Analysis of Trends in Insurance Spendings of Canadian Households

Prepared for:
Professor Bravo Roman
TA. Amethyst He
TA. Ana Carolina Da Cruz
TA. Nathaniel Phleps
TA. Sahab Zandi
TA. Siming Tang

Prepared by:
251359512 - Thomson Lam
251287809 - Emily Berlinghoff
251303637 - Jasmine Vu
251172272 - Kevin Li
Word Count: 2735

April 11, 2025

Table of Contents

List of Figures and Tables

1.0 Executive Summary

We merged DemoStats and HouseholdSpend together, cleaned the data of missing and invalid values, imputed or dropped outliers, and created the target variable for supervised learning. We then used the Elbow method and silhouette scores to find the optimal number of clusters for K Means clustering, and visualized 5 distinct groups using UMAP and PCA based on our findings. We concluded that UMAP served as a better manifold method to interpret the clustering results due to its tendency to emphasize local over global structural relationships compared to PCA. We then trained an elastic net linear regression model using grid search to find the optimal lasso and ridge regularization strength, followed by a XGBoost model with Random Search cross validation to find the best hyperparameters. We compared the performance of both models on the data, and found that high variance, low-bias models like XGBoost generalized much better to the data than linear regression likely due to the non-linear and complex nature of the data. Lastly, we computed and plotted the SHAPley values for XGBoost on a bee swarm SHAP plot, showing that for our model, housekeepers, total money gifts/contributions and support payments, regular mortgage payments, net purchase price of owned secondary residences and water and sewage charges for owned principal residences had the most influence on predictions of the target variable.

2.0 Introduction

2.1 Purpose of the Report

This report outlines the analysis of Canadian households, and prediction of the proportions of income spent on total personal insurance premiums and retirement contributions. The report includes clustering and regression, with elastic net linear regression and XGBoost models.

2.2 Background of the Report

Predicting insurance spending helps with financial planning, policy-making, and targeted marketing. This report uses 2 datasets provided by Environics, DemoStats and HouseholdSpend, for demographic statistics and household spendings to conduct a clustering analysis and produce model predictions. Both datasets contain hierarchical data, where variables of each category contain subcategories, sums, aggregates, averages, medians, etc. DemoStats was collected using a top-down, bottom-up modelling approach feedback loop to refine its projections, including future estimates and current time projections.

2.3 Scope of the Report

This report focuses on analyzing insurance spending among Canadian households using both supervised learning and clustering. The report aims to offer insight into household financial behavior and demonstrate the utility of data-driven methods for predictive modeling in this domain.

3.0 Process and Findings

3.1. Data Preprocessing

We manually filtered for specific features that were of lower hierarchy from DemoStats and HouseholdSpend instead of aggregates and sums of data. Our rationale behind this was to reduce mixed signals, noise and collinearity within the data, enabling the model to pick up on more meaningful patterns. Given the top-down, bottom-up modelling approach used by Environics to ensure the validity of high level data, duplicate columns collected through different means were also dropped, leaving only informative subcategories. All data relevant to creating the target variable was removed for both supervised and clustering datasets to avoid introducing endogeneity, so the models carry no inherent bias and generalize well to unseen data. We merged the datasets and cleaned the columns with over 10% of invalid values and converted columns in

Strings to float 32. We chose to impute outliers using either mean or median based on the skewness of its distribution, given the efficiency of mean and median imputation.

3.2. Clustering and Dimensionality Reduction

3.2.1. Finding Optimal Clusters

This section presents the results of clustering analysis. We randomly sampled 10% of the preprocessed dataset to ensure representativeness and reduce memory usage. Given the high dimensionality of our dataset, we decided to remove noise and redundancy in the data. We dropped all columns with low variance, and filtered those with high correlation based on a variance threshold of 0.01 and a computed correlation matrix. We used the Elbow method and silhouette score to find the ideal number of clusters for K-Means clustering.

