely HSHO002, ECYMTN7584, ECYMTN6574, ECYHNI200P, which were all part of the Shap features from the XGB). Since most of the top linear coefficients also appear in the XGB's most impactful SHAP features, this confirms they are indeed predictive. The only difference is that elastic Net only modeled linear relationships, while SHAP revealed non-linearities as we can see on the respective graphs.

From this we can safely conclude that the problem is non-linear, since the XGB model reached $R^2 = 0.735$, which was much higher than Elastic Net's $R^2 = 0.37$. We also deduce that SHAP dependence plots show non-linear patterns (saturating, curved, etc), and all these visualizations confirm that a non-linear model is indeed better suited for this problem.

total number of words = 2514