Rotman

GROUP Assignment Cover Sheet

Course Code:	RSM 8413	**Student Numbers**: *Please list all student numbers included in this group assignment.*	1004138024 1005976918 1009985624 1005986514 1009969205
Course Title:	Machine Learning Analysis		
Instructor Name:	Gerhard Trippen		
Assignment Title:	Group Assignment 4		
Date:	December 4, 2023		

Academic Integrity Compliance

In submitting this **group** work, we affirm:

- The student numbers listed above are correct and complete.
- The work is original. Due credit is given to others where appropriate and we have acknowledged the ideas, research, phrases etc. of others with accurate and proper citations.
- All members have contributed substantially and proportionally to this assignment.
- All members have sufficient familiarity with the entire contents to be able to sign off on this work as original.
- This is the final version of the assignment and not a draft.
- We have followed any specific formatting requirements set by the instructor.
- We are submitting this work for the correct course, via the specified platform/method (e.g., Quercus).
- We accept and acknowledge that any assignments found to be plagiarized in any way will be subject to sanctions under the University of Toronto's:

  Code of Behaviour on Academic Matters.

We agree that the statements above are true. If we have concerns, we will consult the course instructor immediately.

Optional: you may wish to follow the standard naming convention when saving files:

Complete Course Code (including Section) – GROUP – **Assignment Title**

Example: **RSM1234HF.2021-0108**-Group22-**Homework1**

Executive Summary

Introduction

Fast Retailing is a Japanese company willing to and finding ways to break into the Canadian fashion market. Its two children's brands namely Uniqlo and GU aim to cater to both high and low income brackets. Uniqlo is a casual wear fashion and a slightly pricier brand. GU on the other hand is a pocket-friendly brand for trendy clothes for budget-conscious consumers. Fast Retailing has been provided with a list of potential locations from the business development team wherein the stores of either brand (Uniqlo or GU) can be opened. This study aims to predict the annual median household income of Canadian census tracts (CTs) and to find the best set of demographic characteristics that predict income accordingly. With the help of it, the brand can make a well-informed decision with regard to the purchasing power of the surrounding population of the stores.

Methodology

First, Cluster analysis is performed to segment the CTs into relatively homogeneous groups based on demographic features in order to compare and contrast median household income among clusters. We then predict income for CTs in each cluster using segmentation modeling. The training set is segmented based on the cluster information and develops custom prediction models for each cluster. Each candidate model is first trained to predict income using all training records (e.g. the global model), and then trained individually using K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Linear Regression, and Gradient Boosting Regressor (GBR) on each cluster to predict income for the segments (e.g. the lower-income and higher-income model).

Key Findings

- For the lower income bracket, KNN serves the least MAE (Mean Absolute Error) of 7503.53.
- For the higher income bracket, GBR serves the least MAE (Mean Absolute Error) of 10187.44.
- For final predictions, the segmentation model (KNN and GBR) outperforms the global models in predicting Median Household Incomes.

1.0 Introduction to Business Case

Fast retailing is a public Japanese multinational retail holding company that wants to break into the canadian market. Fast retailing has two children brands namely Uniqlo and GU. Uniqlo is a casual wear brand for individuals seeking high-quality basics and is a slightly pricier brand. GU on the other hand is a pocket friendly brand for trendy clothes for young and budget-conscious consumers. Fast Retailing has been provided with a list of potential locations from the business development team on wherein the stores of either brand (Uniqlo or GU) can be opened.

This study aims to predict the annual median household income of Canadian census tracts (CTs), and to find the best set of demographic characteristics that predict income. Cluster analysis is first performed to segment the CTs into relatively homogeneous groups based on demographic features in order to compare and contrast median household income among clusters. We then predict income for CTs in each cluster using segmentation modelling.

With this prediction, the brand can make a well-informed decision with regard to the purchasing power of the surrounding population of the stores.

2.0 Data Preprocessing

2.1 Data Description

The training data used in this analysis is extracted from the 2021 Canadian Census provided by Statistics Canada. The training data includes 17 demographic characteristics and household income for 5371 census tracts, which are small geographical areas having a population of approximately 2000 to 8000 people (*Census tract: Detailed definition* 2018). The test data contains 855 records and all features in the training data, except for the median household income.