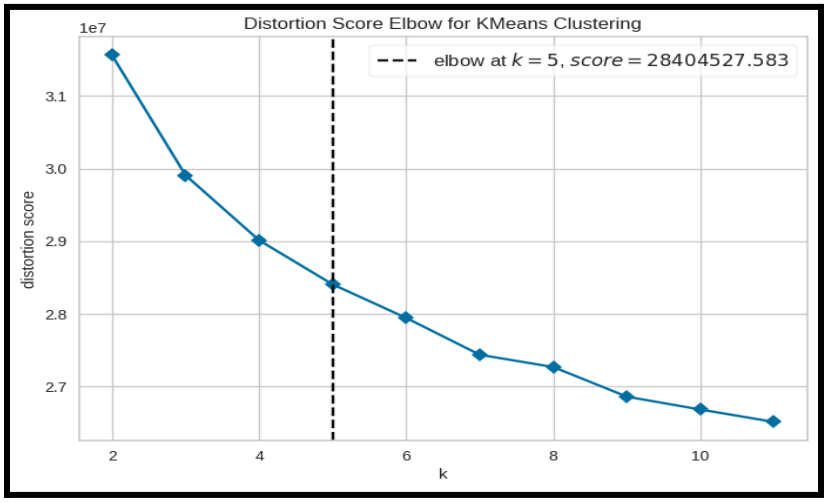

Figure 1.1 The ideal number of clusters shown by the elbow method.

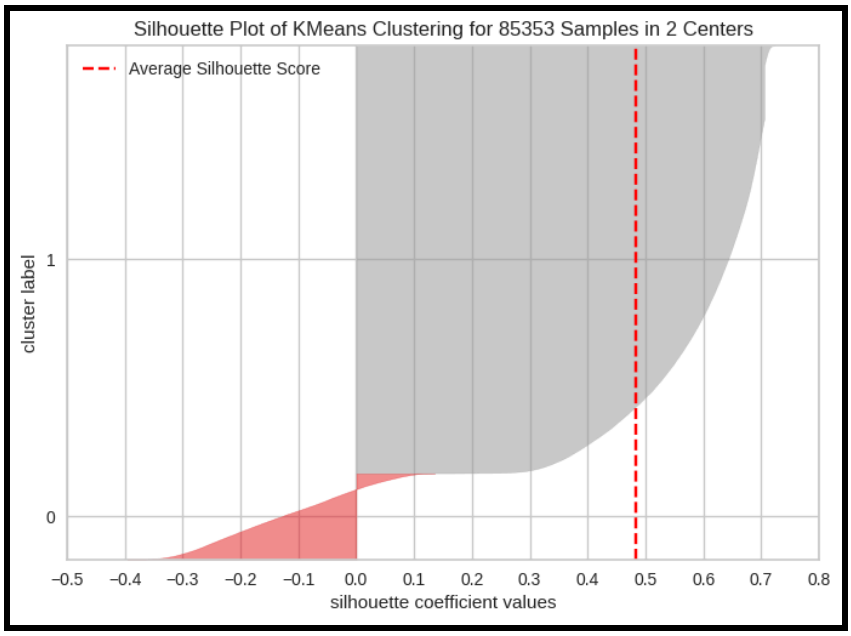

Figure 1.2 The ideal number of clusters shown by the silhouette method.

The Elbow method gives an ideal number of clusters $k = 5$, while the ideal number of clusters according to Silhouette indicates $k = 2$. The disagreement in the value of k is likely due to how the Elbow method favors more compact and localized patterns and often splits more defined substructures into separate clusters. Conversely, the silhouette score emphasizes more globally distinct clusters, and assigns higher scores when there are clearer boundaries between clusters. The Elbow method tends to find more compact regions and split it further into another cluster, likely due to how the Elbow method measures data points's closeness to their assigned centroid, aiming to minimize variance within each cluster. After plotting out each cluster number using PCA projection, we decided to use k=5, as it evenly segregates each cluster into more explainable partitions of the cluster.

