Calvin Cao, cca087

Joonas Dickinson, jdickin7

Kevin Liu, kliu469

Nicholas Heron, nheron2

DS3000

Predicting Retirement and Insurance Contributions by Household

Word Count: 2641

Introduction

In this project, we explore how households in Canada allocate their income, particularly the portion spent on personal insurance premiums and retirement contributions. We will use two datasets provided by Environics Analytics, DemoStats and HouseholdSpend, which provide extensive demographic and economic information aggregated by neighborhood.

We start with data cleaning and feature engineering, converting raw neighborhood-level data into a format suitable for modeling. From there, we identify patterns across neighbourhoods through k-means clustering. We then build and evaluate both regularized linear models (Elastic Net) and nonlinear models (XGBoost) to predict spending. Finally, we interpret our results using SHAP values to better understand which demographic and economic features most strongly influence retirement and insurance spending.

Data Cleaning and Preparation

Before any modeling could begin, our data needed significant transformation. The raw data included aggregate household and demographic statistics by postal code. However, many of these figures - such as total spending - were absolute values dependent on neighborhood size. To make our data meaningful for per-household prediction, we normalized these values by dividing them by the number of households per neighborhood. This step ensured that we weren't simply capturing differences in population size.

Our target variable, the proportion of household income spent on retirement and insurance, was then constructed by dividing insurance-related spending by aggregate household income. We also conducted feature selection to reduce noise and

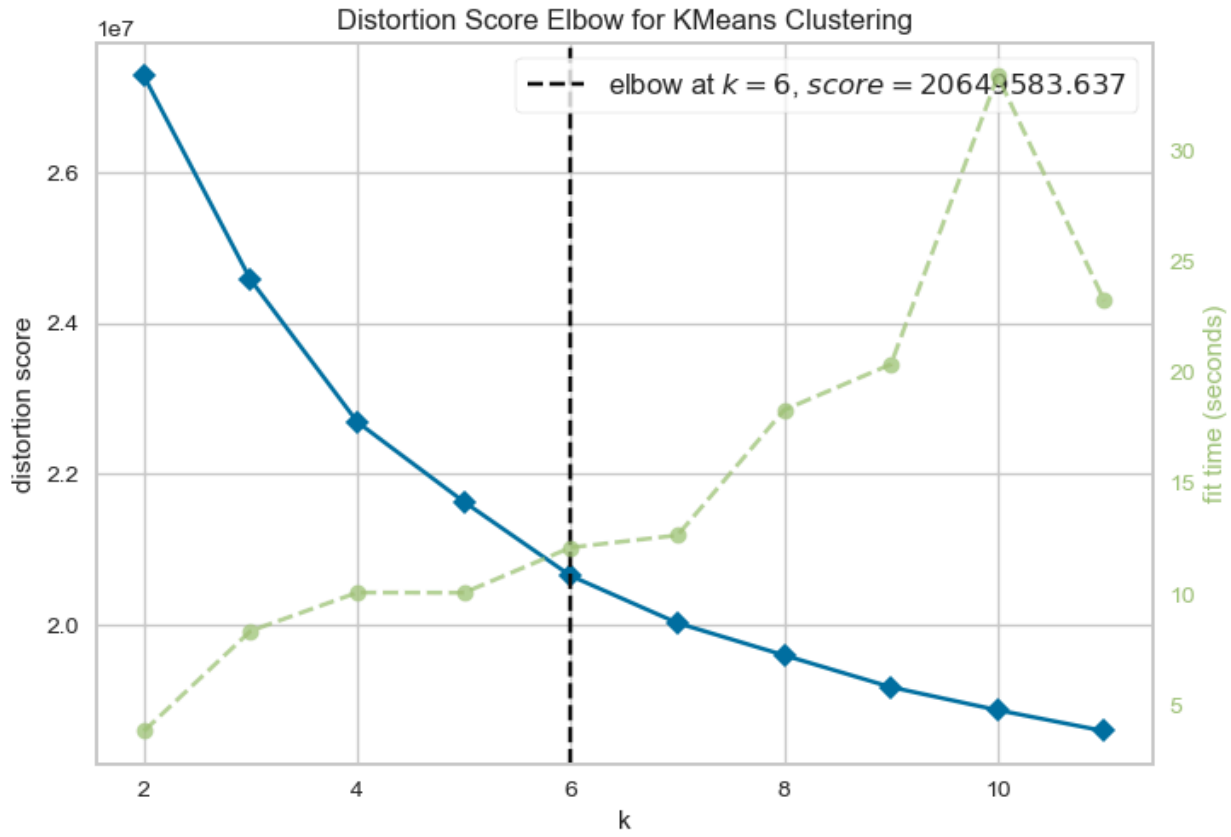
multicollinearity. This included dropping any features directly used to compute the target (like income and insurance variables) as well as features highly correlated with it. Beyond that, we used the datasets' metadata to systematically eliminate "parent" variables in hierarchical categories, keeping only the most granular and informative ones. For example, if both "Total Population" and "Total Male/Total Female Population" were present, we kept the latter.

