AN ANALYSIS OF HOUSEHOLD SPENDING AND RELATIONS TO PROPORTIONS OF INCOME SPENT ON INSURANCE PREMIUMS AND RETIREMENT/PENSION CONTRIBUTIONS IN CANADA

Data Science 3000B – Introduction to Machine Learning

UNDERGRADUATE GROUP 22

OWEN STEVENSON - 251179331

AIDAN DIGNAM - 251202245

AMELIA WALKER - 251143501

ANDY YUAN - 251159206

WORD COUNT: 2473

Executive Summary

Understanding Canadian household spending behaviours is critical for effective policy-making and economic forecasting. Using data from Environics Analytics, this report applies supervised and unsupervised learning techniques to model income allocated to insurance products and pension/retirement contributions.

To uncover insights from the financial behaviour of Canadian households, the data was cleaned and processed through clustering, dimensionality reduction, and regression modeling. Through these processes, two key consumer groups emerged, an urban and a rural group, based on demographic factors and spending characteristics.

Key insights from the data analysis are derived under:

- **Dimensionality Reduction & Clustering:** Principal Component Analysis and Uniform Manifold Approximation and Projection are used to reduce data complexity, illustrating that urban households have higher education, spending, and population density in comparison to the rural group.
- **Regression Analysis:** Elastic Net and XGBoost are applied and highlighted the value of original features in predicting financial behaviours.

Through the various machine learning techniques applied to the data, key variables were uncovered that were shown to be driving forces in spending on insurance products and contributory payments made to pension/retirement accounts. The key variables include:

- Median and average mortgage-to-income ratios.
- Housing and accommodation costs.
- Public sector employment proportion.
- French home language presence and education levels.

Policy and strategic recommendations arose based on the interpretability of the model output. It is evident that policy interventions should focus on urban and high-cost areas where spending behaviours suggest that the income constraints have an impact of savings capacity.

By combining data science techniques with socio-economic insight, this analysis supports more informed, data-driven policymaking and consumer segmentation strategies across Canada. The findings highlight the importance of flexible modeling approaches and robust clustering for capturing real-world financial behaviour patterns and its impact regarding household expenditure on insurance products and pension contributions.

Table of Contents

Clustering & Dimensionality Reduction

Understanding spending and saving behavior of Canadian households is important for developing policy design and economic analysis in Canada. This report explores national-level household data from Environics Analytics, combining spending and demographic metrics for analysis. The objective of this report is to cluster Canadian households into socioeconomic groups using unsupervised machine learning, and to model the proportion of income allocated toward personal insurance premiums and retirement or pension contributions using supervised learning. By combining analysis from dimensionality reduction, clustering, linear regression, and tree-based regression, the results will attempt to explain financial behavior across Canadian communities. Our findings are designed to support more data-driven decisions in the context of income support, retirement readiness, and consumer segmentation.

Data Cleaning

In the cleaning phase, household spending and demographic datasets were merged using the geographic identifier "CODE." With the combined data, redundant columns such as "CODE_right", "GEO", and "GEO_right" were removed, as they contained no additional analytical value. To prevent information leakage in unsupervised modeling, all response-related variables such as pension contributions (HSEP001S) and summary income indicators (HSHNIAGG) were excluded. Missing values, primarily NA's, were imputed using column-wise medians calculated from the training set. String columns were set to numeric types, and all continuous features were scaled using StandardScaler to ensure equal equally to distance calculations. The result is a cleaned and numeric dataset that can be used for clustering and dimensionality reduction.

K-Means Clustering

After cleaning and preparing the data through truncating missing values, standardizing variables, and dropping exogenous features, we can perform K-Means Clustering to identify similar characteristics within our data. First, we sample 400,000 individuals from our dataset to perform K-Means Clustering on. We use a representative large sample to ensure that the number of clusters we identify would be representative of the entire dataset. After applying clustering, we create a range of k clusters, from 2 to 11 to identify which cluster would give us the optimal within-cluster sum of squares (see Exhibit 1). This point occurs where the marginal benefit of an additional k cluster slows down. From this point, we identify 5 clusters as our optimal number of clusters. To compare this, our research conducted a silhouette analysis, which can provide better insights onto cohesion and separation between different clusters. The method measures within-cluster similarity, and the difference between clusters. We test 2 to 7 different cluster numbers- and find that with 2 clusters we obtain the highest silhouette score. This indicates, that under silhouette scoring methods, 2 clusters give the most well-defined clusters while along considering the degree of separation. Unfortunately, the elbow and silhouette method for identifying the optimal number of

clusters disagree. Likely, our data has overlapping or uneven clusters, and thus, the silhouette method will select a fewer number of clusters when compared to elbow. In our research, to reflect our desire for cluster quality, and for easy interpretability, we will continue our analysis with 2 clusters.