The data focuses on the number of occupied dwellings in each census tract by various criteria, including the construction periods, type (e.g. houses, apartment), tenure (e.g. renter, owner). Broader features such as total population and total household are also included. Please refer to **Appendix 1: Table 1** for a detailed description of all predictors that are considered.

Initial data screening identifies 126 missing values in the training set, which were dropped from the analysis as these observations do not add meaningful information. The resulting training data contains 5245 records. The test data contains no missing values.

2.2 Exploratory Data Analysis

2.2.1 Correlation Matrix

Figure 1 displays the correlation matrix among predictors, based on which we made several key observations. First, predictors within a sub-category (as defined in the previous section) tend to be correlated. For instance, total population shows strong positive correlation with total household, number of houses show moderate negative correlation with number of apartments, and total number of owners negatively correlates with number of renters. Hence, we select only a subset of predictors from each category as inputs to the model in order to avoid introducing redundant information.

Second, we observe a very strong positive correlation between total household, total occupied dwellings and total households for tenure. This is as expected as all three variables correspond with the total number of households living in a census tract. Again, we keep only one of these three variables in the model.

The third conclusion drawn from the correlation matrix is that median household income of a CT is negatively correlated with the number of households who are renters of living in apartments. This suggests that home-owners, and households living in houses, tend to have higher annual income.

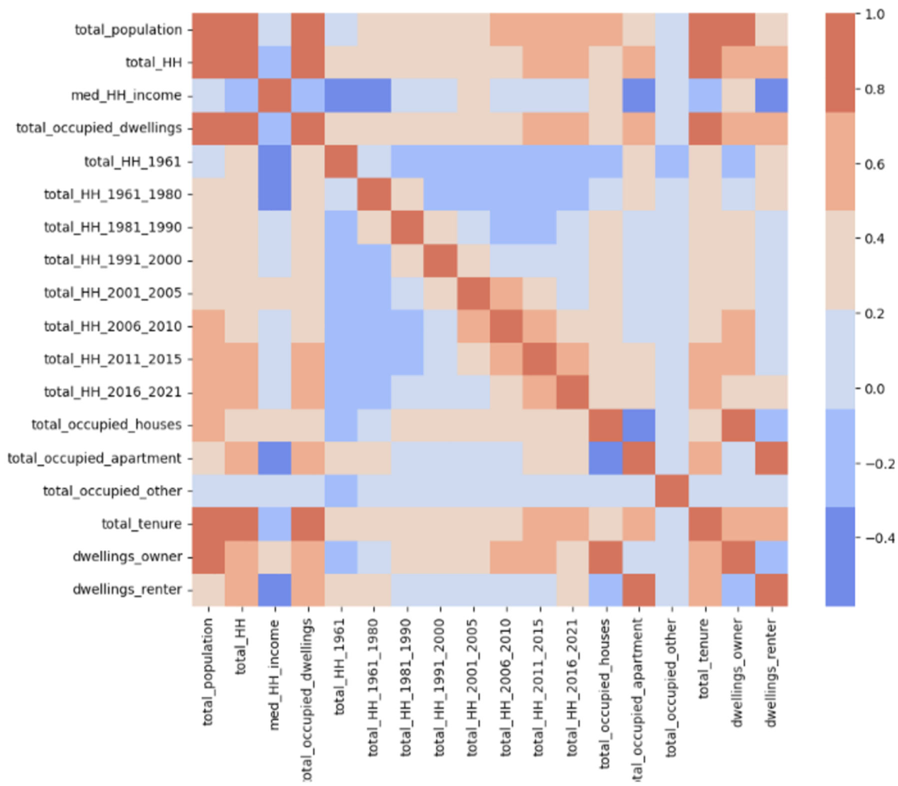

Figure 1. Pairwise correlation heatmap among predictors and the target variable. Darker blue color represents stronger negative correlation, and a darker red color represents stronger positive correlation.

2.2.2 Variable Selection

Based on the heatmap, we selected 6 demographic features that show the most evident correlation with median household income as input to the segmentation model. **Table 1** summarizes and describes the subset of predictors that are included in the analysis.