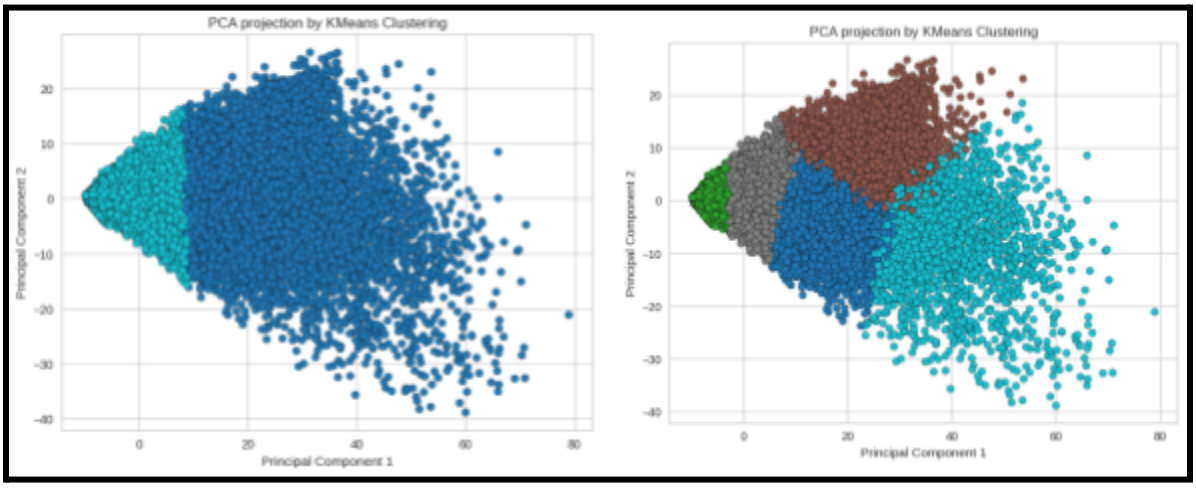

Figure 1.3 The cluster number of 2 and 5 plotted by PCA. The right plot shows the PCA projection for k=5, the left for k=2.

3.2.2. Analyzing Using Principal Component Analysis

For further analysis, we calculated the average value of each of the first 3 components for each cluster which was labeled with k-means clustering using the found ideal number of clusters of 5.

PC 1:
The features included in this PC describe a middle-aged, wealthy demographic that work in business finance administrations. We interpreted PC 1 to represent smaller sized households with high income and spending.

PC 2:
This PC consists of foreign populations and minorities. PC 2 represents immigrants and first generation residents.

PC 3:
This principal component mostly contains features that relate to renting and home investments. Thus, PC 3 represents low to middle income renters.

With the 3 components established, we now interpret each cluster using the table of average values:

Cluster #	PC 1	PC 2	PC 3
0	13.55	-3.52	0.03
1	-7.09	-0.06	0.09
2	22.95	10.12	-0.63
3	1.00	0.09	-0.28
4	40.12	-8.25	2.14

Table 1.1 The averages of the first 3 PCs for each cluster.

Cluster 0: Since this cluster has above average wealth, less immigrant presence, and small renter signal, this cluster consists of upper-middle-class and native residents that are homeowners.

Cluster 1: Since this cluster has below average wealth, little immigrant presence, and slight renter signal, then this cluster consists of lower-income native renters with modest housing.

Cluster 2: Since this cluster is very wealthy, has high immigrant presence, and negative renter value, then this cluster consists of rich immigrant households.

Cluster 3: Since this cluster has slightly above average wealth, neutral immigrant presence, and slightly less renter signal, then this cluster consists of middle-class mixed communities with balanced wealth, homeownership and diversity.

Cluster 4: Since this cluster has extreme wealth, very low immigrant presence, and strong renter/home investment theme, then this cluster consists of wealthy, native renters.

3.2.3. Analyzing Using UMAP

We used UMAP to reduce the data to 2 dimensions. We chose ‘n_neighbors=30’ to balance local and global structure and ‘metric=’cosine’’ to capture angular relationships between high dimensional features.