Linear Dimensionality Reductions

From the PCA cluster plot, we can see that PC1's scale is much larger than PC2's (see Exhibit 2 for visualization). This indicates that it can explain more variability in the data than does PC2. Unfortunately, the clusters for the data are not clearly defined from this plot. Overlapping points appear between cluster 1 in yellow, and cluster 0 in purple. This indicates that the data has no defined clusters, and that drawing clear lines between Canadian households can be quite challenging. For PC1, the average value is -0.5719 for cluster 0, and 483.3834 for cluster 1. For PC2, the average value is 0.0609 and -56.9878 for cluster 0 and 1. Lastly, from the third cluster, the values are -0.000039 for cluster 0 and 0.0364 for cluster 1. From the average component value, population statistics, education levels, income, and spending related features, cluster 1 tends to be higher than cluster 0. This increase in population size, spending, education, presence of different languages, among other demographic values can indicate that the two clusters represent a rural (cluster 0), and urban (cluster) affinity. Since these clusters are exact, there is still some overlap and close features between them.

Uniform Manifold Approximation and Projection

The Uniform Manifold Approximation and Projection (UMAP) uses the optimal parameters to project the high-dimensional data into two-dimensional space. The plot Exhibit 3 illustrates that most data, labeled as class 0, are compactly clustered in the center of the plot, with some sparsely distributed data points belonging to class 1. The silhouette score of 0.204 indicates a poor clustering structure and low inter-cluster separation, suggestion significant overlap between the classes. Further, the low silhouette score indicates that there is weak class distinction, aligning with the small number of points labeled as class 1, showing that there is a class imbalance within the data. The UMAP projection provides a compressed view of the data's structure, however, further dimensionality reduction techniques should be employed to enhance class distinction and improve clustering performance to gain a better understanding of the data.

Regression

Regularized Elastic Net Linear Regression

The elastic net model combines ridge and lasso regression by leveraging L1 and L2 penalties. This provides a balance between multicollinearity and selects relevant variables, providing stable and interpretable coefficients. To prepare the data, categorical variables were numerically encoded and

dimensionality reduction via Principal Component Analysis (PCA) was performed. Then, a model with and without PCA transformations were fitted.

Cross validation was used to tune the hyperparameters of the model. The optimal model, with PCA, returned an alpha of 0.0016 and an L1 ratio of 0.8. The optimal model, without PCA, had an alpha of 0.001 and an L1 ratio of 0.4. Overall, the hyperparameters optimize the performance of the model by balancing sparsity and coefficient shrinkage.

Regarding model evaluation, an R^2 of 0.0743 and a mean squared error (MSE) of 0.00002 for the PCA model and R^2 of 0.3221 and a mean squared error (MSE) of 0.0001 for the non- PCA model. This performance can be visualized through scatter plots of the predicted values against the actual values (see Exhibit 4 for the PCA-based prediction plot). Based on the plot of the regression line, the non-PCA model clearly demonstrates a stronger fit to the data (see Exhibit 5). We can infer that the non-PCA fit has string accuracy than the PCA model. Under bootstrapping, the PCA model had a 95% confidence interval for R^2 of (0.071, 0.077) and a non-PCA model 95% confidence interval for R^2 of (0.319, 0.326). We can infer that the PCA model has poor predictive power compared to the non-PCA model.

For the PCA-based model, the most significant features are average values of household income and mortgage-to-income ratios. The non-PCA model identified median mortgage-to-income ratio (ECYMTNMED), public sector employment proportion (ECYINDPUBL), average mortgage-to-income ratio (ECYMTNAVG), principal accommodation costs (HSCC014), and additional housing costs (HSTR052) as the most impactful predictors.

Overall, the non-PCA elastic net model performed stronger than the PCA elastic net model. Therefore, it is important to always analyze the original feature set before reducing dimensions.

XGBoost Model

XGBoost is a sequential learning algorithm that builds trees iteratively, correcting errors from each previous iteration. This method is useful for analyzing high-dimensional and non-linear datasets (see Exhibit 6). To prepare the data for modelling, categorical variables including demographic indicators were given numerical values. Then, PCA was applied to decrease the dimensionality of the data, where the first 10 principal components were preserved. Cross validation was used to tune the parameters of the XGBoost model, where many hyperparameters were explored to find the optimal level of model performance. The optimal model included 124 trees, with a maximum depth of 8 and learning rate of 0.038. This solution results in the best predictive accuracy, avoiding overfitting and ensuring efficiency.

When evaluating the optimal model, the MSE value was 0.0001 and the R^2 returned 0.7721. These results demonstrate strong performance, which can be observed in the scatter plots of actual vs. predicted value (see Exhibit 7). 100 bootstrap iterations were performed to obtain a 95% confidence interval the performs stronger than the elastic net model, with an R^2 confidence interval

of (0.319, 0.326). Moreover, SHAP values were used to understand variable importance and how specific predictors contribute to the predictions of the model. This process identified that the following predictors are significant to the model predictions: Median Mortgage-to-Income Ratio (ECYMTNMED), Average Mortgage-to-Income Ratio (ECYMTNAVG), Shelter Cost (HSSH033), French Home Language (ECYHOMFREN), and Education Level (College Degree, ECYEDACOLL). Thus, these variables have a strong impact of financial behaviours across Canadian households (Exhibit 8).

Compared to the elastic net model, XGBoost outperformed. This is caused by the model's ability to realize relationships across non-linear data. Both models identified similar attributes that are significant to the model, but XGBoost allows for better interpretability and accuracy in understanding these relationships.