2.3 Data Cleaning

2.3.1 Converting Counts into Proportions

The predictors of interest can be categorized into 4 types: basic demographic information (ex. total household), construction period of occupied dwellings, dwelling structure and tenure of household. In each category, there is a specific column dedicated to representing the total count of the other columns within that category. To illustrate, the *Total Households for Tenure* is the aggregated counts from *Dwellings by Tenure Owner* and *Dwellings by Tenure Renter*. Therefore, we express the counts of dwellings satisfying particular criteria as a proportion relative to the total counts within that category. Converting into proportions serves as a dimension reduction technique by allowing us to include only one of highly correlated predictors in the model, and represent the other by one minus the proportion.

2.3.2 Min-Max Standardization

To prepare the data to be inputted into the cluster and segmentation models, we standardized all predictors with min-max normalization as shown in (1). Normalization ensures that all predictors range from 0 to 1, which account for the difference in scales of the original values.

$$X^* = \frac{X - min(X)}{max(X) - min(X)} \quad (1)$$

To ensure validity and interpretability of prediction results, the outcome variable (i.e. median household income) does not undergo standardization and retain its original scale.

Table 1. Predictor summary and correlation with the outcome variable (median household income)

Type	Original Column Header	Description	Correlation with Response
Basics	Total Households	Total number of households	-0.15
Construction Period	Total Households For Period Of Construction Built Before 1961	Occupied private dwellings built between the specific construction period	-0.38
	Total Households For Period Of Construction Built Between 1961 And 1980		-0.36
Household Structure	Total Households For Structure Type Houses	Total number of households that occupy houses	0.44
	Total Households For Structure Type Apartment, Building Low And High Rise	Total number of households that occupy apartments	-0.53
Tenure	Dwellings by Tenure Renter	Total number of households that are renters	-0.58

Rotman

3.0 K-Means Clustering

We first employ the K-means algorithm to cluster the census tracts into relatively homogenous groups based on the predictors identified in **Table 1**. The resulting clusters would allow us to explore the potential differences in median income among CTs. As K-means is an unsupervised learning algorithm, the median household income variable is excluded from the model.

3.1 Optimizing Number of Clusters k

3.3.1 Elbow Method

The Elbow method is initially used to determine the optimal number of clusters *k*. The Elbow method calculates the within-cluster sum of squares (WCSS), or the squared distance between each point and the cluster centroid, for various numbers of clusters (Saji, 2023). Ideally, the value of *k* after which the decrease in WCSS plateaus will be selected as the optimal number of clusters. However, the Elbow method produces a rather smooth curve (see **Appendix 2: Figure 1**). While the curve suggests $k = 2$, it is less clear whether there exists another turning point after which the rate at which WCSS decreases starts to approach 0. We therefore seek further insight from the Silhouette method in choosing the best *k*.

3.3.2 Silhouette Analysis

The silhouette score quantifies the distance between an observation and its cluster center, in contrast to its distance to the next closest cluster. In other words, the silhouette scores measure the cohesion within a cluster and separation between clusters, with higher values indicating better clustering. Silhouette score is calculated for *k* values ranging from 2 to 6, and is maximized for $k = 2$ (see **Appendix 2: Figure 2**). This result is in line with the optimal *k* chosen by the Elbow method.

3.2 Cluster Summary and Description

The K-means algorithm with $k = 2$ yields two clusters of unequal sizes. Namely, the algorithm identified a lower-income cluster and a higher-income cluster with 1860 and 3390 CTs, respectively. Further insights into the two clusters are obtained after conducting statistical analysis and visualization.

3.2.1 Distribution of Median Household Income

Table 2 summarizes the distribution of median household income of CTs in the lower-income and higher-income clusters. The CTs within the higher-income cluster have an average median household income of $91,178, whereas the lower-income cluster shows a lower average of $60,907. Furthermore, the higher-income cluster displays both a higher maximum and minimum median income compared to the lower-income cluster. These findings indicate a considerable disparity in income between the two clusters, suggesting that income can serve as a distinguishing factor for categorizing CTs.

Table 2. Median household income distribution in lower-income (red, 0) and higher-income (blue, 1) clusters.