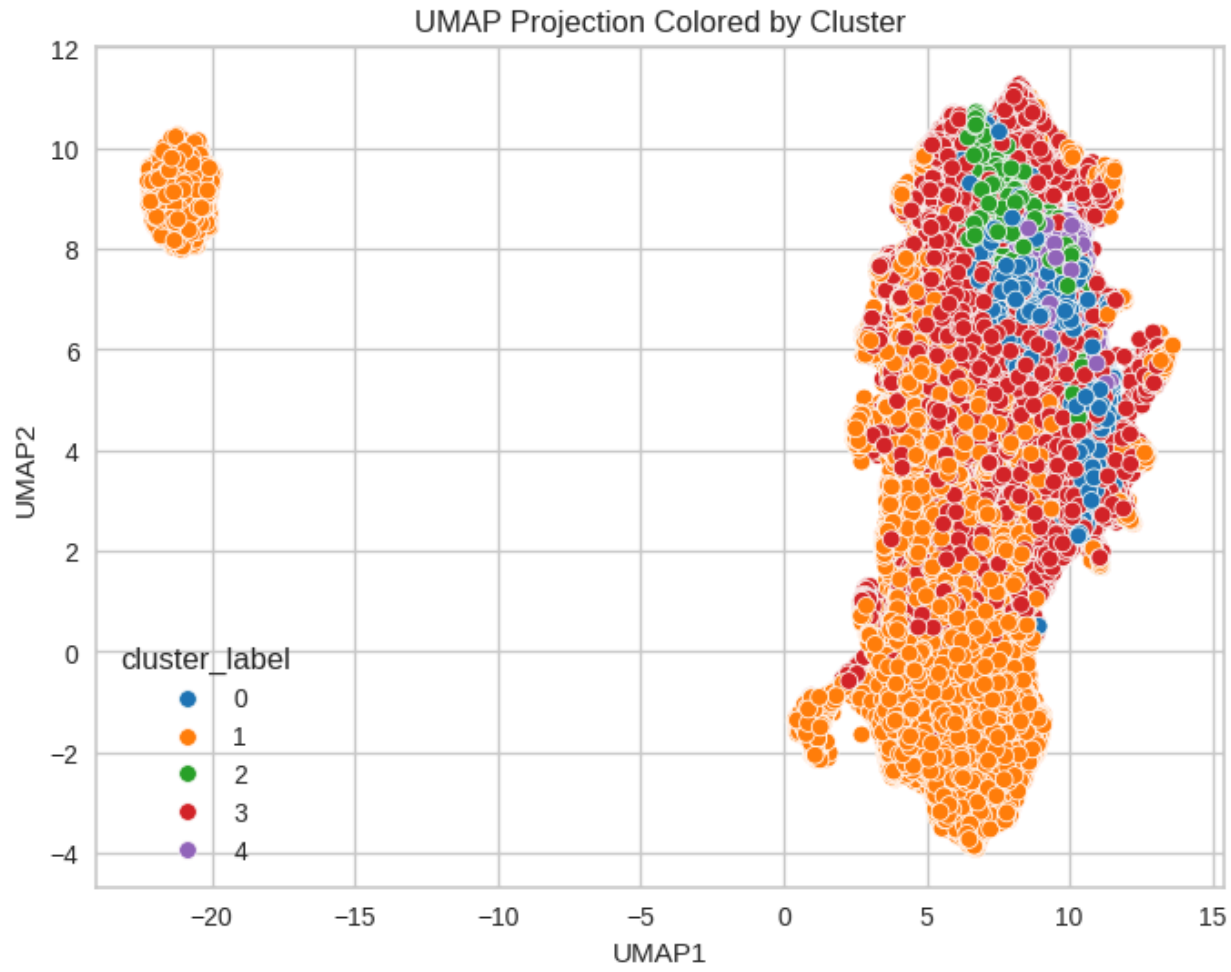

Figure 1.4 UMAP Manifold Projection with 5 cluster labels.