Interpretations

The results from the Regularized Elastic Net Linear Regression and the XGBoost model exhibit a contrast in predictive performance based on the data at hand. The Elastic Net Regression without PCA yields an R^2 value of 0.3221 whereas the model with PCA has a significantly lower R^2 value of 0.0743. This indicates that variation within the model is better explained without the use of PCA, suggesting that essential information was removed during the dimensionality reduction process. Comparing the Elastic Net Regression with the XGBoost model, there is an observed stronger performance with the XGBoost method as it yields an R^2 value of 0.7721. The low MSE and narrow confidence intervals reinforce the reliability of the output from this model. Furthermore, the SHAP values produced by the XGBoost model allows for higher levels of interpretability, leading to the identification of key factors affecting financial behaviour such as average mortgage-to-income ratios, shelter costs, and education levels. While the Elastic Net Regression without PCA provides a reasonable amount of model interpretability, the XGBoost model outperforms in terms of explanatory power and accuracy. Overall, the XGBoost model should be used over the Elastic Net Regression when making decisions for policies regarding the spending behaviours of Canadians on insurance related products and pension contributions.

Conclusions

The XGBoost model can explain far more of the variation in the response variable compared to the standard linear model. This indicates that the relationship between the proportion of income spent on retirement, insurance or pension contributions is non-linear in nature. Thus, to better explain this relationship, XGBoost (non-linear model) is preferred. The performance results are summarized in the following table.

Model	R^2	MSE	95% CI (R^2)
Elastic Net (PCA)	0.0743	0.00002	(0.071, 0.077)
Elastic Net (Non-PCA)	0.3221	0.0001	(0.319, 0.326)
XGBoost (PCA)	0.4066	0.0001	(0.403, 0.410)
XGBoost (Non-PCA)	0.7721	0.0001	(0.770, 0.774)

From the SHAP values, when comparing elastic net to XGBoost, both models identify similar features as the most impactful (Exhibit 10). When using data without applying PCA, additional amounts paid to landlord (HSSH054) and principal accommodation cost (HSSH002) dominate the SHAP importance chart (effects shown in Exhibit 8). These factors could be related to income effects, where consumers spending more money on rent as a fixed value have higher income, are paying more into insurance schemes to protect a high income and lifestyle spending.

To conclude, the analysis identified two key clusters for Canadian consumers: a rural group and an urban group. From the provided data on Canadian household spending and characteristics, the models identified 10 features that explain about 70% of the variation in the data. Following this, a Linear Regression and an XGBoost model were used for predicting the proportion of income spent on insurance and pension contributions. The analysis found that the XGBoost model performed stronger, indicating a non-linear relationship between the proportion and the additional features. Using SHAP values, significant variables—including the Median Mortgage-to-Income Ratio (ECYMTNMED), Average Mortgage-to-Income Ratio (ECYMTNAVG), Shelter Cost (HSSH033), French Home Language (ECYHOMFREN), and Education Level (College Degree, ECYEDACOLL)—can be used to inform decision-makers about specific groups or characteristics of Canadians who may be contributing or saving too little for emergencies or retirement (Exhibit 11). This information can guide targeted policy decisions across Canada.