Cluster	0 (Lower income cluster)	1 (Higher income cluster)
Cluster size	1855	3385
Maximum median income	128000.00	240000.00
Average median income	60907.37	91178.55
Minimum median income	23400.00	40400.00
Standard deviation	13876.58	20657.45

3.2.2 Summary of Demographic Features for the Clusters

Figure 2 showcases the percentage of occupied apartments (versus houses), owners (versus renters) and occupied dwellings built before 1961 and between 1961-1980. CTs in the higher-income (blue) cluster appear to have more households living in houses, and less living in apartments. In addition, the higher-income census tracts also have more households that are owners of dwellings, whereas the lower-income have fewer owners and more renters. We also observe that a smaller proportion of occupied dwellings in the higher-income CTs are built before 1961, or between 1961 and 1980, compared to the lower-income CTs. We can conclude that the lower-income cluster are mostly apartment renters living in relatively older dwellings, whereas the higher-income cluster are mainly house owners living in relatively newer dwellings.

After comparing the above insights on the census tracts and the list of potential opening locations provided by the business development team, we can determine which locations are more suitable for opening UNIQLO stores, and which are more suitable for GU stores. In other words, we should open UNIQLO stores in areas that have a higher median household income and GU stores with a lower median household income.

4.0 Segmentation Modelling

To enhance prediction performance, we segment the training set based on the cluster information and develop custom prediction models for each cluster. We consider a basket of candidate regressors, including linear regression, KNN, ANN, and gradient boosting regressor (GBR). Each candidate model is first trained to predict income using all training records (e.g. the global model), and then trained individually on each cluster to predict income for the segments (e.g. the lower-income and higher-income model).

For KNN, ANN and GBR, hyperparameter tuning is done through grid search with 5-fold cross validation. The model is cross-validated over an arbitrary set of feasible hyperparameter values, and the optimal value would yield the lowest average validation error. We choose the model that yields the highest prediction accuracy (lowest mean absolute error) as the best regressor.

The overall prediction accuracy of the segmentation model is assessed by the weighted average MAE of the best candidate models on each cluster. The clusters are weighted by their relative proportions in the dataset. Specifically, the lower-income cluster (n=1855, 35.4%) receives a weight of 0.35 and the higher-income cluster (n=3385, 64.6%) receives a weight of 0.65.

4.1 Linear Regression

Our initial attempt is to fit linear regression to predict median household income. We first fit a global linear model on all training records. To account for non-linearity in median income and the predictors, we also fit a linear regression on the *square-root* and *log* income. The global model yielding the highest adjusted R-squared is then applied to each cluster. Model fit is assessed within each segment.

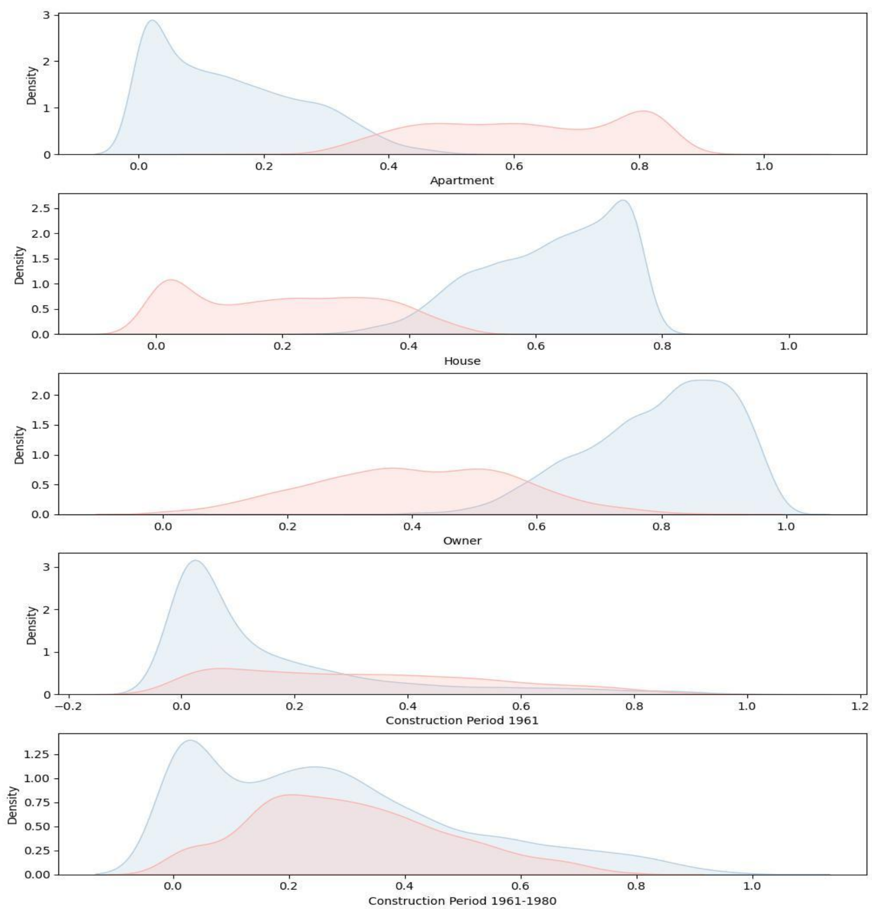

Figure 2. Comparing demographic features (proportion of occupied dwellings by dwellings structures, type of tenure and construction periods) between the two clusters.