Further, we removed features with over 5% negative or missing values, as well as rows where population was zero or where most entries were invalid. These likely represented areas without sufficient data coverage. Some columns, especially in the spending dataset, contained extreme outliers and invalid negative values. We handled this in two steps: first by imputing negatives with the median of valid values, and then by clipping outliers using a 3×IQR rule. Finally, the data was standardized using z-scores to bring all features onto the same scale, which is known to improve the performance of most of the models we will train.

K-Means Clustering

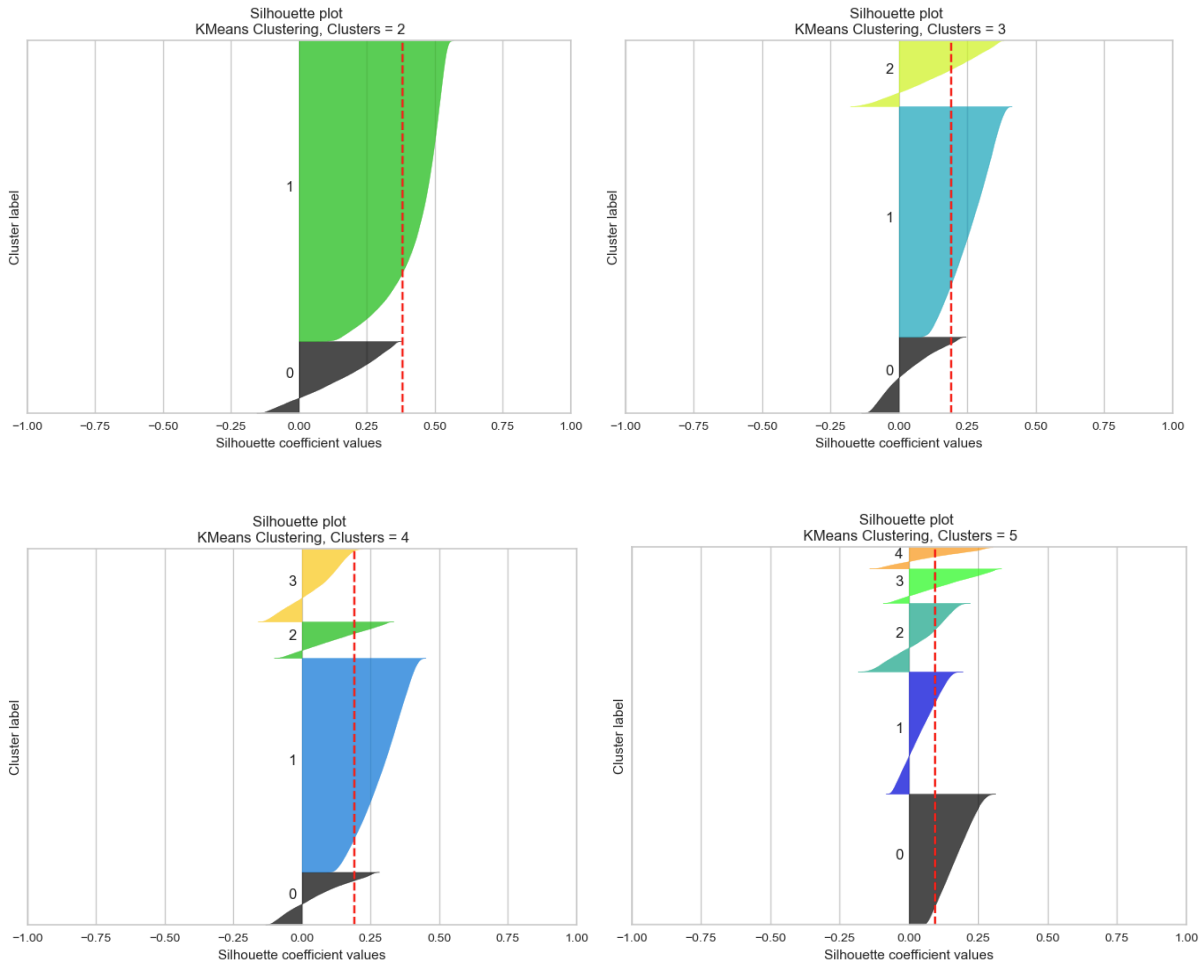
Before diving into supervised prediction, we explored the structure of the data using k-means clustering, excluding the target variable. Given the high dimensionality and large quantity of samples, we first applied preprocessing to a 10% sample of the dataset to keep training time reasonable. This sampled subset was processed with the same pipeline used earlier: negative imputation, outlier clipping, and standard scaling.
To determine the ideal number of clusters, we used the elbow method and silhouette analysis.