Word Count: 2473

Exhibits

Exhibit 1: K-Means Elbow Score

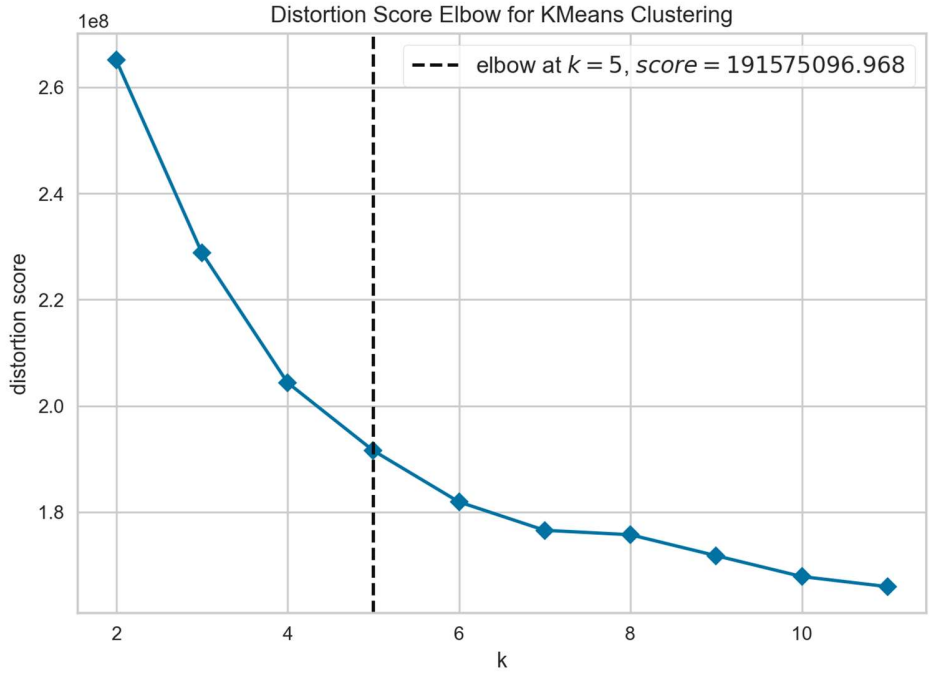

Exhibit 2: PCA Scatterplot by Colour & Cluster

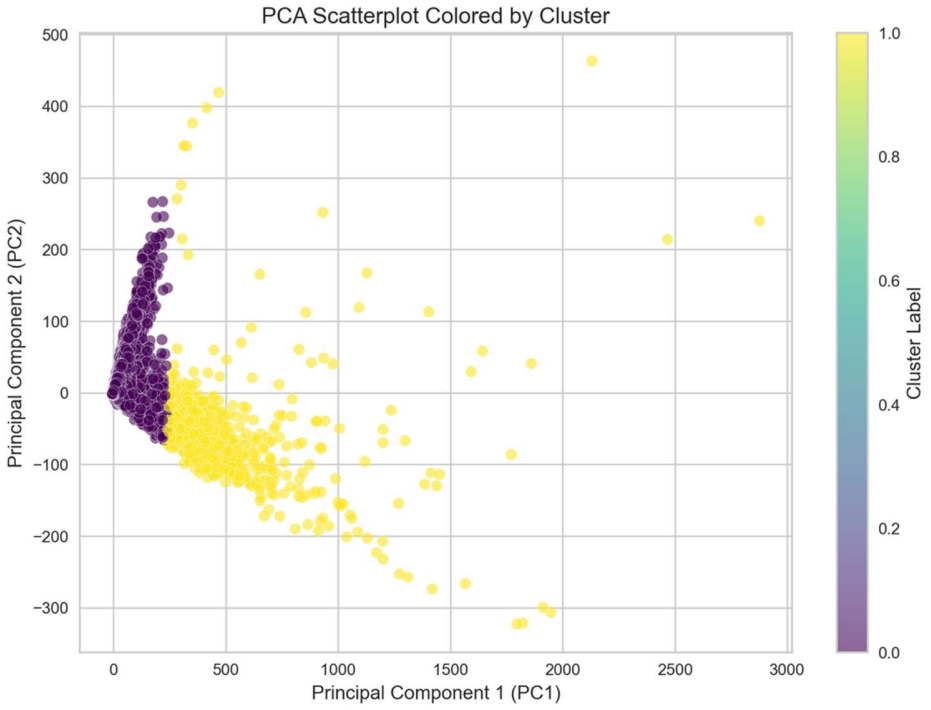