4.2 K Nearest Neighbours Regression

In addition to linear regression, a weighted KNN regressor is also built to predict income. The training records are weighted by their distance to the new data, assigning farther away observations with less weight. The optimal number of neighbours *k* is chosen through a 5-fold cross-validation grid search over

k values ranging from 1 to 30. The k value resulting in the lowest average mean absolute error (MAE) is selected as the optimal k.

4.3 Neural Network

We consider a sequential neural network as the third candidate model, which would have one input layer, one or multiple hidden layers, and an output layer. To account for a limited training set size and avoid overfitting, we primarily consider simpler networks with one to two hidden layers. The RELU activation function is adopted for all hidden layers, and a sigmoid activation function is used for the output layer. The learning rate is fixed at 0.001. In addition, we consider the mean squared error as the loss function, with the Adam classifier used for back-propagation.

The network structure was explored through a grid search with 5-fold cross-validation. Various model configurations were considered, including different batch sizes (e.g. 32, 64, 128) and number of hidden layers. A single hidden layer with 4, 8, 12, and 24 nodes, as well as two hidden layers with 24 and 12 nodes, and 36 and 18 nodes are considered. To determine the optimal number of epochs is chosen to be the value that minimizes the validation MAE.

4.4 Gradient Boosting Regressor

The fourth candidate model is the gradient boosting regressor (GBR), which is an ensemble learning method that aggregates the prediction of a sequence of weaker models to build a stronger model. In GBR, each model is built sequentially to account for the error of the previous model (Masui, 2022). The hyperparameters are tuned through a 5-fold cross-validation grid search. The grid search is run over learning rates ranging from 0 to 1, maximum depths from 2 to 5, and various numbers of boosting stages ranging from 100 to 1000. This tuning process was repeated for each cluster to identify the optimal hyperparameters for each subgroup.

5.0 Results and Conclusion

The mean absolute errors from 5-fold cross-validation for the best candidate models are listed in **Table 3**. Detailed discussion of model performance is provided in the following sections.

Table 3. Mean absolute errors for the global and segmented optimal models. The baseline MAE from global linear regression on original-scaled income is **10437.30**. Best MAE on each cluster is *italicized*.

MAE	**LR (log)**	**KNN**	**ANN**	**GBR**
Global	0.13	9230.82	80442.68	9276.99
Lower-income	0.13	***7303.53***	60906.37	7346.44
Higher-income	0.12	10321.78	91177.55	***10187.44***

5.1 Linear Regression

We fit three candidate global regression models on median income, square-root income and log-transformed income using all predictors. Regression on original-scaled median income yields an adjusted R-squared of 0.63 and an MAE of $10437.30. While the original-scaled linear model does not achieve the best fit, we use its MAE as a baseline of comparison with other models.

Global linear regression on log-transformed income (log model) achieves the highest adjusted R-squared of 0.69 among all three candidate global regression models. We thus proceed to fit the log model individually on each cluster to assess the model's performance in segmented data.

The log model applied to the lower-income cluster results in a MAE of 0.13, while the higher-income cluster exhibits a slightly lower MAE of 0.12. This implies that, on average, the predictions from the log model deviate from the true log income values by approximately 0.12 to 0.13 units.

However, despite the slight reduction in MAE for the segmented models compared to the global log model (MAE = 0.13), the adjusted R-squared for the segmented data were 0.47 and 0.42 for the lower and higher-income clusters, respectively. The notable decrease in model fit compared to the global model undermines the effectiveness of segmentation in improving predictive accuracy of linear models.

5.2 K Nearest Neighbours

The global KNN model is optimized with k=19, yielding a cross-validation MAE of $9230.82. In other words, the predicted income based on the 19 nearest neighbours deviates, on average, by $9231 from the true median household income.