The UMAP projection mostly aligns with the clusters visualized by PCA, but the relationships between local data points are more emphasized. This is because the local structure is prioritized at the cost of distorting global data structure. The clusters are closely aligned and slightly overlap with each other compared to PCA. Although the PCA plot shows a more distinct and clear boundary between clusters, UMAP pulls points from separate clusters together in the low dimensional embedding compared to PCA, which projects data to capture the directions of maximum global variance at the cost of interpretability. Thus, UMAP compresses or distorts distances between clusters to improve representations of local structures. Although UMAP blurs the cluster boundary and causes overlaps, we find UMAP to be a better representation of the data, due to how information about local structure is preserved in the low dimensional embedding compared to PCA's, even though PCA's exaggerated separation of clusters makes it easy to visualize distinct clusters as a result of the uncorrelated principal components. UMAP can display more detailed relationships between data points that may not be projected in PCA, which we considered more vital to our analysis of the data. The manifold results also aligned with our expectations.

3.3. Training Regularized Elastic Net Linear Regression

3.3.1. Creation of the Target Variable (A-i)

We computed the target as the ratio of total spending on insurance and retirement to total household income. We then removed both features after creating the target variable, ensuring that any patterns learned are based on independent explanatory variables.

3.3.2. Necessary Data Transformations (A-ii)

To improve model performance and computational efficiency, we applied a StandardScaler within our Pipeline to center our features around 0 with the same order of variance.

3.3.3. Parameter Grid & Outcomes (A-iii)

We performed hyperparameter tuning using a grid search to explore combinations of regularization strengths for Lasso and Ridge penalties. This serves to ensure the optimal feature regularization is reached. The cross-validated model selected alpha = 0.01 and l1_ratio = 0.1, indicating a strong preference for Ridge regularization over Lasso. This suggests the model benefits more from minimizing variance and retaining most features rather than enforcing sparsity. On the test set, the tuned Elastic Net model achieved an R^2 of 0.2067 and a bootstrapped 95% confidence interval of [0.2032, 0.2098], indicating moderate predictive performance. While it outperformed a basic linear regression model, the improvement was modest, suggesting some degree of underfitting due to the simplicity and high bias of linear regression.