The distortion score test found the elbow at k=6.



For the silhouette analysis, values k = 2 to k = 12 were explored. Overall, the clusters were of low quality. As k increased, the silhouette chart showed many points had low or even negative silhouette scores, suggesting overlapping clusters and unclear group boundaries. This is likely due to the curse of dimensionality - as the number of features grows, distance metrics like Euclidean distance (used by k-means) become less meaningful. Additionally, k-means assumes spherical, equally sized clusters, which may not hold for this real-world socioeconomic dataset. Preprocessing steps - such as applying PCA or switching to alternative clustering techniques - might have yielded better clusters.

The plots suggest that the least bad clustering was k = 2; it is the only plot for which there isn't a significant amount of points with negative silhouette scores and the clusters are at least approaching balance in terms of number of points per cluster above or below the overall average silhouette.



Dimensionality Reduction

To explore the structure of the dataset, we used dimensionality reduction techniques.

PCA

We reduced the dataset to five components using PCA and analyzed the 3 that explained the most variance.

Component 1, which we labeled "Demographics", captured the majority of the variance in the target variable. It's driven by variables like total household population, employment levels, household income, and education. Essentially, this component reflects the overall demographic and socioeconomic profile of the household.

Component 2, labeled "Spending", included variables related to day-to-day expenses such as food, transportation, shelter, and household operations. This component summarizes general spending behavior across categories.

Component 3, which we called "Elderly", was characterized by variables tied to senior population segments, like the number of household maintainers over 75, people over 65 living alone, and age-specific demographic indicators. It seems to capture the presence and lifestyle of older adults in a household.