Upon segmentation, the lower-income model is optimized with k=13, resulting in a reduced cross-validation MAE of $7303.53. This reduction in validation errors suggests improved performance of the KNN model on the lower-income cluster.

The higher-income model was optimized with k=19, but it results in a higher cross-validation MAE of $10321.78. The increase in MAE suggests worse performance of the KNN model on the higher-income cluster compared to the global model.

5.3 Neural Network

5.3.1 Grid Search Outcome

Based on a grid search with 5-fold cross-validation, the optimal MAE for both the global and segmented models is achieved with a *batch size* of 32. Further investigation over various numbers of *epochs* shows that the reduction of MAE plateaus at an epoch size of around 10 (refer to **Figure 3**). The ANN models are thus trained with 10 epochs for both the global and segmented datasets.

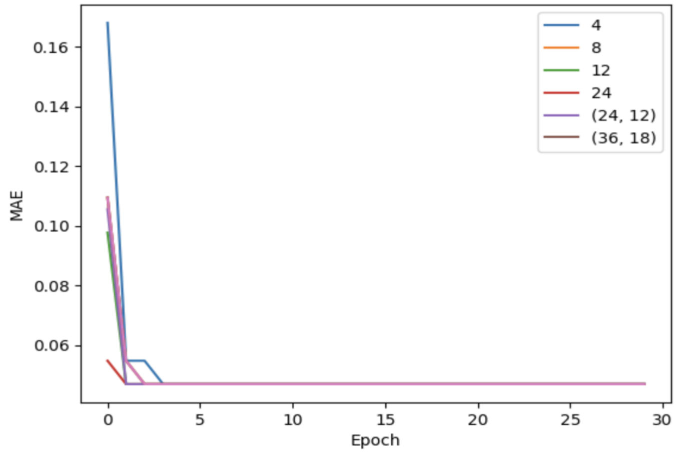

Figure 3. Global ANN mean absolute validation error by layer structure and epoch sizes, with fixed batch size = 32 and learning rate = 0.01

5.3.2 Network Optimization

The global ANN, optimized with an input layer comprising 6 nodes (one for each predictor), a single hidden layer with 4 nodes, and an output layer, exhibited a cross-validation MAE of $80442.68. This MAE is notably higher than the validation errors obtained from linear regression and K Nearest Neighbours (KNN). It is important to note that the observed increase in MAE is anticipated, considering the relatively limited amount of data available for training the neural network.

For Cluster 0, the ANN was configured with a single hidden layer comprising 24 nodes, resulting in a MAE of $60906.37. While this demonstrated better performance (lower MAE) compared to the global model, it is noteworthy that the MAE remains significantly higher than that observed with linear regression and KNN models.

In the case of Cluster 1, the ANN was configured with two hidden layers, consisting of 36 and 18 nodes, respectively. Despite the additional complexity, the model yielded a higher MAE of $91177.5 compared to the global model. This outcome suggests that the ANN's performance on Cluster 1 is suboptimal when compared to the simpler global ANN configuration.

5.4 Gradient Boosting Regressor

The global GBR model is optimized with a depth of 5, a learning rate of 0.0183, and 500 boosting stages. This configuration yielded a cross-validation MAE of $9276.99. Although the MAE is lower than that achieved by linear regression and the neural network, it is higher than the K Nearest Neighbours (KNN) model. The fit on the global data is reasonably good with a R-squared of 0.69.

For the lower-income cluster, the GBR is optimized with a depth of 5, a learning rate of 0.0264, and 200 boosting stages. We see a notable reduction in MAE, achieving $7346.44. However, it remains higher than the error achieved by the KNN model on the lower-income cluster.

The higher-income model was optimized with a depth of 4, a learning rate of 0.0183, and 500 boosting stages. The model yields a MAE of 10187.44, the lowest among all candidate models on the higher-income cluster. However, the model performance on this segment is still higher than on the global data.

Extensive model training outcomes are summarized in **Appendix 3**.

5.5 Overall Prediction Accuracy

In the evaluation of candidate models, the KNN model achieves the lowest MAE of 7303.53 on the lower-income cluster. Conversely, the GBR achieves the optimal MAE of 10187.44 on the higher-income cluster. Thus, we adopt the **KNN model on the lower-income cluster** and **GBR model on the higher income cluster**. The weighted average MAE is 9187.07, which is lower than the MAE achieved by all global models. The results suggest that the segmentation model outperforms the global models in predicting median household income.