3.3.4. Elastic-Net Results (A-iv)

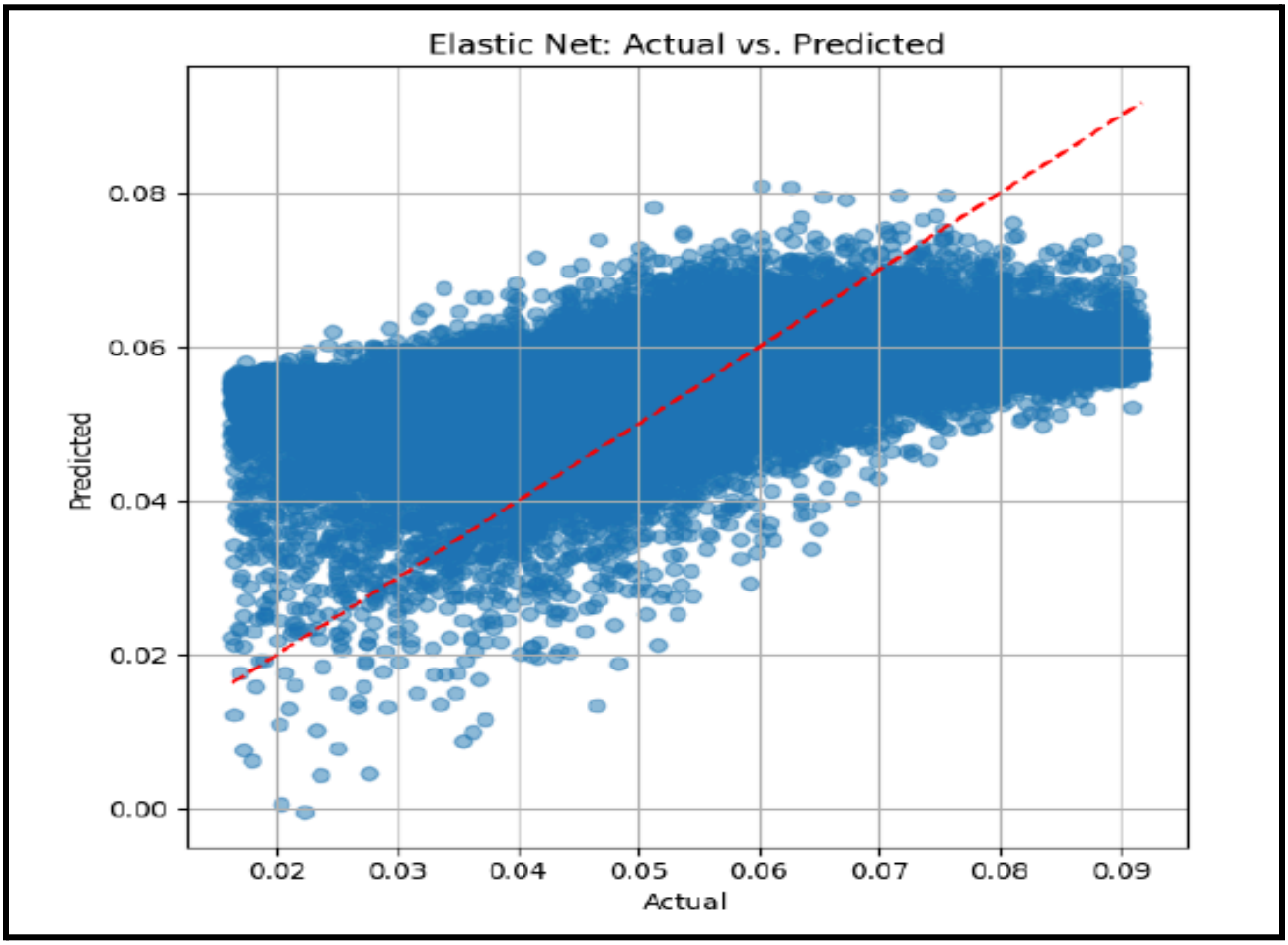

Figure 1.5 Plot of predicted and actual values from ElasticNet.

Model Performance:

- Test Set MSE: 0.0002
- Test Set R^2: 0.2067
- Bootstrapped 95% CI for R^2: [0.2033, 0.2099]

We quantified performance using mean squared error (MSE) as the loss function and the coefficient of determination (R^2 score) to measure the proportion of explainable variance by the model. We then calculated a bootstrapped 95% confidence interval for R^2 using 1,000 resampled test sets.

3.3.5. Top 5 Most Important Features + Interpretation (A-v)

	Code	Feature	Coefficient
1	HSMG001S	Total money gifts, contributions and support payments	-0.003588
2	HSHO002	Housekeepers, cleaners, house-sitters	-0.003434
3	HSSH011	Regular mortgage payments	0.002166
4	HSWH041S	Net purchase price of owned secondary residences	0.001868
5	HSSH018	Legal fees related to the dwelling	0.001375

Table 1.2 The 5 most important features and their meanings.

Since Elastic Net applies regularization, many coefficients are shrunk toward zero, allowing us to focus on the most meaningful predictors. We ranked features by relative importance in predicting the target variable using absolute values of their coefficients. The top five most influential features included HSMG001S, HSHO002, HSSH011, HSWH041S, and HSSH018. Positive and negative coefficients indicate increases in those variables are associated with a higher and lower proportion of income spent on insurance and retirement respectively.

3.4. Training XGBoost

3.4.1 XGB Model Summary (B-i)

Due to computational constraints, we applied Random SearchCV to tune hyperparameters. We evaluated 50 random combinations from the grid. After identifying the best estimator, we assessed its performance on the test set using MSE and R^2, and calculated a 95% bootstrapped confidence interval to quantify the model's stability.

3.4.2. Parameter Grid & Rationale (B-ii)

To optimize the performance of the XGBoost model, we constructed a well-rounded hyperparameter grid to balance model flexibility with computational efficiency. The range of additive learners allows exploration of shallower and deeper ensembles, while shallow tree depth values between 1 and 3 reduce overfitting and encourage simple trees. The `learning_rate` values span from 0.01 to 0.3 to balance fine-grained updates and faster convergence. Subsampling rates for rows and features from 0.2 to 1.0 are included to introduce stochasticity, which prevents

overfitting. The minimum loss reduction for node splitting is varied to regulate tree growth more conservatively. Finally, a wide range of values for ‘reg_lambda’ and ‘reg_alpha’ is provided to test how much regularization is needed to penalize overly complex models.