Component 1 "Demographics" Top 10 Variables by Explained Variance

Variable	Variance
Total Household Population	0.0703
With Income	0.0701
Non-Indigenous Identity	0.0700
Total Census Families	0.0696
Total Population	0.0696
One-Family Households Without Additional Persons	0.0687
Employed	0.0681
Total Households	0.0675
High School Certificate Or Equivalent	0.0664
Aggregate Household Income (Current Year $)	0.0661

Component 2 “Spending” Top 10 Variables by Explained Variance

Variable	Variance
Household operation	0.1104
Food	0.1098
Household furnishings and equipment	0.1096
Recreation	0.1089
Transportation	0.1087
Premiums for homeowners' insurance	0.1060
Shelter	0.1052
Restaurant dinners	0.1043
Income tax	0.1031
Personal care	0.1024

Component 3 “Elderly” Top 10 Variables by Explained Variance

Variable	Variance
Maintainers 75 To 84	0.1141
People 65 Years Or Over Living Alone	0.0997
French	0.0996
Third Generation Or More	0.0989
Wood and other fuel for heating and cooking for owned principal residence	0.0941
Females 70 To 74	0.0916
Household Population 80 To 84	0.0900
French Only	0.0855
Landline telephone services	0.0842
Maintainers 85 Or Older	0.0809

When plotting the PCA-reduced data, we see that the data is quite spread out, which means Component 1 and 2 each capture a fair amount of variance. When coloring points by kmeans cluster assignment (k=2), we saw a very clear separation almost entirely along Component 1. The clusters were split cleanly, suggesting that the main driver of variation between clusters is the demographic profile of the household rather than spending. This interpretation is supported by the fact that the average values for

Component 1 in each cluster differed significantly, while Components 2 and 3 showed almost no distinction.

Average Value of Components Per k-Means Cluster

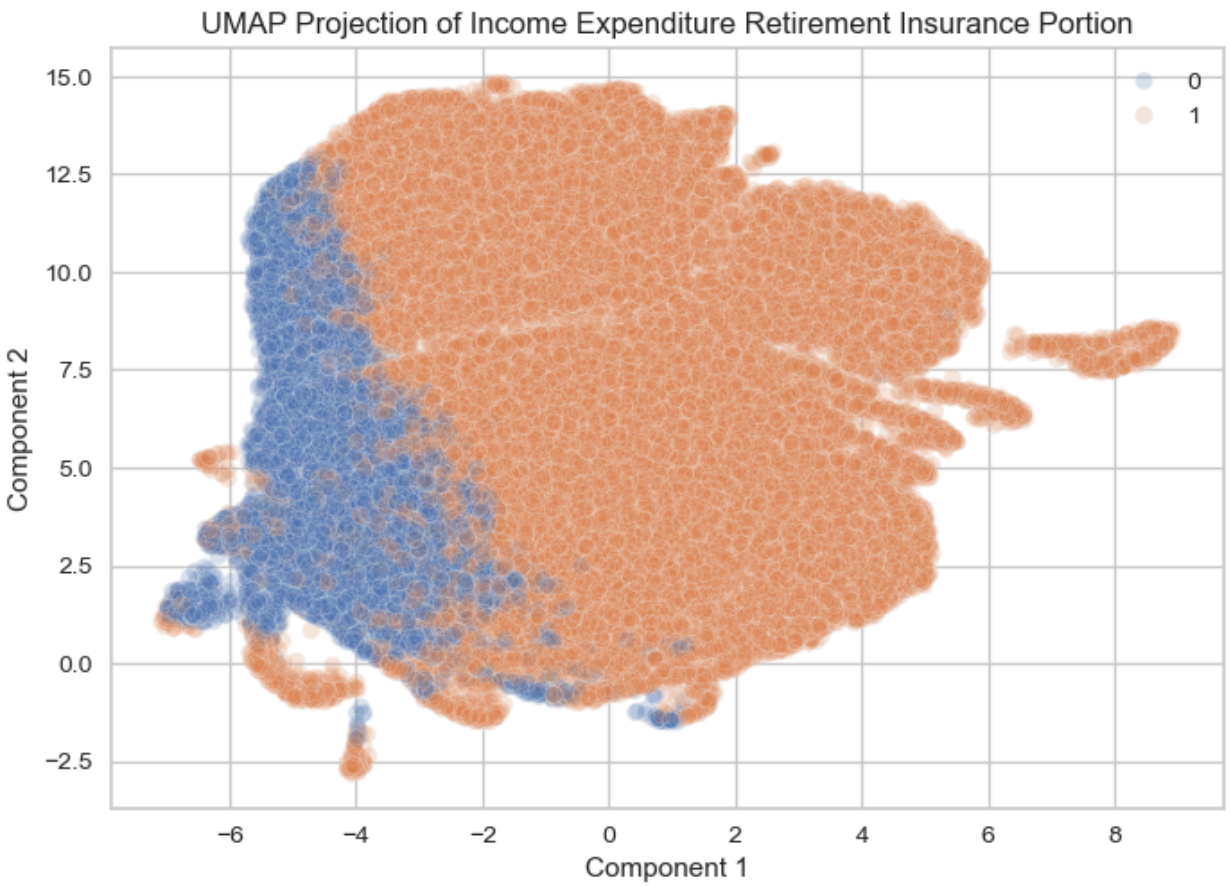
	Component 1	Component 2	Component 3
Clust 1	24.712773	-0.186233	-0.318945
Clust 2	-5.900950	0.044469	0.076158