Rotman

6.0 References

Census tract: Detailed definition. Statistics Canada: Canada's national statistical agency / Statistique Canada : Organisme statistique national du Canada. (2018, September 17). https://www150.statcan.gc.ca/n1/pub/92-195-x/2011001/geo/ct-sr/def-eng.htm

Masui, T. (2022, February 12). *All you need to know about gradient boosting algorithm − Part 1. regression*. Medium. https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Saji, B. (2023, September 20). *Elbow method for finding the optimal number of clusters in K-means*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

7.0 Appendices

***Appendix 1**: Data Description and Distribution of Variables*

A1.1 Data Description

Table 1: Description of input and output variables

Type	Original Column Header	Variable Name	Description
Output	Median Household Income (Current Year $)	med_HH_income	Annual median household income
Basics	Total Population	total_population	Total population in CT
	Total Households	total_HH	Total number of households
Construction Period	Total Households For Period Of Construction	total_occupied_dwellings	Occupied private dwellings (total household) for period of construction
	Total Households For Period Of Construction Built Before 1961	total_HH_1961	Occupied private dwellings built between the specific construction period
	Total Households For Period Of Construction Built Between 1961 And 1980	total_HH_1961_1980	
	Total Households For Period Of Construction Built Between 1981 And 1990	total_HH_1981_1990	
	Total Households For Period Of Construction Built Between 1991 And 2000	total_HH_1991_2000	
	Total Households For Period Of Construction Built Between 2001 And 2005	total_HH_2001_2005	
	Total Households For Period Of Construction Built Between 2006 And 2010	total_HH_2006_2010	
	Total Households For Period Of Construction Built Between 2011 And 2015	total_HH_2011_2015	
	Total Households For Period Of Construction Built Between 2016 And 2021	total_HH_2016_2021	
Household Structure	Total Households For Structure Type Houses	total_occupied_houses	Total number of households that occupy houses
	Total Households For Structure Type Apartment, Building Low And High Rise	total_occupied_apartment	Total number of households that occupy apartments
	Total Households For Structure Type Other Dwelling Types	total_occupied_other	Total number of households that occupy other type of residences (non-house,

			non-apartment)
Tenure	Total Households for Tenure	total_tenure	Total households for tenure
	Dwellings by Tenure Owner	dwellings_owner	Total number of households that are owners
	Dwellings by Tenure Renter	dwellings_renter	Total number of households that are renters

A1.2 Cluster Description

Table 2 shows a numerical summary comparing the distribution of standardized predictors in the lower (0) and higher-income (1) clusters.

Table 2: Cluster description of standardized predictors. A value closer to 1 indicating larger values on the original scale of a predictor.

Predictors	Cluster	Minimum	Mean	Maximum	Standard Dev.
Total household	0	0.0058	0.2534	1	0.1205
	1	0	0.2078	0.9520	0.0982
Dwellings before 1961	0	0	0.3198	0.9936	0.2286
	1	0	0.1510	1	0.2012
Dwellings between 1961 and 1980	0	0	0.3052	0.8047	0.1675
	1	0	0.2736	1	0.2268
Houses	0	0	0.1980	0.6099	0.1432
	1	0.0185	0.6240	1	0.1092
Apartments	0	0.1685	0.6194	1	0.1578
	1	0	0.1435	0.5263	0.1164
Owner	0	0	0.4074	0.8852	0.1617
	1	0	0.7955	1	0.1178

***Appendix 2**: Optimization and Results of K-Means Clustering*

Elbow Method.

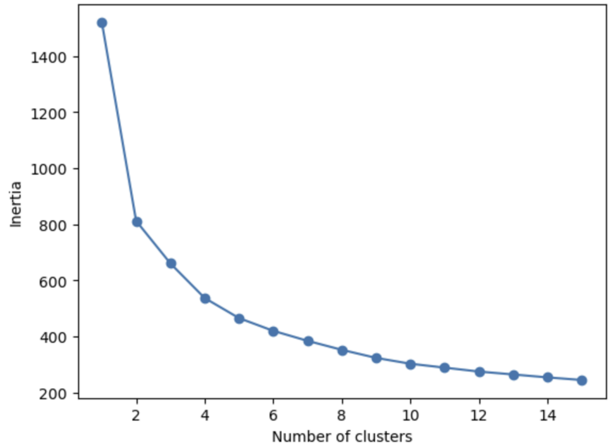

Figure 1: Elbow method, showcasing inertia (within-cluster sum of squares) by the number of clusters.