3.4.2. XGB Results and CI Comparison against Elastic Net (B-iii)

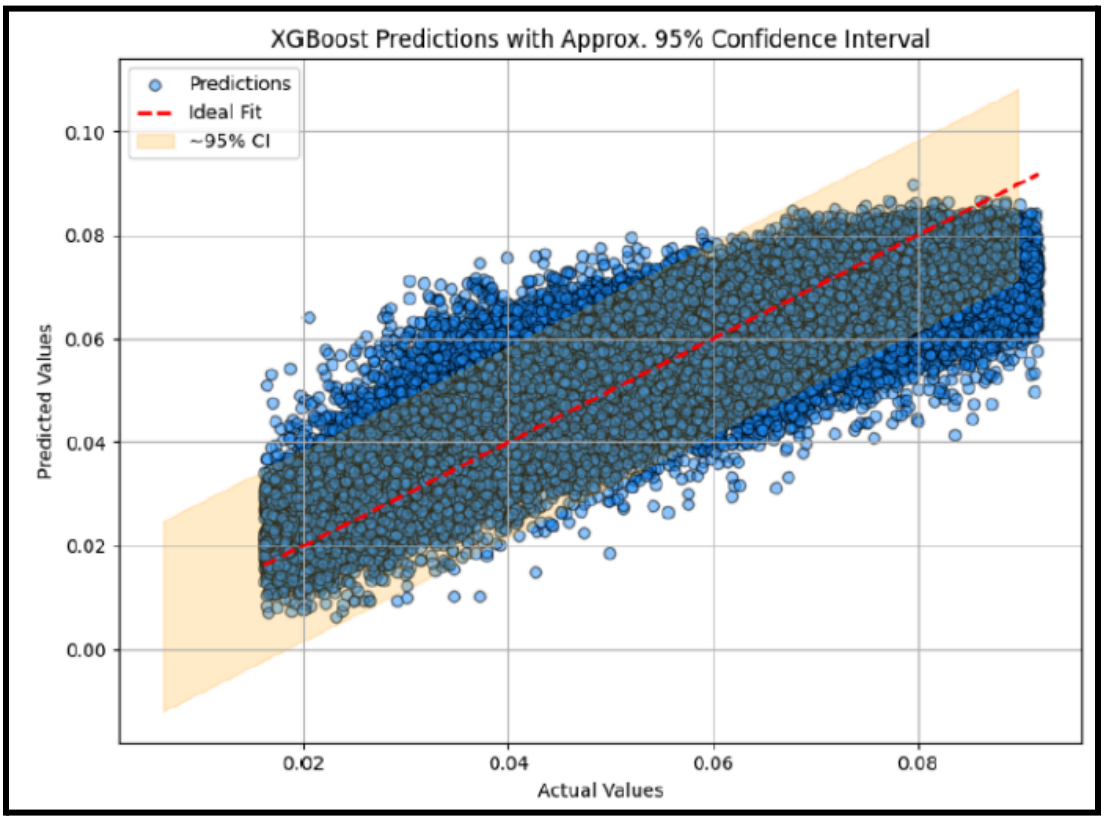

Figure 1.6 Plot of predicted and actual value from XGBoost.

Test Set MSE: 0.0001
Test Set R^2: 0.6651
Bootstrapped 95% CI for R^2: [0.6609, 0.6695]

The resulting scatterplot shows a strong, tight alignment of predicted values along the ideal fit line, with minimal dispersion and consistent variance. This performance is reflected quantitatively in the test set R^2 score of 0.6651 and a bootstrapped 95% confidence interval of [0.6609, 0.6695], indicating that the model captures a substantial portion of the variance in the data with reliable generalization. Compared to the Elastic Net model, XGBoost demonstrates superior predictive power. The elastic net linear regression struggled to capture the complex, non-linear relationships present in the dataset, causing under-fitted predictions. In contrast, XGBoost’s ability to model high variance data and nonlinearities results in much more accurate and stable predictions, making it the more robust choice for this regression task.