This dominant role of demographics raises some questions about how meaningful the clusters are. Rather than representing two fundamentally different household types across multiple dimensions, the clustering appears to reflect a simple threshold set at an arbitrary point on the demographic spectrum. There's little evidence of interaction between spending behavior and demographic profile in how the clusters are formed.

UMAP

To test whether nonlinear structure might reveal more nuanced patterns, we also applied UMAP. We used a grid search to explore combinations of n_neighbors (5 to 200) and min_dist (0.1 to 10), and ultimately settled on 100 and 0.5, respectively. Lower values for these two variables tended to produce scattered and arbitrary groupings that didn't align well with the cluster structure observed using PCA and k-means.

The UMAP projection ended up looking very similar to the PCA plot. It produced a nearly identical clustering split, though the separation was slightly more diagonal. This suggests that UMAP gives a bit more weight to Component 2 than PCA does, possibly picking up subtle patterns in spending behavior that are otherwise overshadowed by the strong demographic signal, though both clusters ultimately agree that Component 1 is the most significant.



Elastic Net Linear Regression Model

After data cleaning, we trained a linear regression model using an elastic net penalty. To fine-tune the model, we performed a grid search over two key parameters: the alpha value, which controls the overall strength of the regularization, and the L1 ratio, which determines the balance between L1 (Lasso) and L2 (Ridge) penalties.

For the alpha parameter, we tested a range of values: 0, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1, and 1. The grid search revealed that the model performed best with the smallest non-zero value tested, 0.000001. This suggests that our features are already quite informative, and the model did not benefit significantly from strong regularization.

The L1 ratio was tested using the values 0, 0.1, 0.25, 0.5, 0.75, 0.9, and 1. According to the grid search, the model achieved the best performance when the L1 ratio was set to 1, indicating that a pure Lasso regression approach outperformed both Ridge regression and combinations of the two. While Lasso regression is utilized for its ability to remove irrelevant features by shrinking their coefficients to zero, the very low alpha value used in this case likely meant that no features were eliminated.

Once the best parameters were identified through the grid search, we trained the final model on the training dataset and used it to predict the values on the test set. The model achieved a mean squared error (MSE) of approximately 5.991×10^{-5}, with a 95% confidence interval ranging from 5.921×10^{-5} to 6.063×10^{-5}. This low MSE indicates that the model's predictions for household spending on insurance and retirement were close to the actual values. Additionally, the model produced an R-squared value of 0.777, with a 95% confidence interval between 0.7742 and 0.7796. Since an R-squared value above 0.7 is typically considered strong our model performed quite well. This result suggests that the model can explain roughly 77% of the variance in household spending on insurance and retirement.