Exhibit 3: UMAP Plot of Data

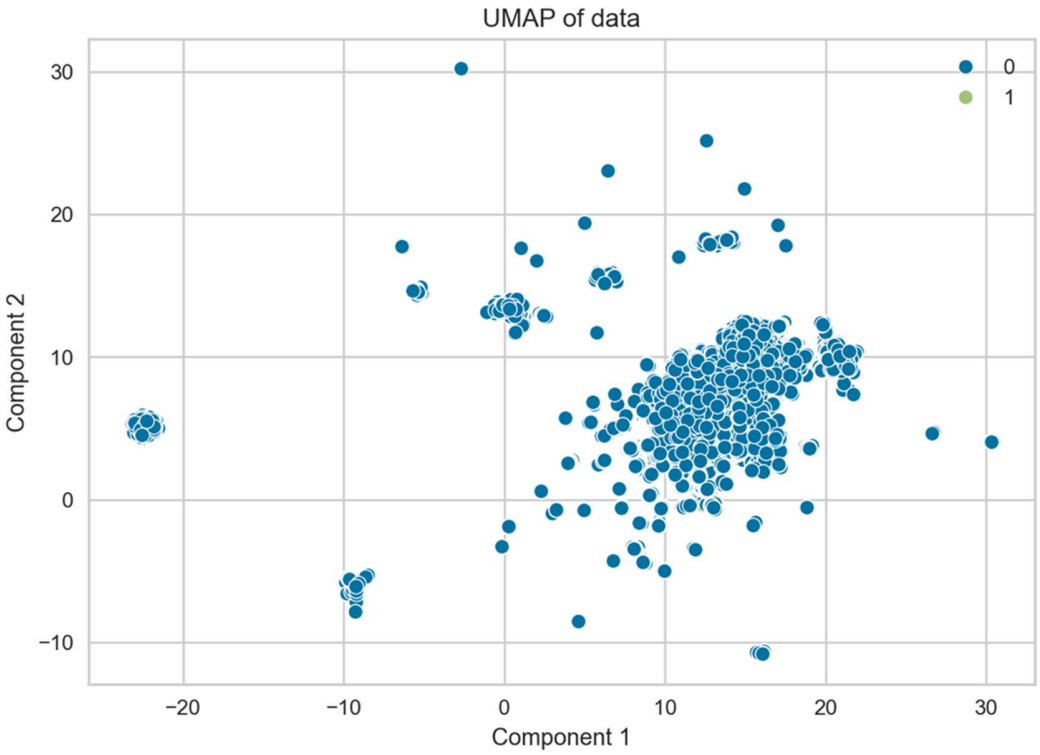

Exhibit 4: Scatter Plot of Predicted vs Actual on Test – PCA

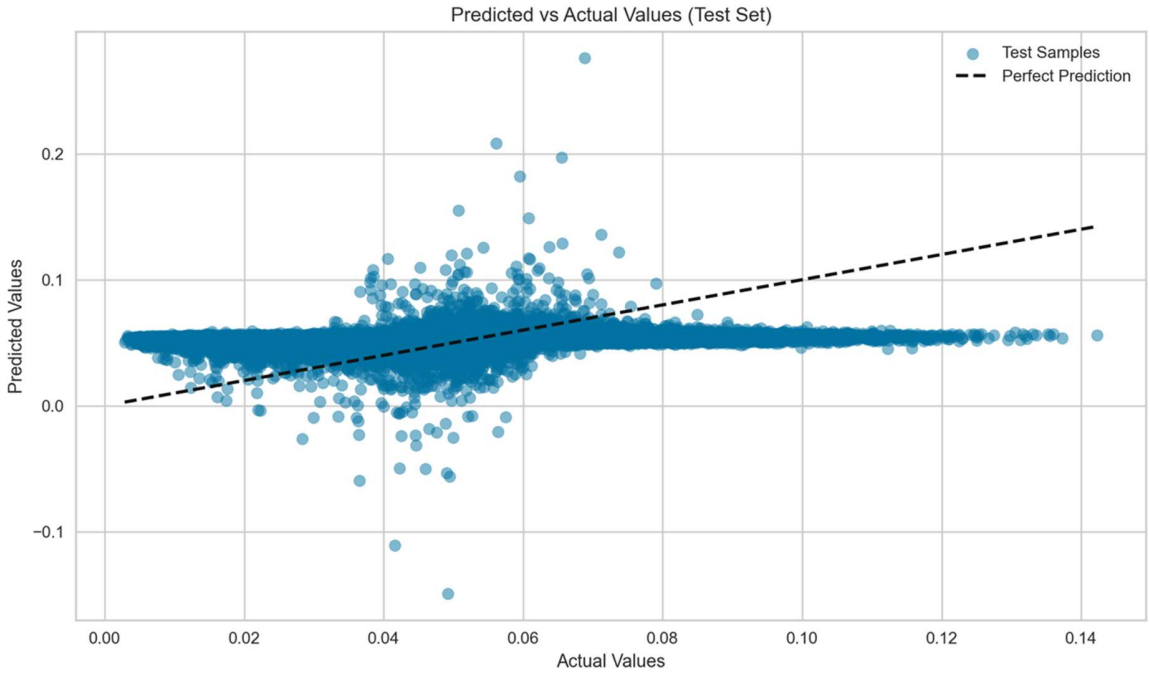