3.4.3. Variable SHAP Results & Interpretation (C)

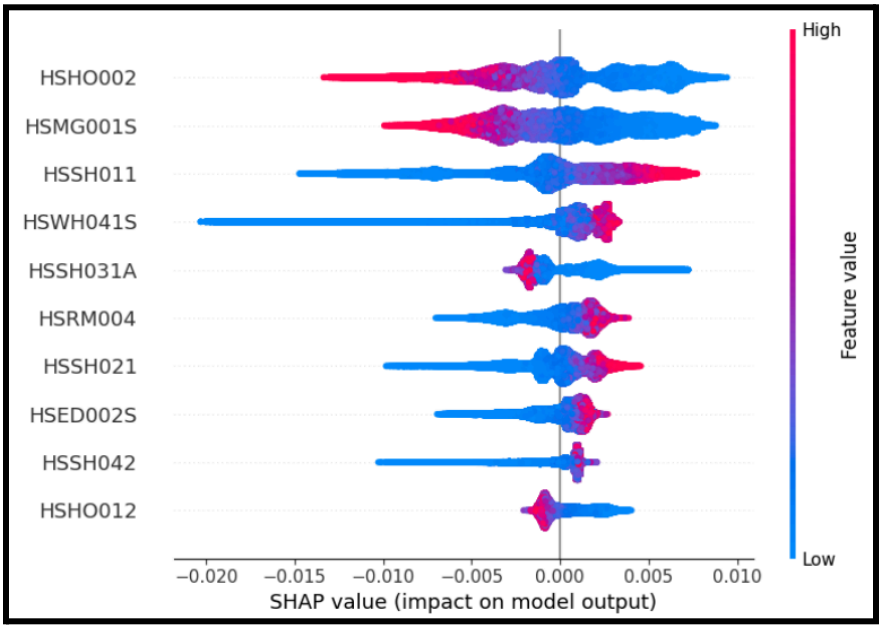

Figure 1.7 Plot of SHAP values for features.

We plotted a bee swarm SHAP plot to quantify the contribution of each feature to the model's output. The most influential variables include:

HSHO002: Housekeepers, cleaners, house-sitters
HSMG001S: Total money gifts, contributions and support payments
HSSH011: Regular mortgage payments
HSWH041S: Net purchase price of owned secondary residences
HSSH031A: Water and sewage charges for owned principal residence

These features consistently have the greatest impact on the prediction of the target variable. Higher values of HSHO002 and HSMG001S in red generally push the model's predictions upward, indicating that households with higher values in these features tend to spend a greater proportion of income on insurance and retirement. In contrast, features like HSSH011 and HSWH041S show complex behavior with high and low values affecting the output in different directions. The SHAP values reveal a richer structure in the variables' influence on predictions. This supports our conclusion that the problem is inherently nonlinear and better suited to tree-based models like XGBoost, which can capture feature interactions and variable importance more flexibly.

4.0 Conclusion

This report conducted an extensive analysis based on demographic statistics and household spending data for dissemination areas provided by Environics. After processing the data, clustering, principal component analysis and UMAP was conducted on the data without the target variable and variables used to create it, where we derived trends and demographics in our data based on the clusters formed. An elastic net linear regression and XGBoost tuned via RandomSearchCV was then trained on the data, with the XGBoost model having the superior performance on the test set. SHAP values for variables were then computed from the XGBoost model to interpret how each feature contributed to the predictions, where the results suggested that the model captured meaningful relationships from the training data.

References

Environics Analytics. (2024). *HouseholdSpend – Variables List* [HouseholdSpend 2024 - Variables List.xlsx]. Environics Analytics. https://environicsanalytics.com/docs/default-source/can---variable-lists/householdspend-variables-list.xlsx

Environics Analytics. (2024). *DemoStats – Variables List* [DemoStats 2024 - Variables List.csv]. Environics Analytics. https://environicsanalytics.com/docs/default-source/can---variable-lists/demostats---variables-list.csv