The top five variables contributing to the model, along with their respective coefficients, provide some insight into the patterns observed. The variable HSTX001: Income tax had a coefficient of -0.009178281, indicating that higher income tax payments are associated with lower spending on insurance and retirement. HSHC003: Prescribed medicines and pharmaceutical products had a coefficient of -0.015387753, suggesting that households with higher prescription expenses may have poorer insurance coverage or more health-related financial burdens, leaving less money available for retirement savings. In contrast, HSHC012: Prescription eyewear had a positive coefficient of 0.011309660, which could indicate that households able to afford eyewear may have better insurance coverage overall. HSME001S: Miscellaneous expenditures had a positive coefficient of 0.007342417, supporting that households with

more disposable income can allocate more toward insurance and retirement. HSTR001S: Transportation had a positive coefficient of 0.006880272, possibly reflecting the idea that households with higher transportation costs, such as those owning expensive vehicles, tend to have higher incomes and are thus able to invest more in retirement and insurance.

XGB Model

After implementing the Elastic Net model, we trained an XGB model to try and capture any nonlinear relationships in the data that the linear regression model may have missed. Considering both accuracy and training time, we used a parameter grid to test different combinations of the three most impactful hyperparameters.

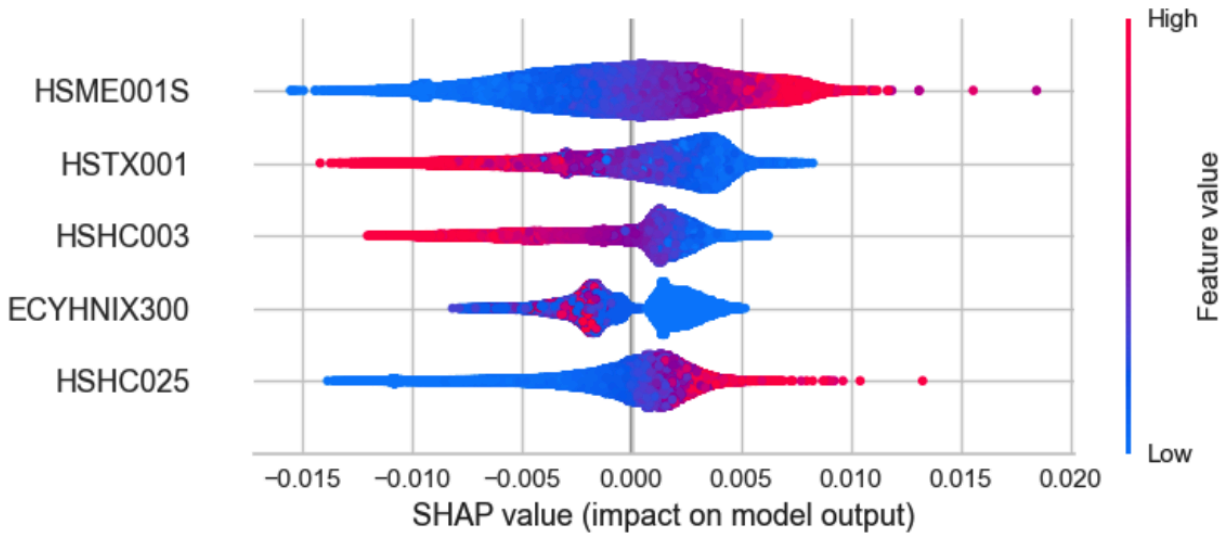
For the parameter grid, we used 5-fold cross-validation to test three parameters: max_depth, learning_rate, and n_estimators. These choices affect how deep each tree can go, how fast the model learns, and the total number of trees to build. Specifically, we tested these values for max_depth: 3, 5, 7, 9, or 11. Deeper trees allowed the model to learn complex patterns at the risk of overfitting. We also tested five values from 0.001 to 10 for the model's learning rate. This helped us find the perfect balance between underfitting and overfitting. Lastly, we tested six values from 50 to 2000 for the number of estimators to build. More trees can improve accuracy but take longer to train.