Exhibit 5: Scatter Plot of Predicted vs Actual on Test – Non-PCA

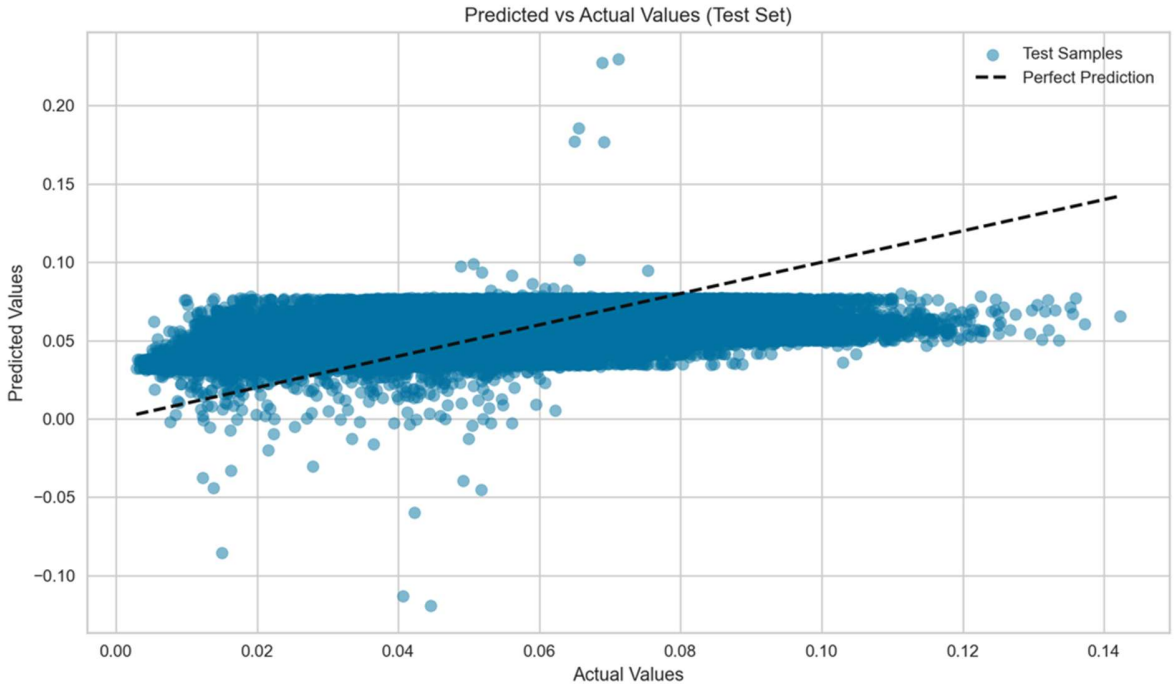

Exhibit 6: Prediction Accuracy XGBoost – PCA

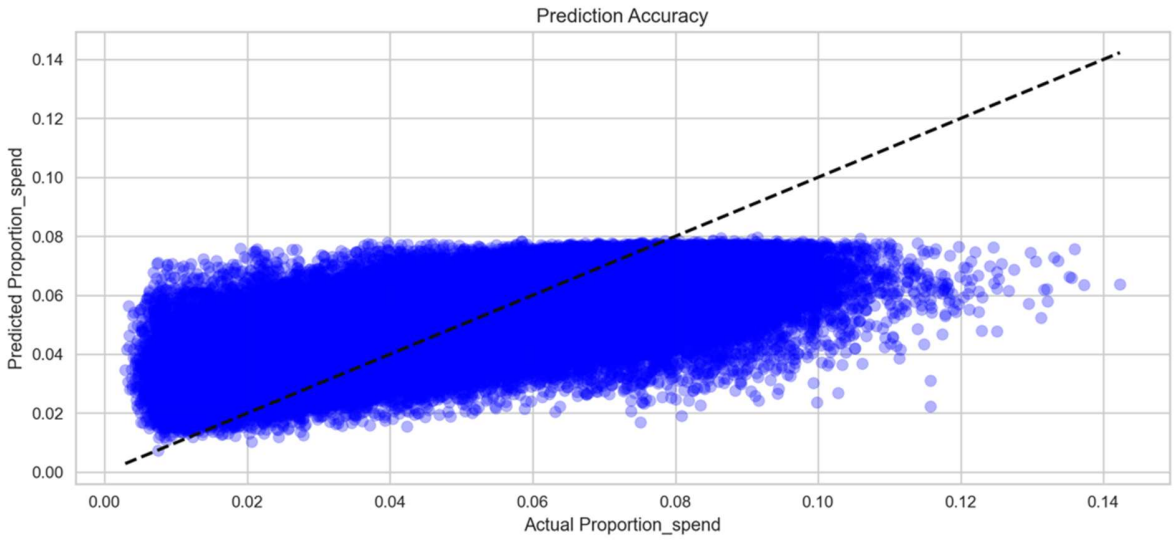