Silhouette Method.

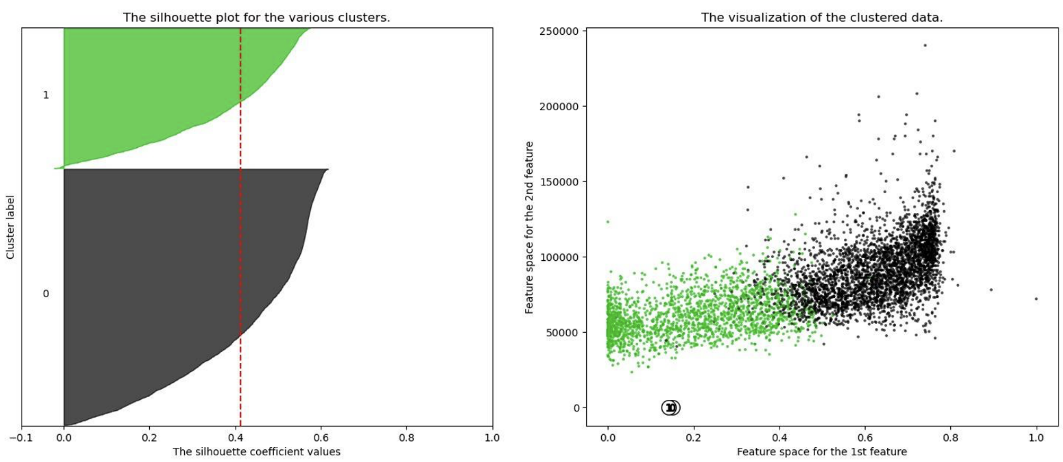

Figure 2. Silhouette analysis for K-means clustering on the training data, with $k = 2$.

Rotman

Appendix 3*: Model Results*

A3.1 Linear Regression

Global model. Measures on goodness of fit and validation accuracy for global linear models are listed in Table 1, where Model 1, 2, and 3 are fit on median household income, square-root income, and log income, respectively.

Table 1: Global linear regression results.

	Model 1	**Model 2 (Sqrt)**	**Model 3 (Log)**
R-squared	0.64	0.67	0.69
Adjusted R-squared	0.63	0.67	0.69
5-fold CV error (average MAE)	10437.30	17.90	0.13

Segmentation models. Measures on goodness of fit and validation accuracy for the *log* model on the lower-income and higher-income clusters is shown in Table 2.

Table 2: Segmented linear regression (on log median income) results, in comparison with the global model.

	Global	**Lower-Income**	**Higher-Income**
R-squared	0.69	0.48	0.43
Adjusted R-squared	0.69	0.47	0.42
5-fold CV error (average MAE)	0.13	0.13	0.12

A3.2 KNN

The optimal k and cross-validation error for the global and segmented KNN model is summarized in **Table 3**. We see that the lower-income model yields a lower cross-validation error than the global model.

Table 3: KNN results.

	Global	**Lower-Income**	**Higher-Income**
Optimal *k*	19	13	19
5-fold CV error (average MAE)	9230.82	7303.53	10321.78

A3.3 ANN

The optimal neural network design and validation errors are summarized in **Table 4**.

Table 4: Average mean absolute error from 5-fold cross-validation for various ANN structures. ANN is built using a batch size of 32 and 10 epochs.

		Global	Lower-Income	Higher-Income
Optimal Design	Hidden layers	1	1	2
	Nodes	4	24	(36, 18)
5-fold CV error (average MAE)	-	80442.68	60906.37	91177.55

A3.4 Gradient Boosting Regressor

The optimal GBR hyperparameters and validation errors are summarized in **Table 5**.

Table 5: Grid search result for gradient boosting regressor.

	Global	Lower-Income	Higher-Income
Learning rate	0.0183	0.0264	0.0183
Boosting stages	500	200	500
Max Depth	5	5	4
Validation R-squared	0.69	0.55	0.43
5-fold CV error (average MAE)	9276.99	7346.44	10187.44