The optimal values for the parameters after the grid search were the following: a moderate learning rate of 0.1, maximum depth of 7 for each tree, and a total of 1500 trees. This combination allowed the model to learn gradually while still capturing enough complexity without overfitting to noise in the training data.

Using this configuration, we trained the XGB model and calculated the confidence intervals for both MSE and R^2, which showed better performance compared to the Elastic Net. For the MSE, the CI was between 1.5×10^{-5} and 1.6×10^{-5}, showing that its predictions were much closer to the real values than Elastic Net's predictions.

The confidence interval for R^2 of the XGB model was between 0.9413 to 0.9430, both higher and narrower than the Elastic Net's interval. Specifically, the confidence interval range decreased from 0.0054 to 0.0017. This means that the XGB model not only explained more of the variation in the household spending differences on retirement and insurance, but also did so with more consistency across different bootstrap samples.

We also created a SHAP summary dot plot to show the five most important variables that helped our XGB model predict the percentage of income spent on retirement and insurance.



The first variable in the SHAP plot above is miscellaneous expenditures (HSME001s). High values appear on the right of the plot, which means households that spend more on miscellaneous goods and services tend to have more financial flexibility to spend more on services like retirement and insurance.

The second variable is income tax (HSTX001). High values appear on the left of the plot, which means households that spend more on paying income tax tend to contribute less to retirement and insurance.

The third variable is prescribed medicines and pharmaceutical products (HSHC003), with a similar plot shape as income tax. This means that households that need to budget for medication may have less money left to contribute to long-term retirement savings and insurance.

The fourth variable is household income ranging between $200,000.00 to $299,999.00 CAD (ECYHNI300), which represents upper-middle class households. This household demographic mostly consists of dual income professionals, so they likely have existing retirement plans sponsored by their employer.

Lastly, the fifth variable is disability expenses (HSHC025). The high values are on the right, which means households that spend more on disability support also spend more on insurance, which makes sense.

Compared to the results from the Elastic Net model's coefficient analysis, the XGB model's results match for three out of five variables. Both models chose miscellaneous spending, income tax, and prescription medicine costs as key variables. However, the Elastic Net model chose prescription eyewear and transportation that the XGB model did not consider as important. Similarly, the XGB model chose upper-middle class households and disability expenses that Elastic Net overlooked.

This difference in variables comes from how the two models handle relationships between variables and the target. Elastic Net assumes each variable has a constant, fixed effect across all households, so even a small but steady effect can signal that it is an important variable. In contrast, XGBoost doesn't assume the same fixed effect everywhere. So in XGBoost, if a variable like transportation spending only helps explain outcomes for a small group, such as households with lower income, it may only be used in a few branches of the model. In contrast, Elastic Net would still give it a coefficient based on its overall average effect across the entire household spending dataset, even if the impact applies to a small group.

After analyzing how different household spending categories relate to retirement and insurance contributions, we believe the relationship is nonlinear. One example of how the variables are nonlinear can be seen by how middle class households tend to contribute more to retirement and insurance than lower-middle class households, but that doesn't mean contributions keep rising as household income rises. It could be due to reasons such as upper-middle class households having trust in their existing

employer-sponsored retirement plans. Therefore, a flexible model like the XGBoost is well suited at capturing the real world factors that affect household financial decisions.

Conclusion

In conclusion, the linear regression model was able to effectively predict the income spent on insurance and retirement. This is supported by the low mean squared error, indicating that the model's predictions were close to the actual values, and the high R-squared value, which shows that a large portion of the variance in household spending on insurance and retirement can be explained by the model. However, due to the nonlinear relationship between the variables and their effect on a household's retirement and insurance spending, the XGB model was able to achieve even better R-Squared and mean squared error values, and was overall the best model for predicting proportion of income spent on insurance and retirement.