Exhibit 7: Prediction Accuracy XGBoost – Non-PCA

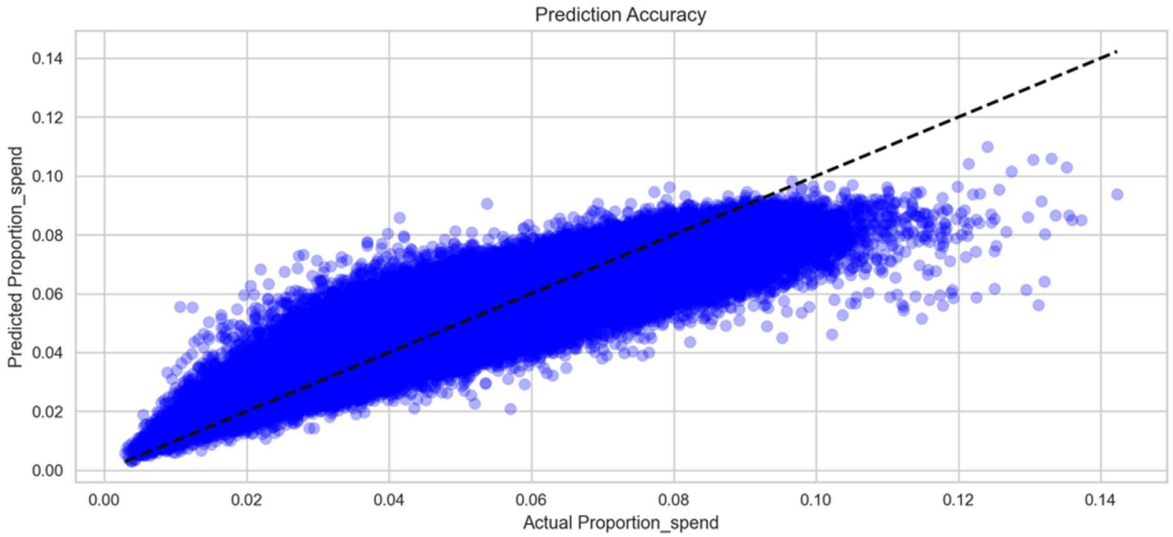

Exhibit 8: ShapValues Elastic Net – PCA

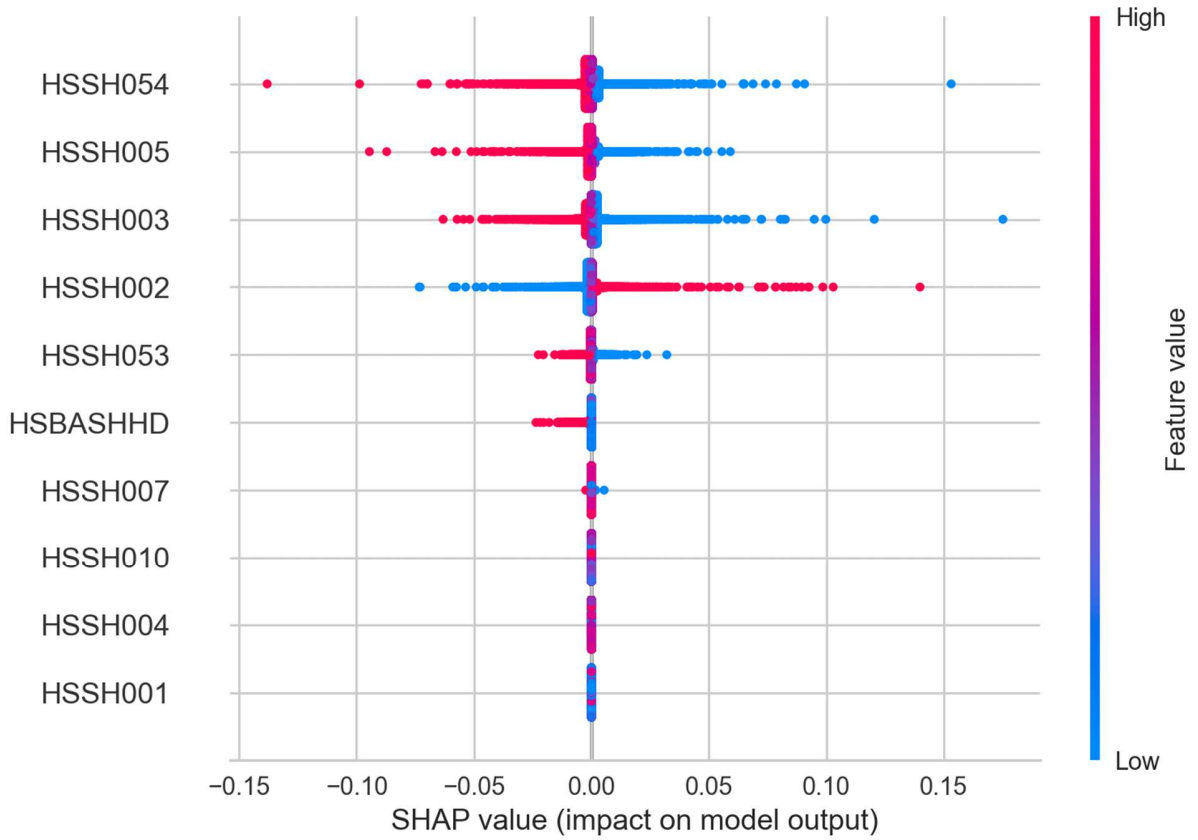

Exhibit 9: ShapValues Elastic Net – Non-PCA

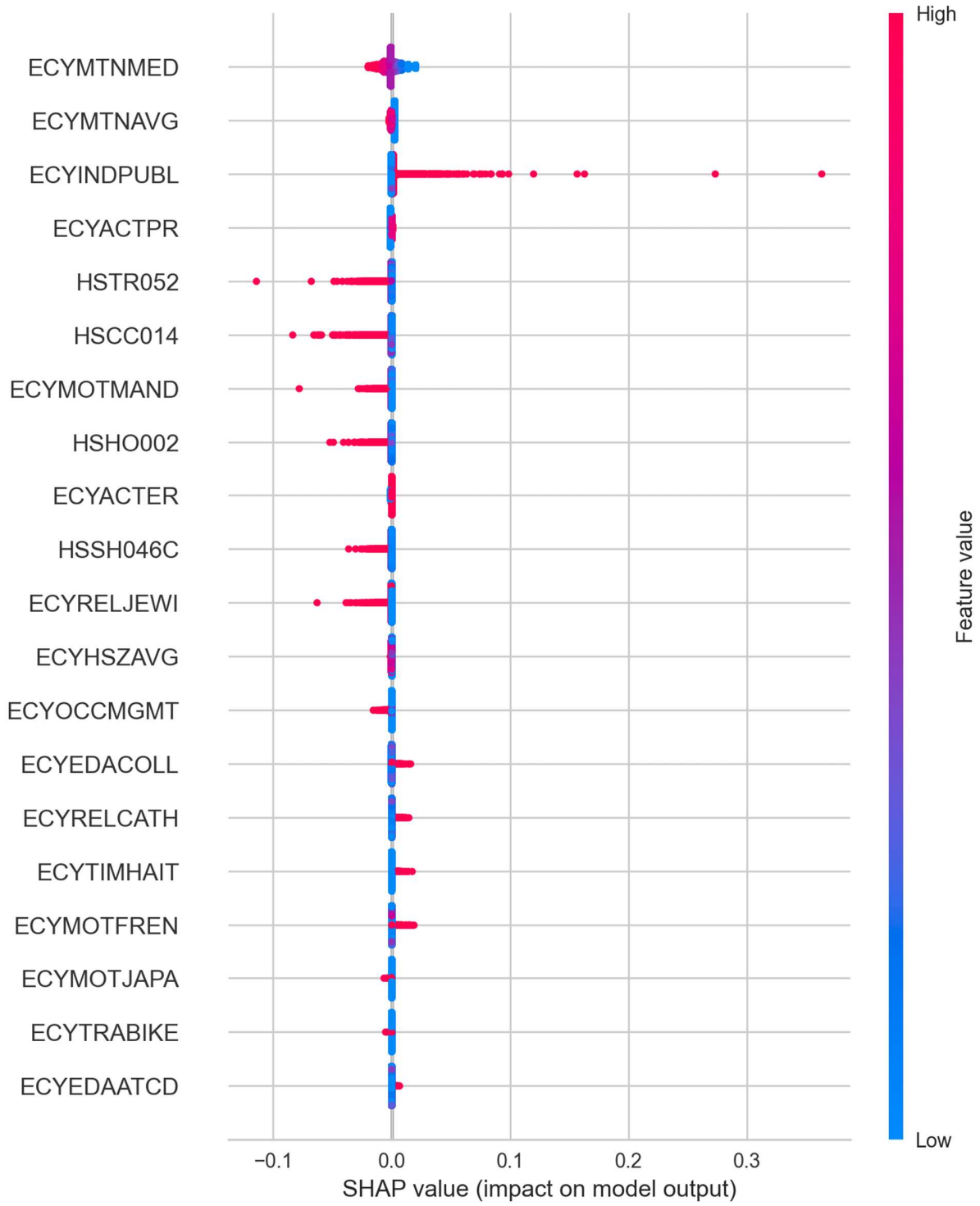

Exhibit 10: ShapValues XGBoost – PCA

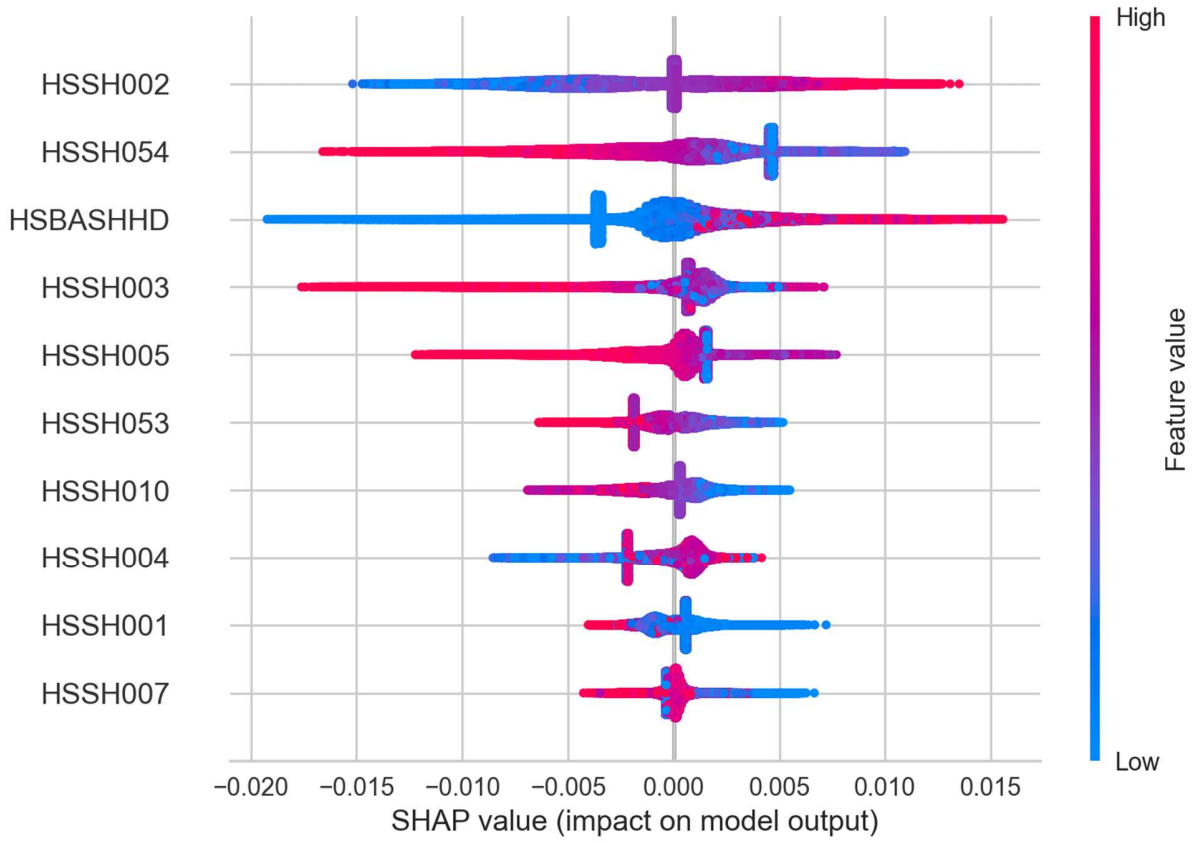

Exhibit 11: ShapValues XGBoost – Non-PCA