Coursework – Canadian Households

This coursework is focused on studying Canadian households and their spending/saving habits. For this, we have been exclusively provided with the data from Environics Analytics, a leading company that has "Canada's largest collection of high-quality, actionable data". We have been provided with an exclusive license to use this data for the course. Please note that **THIS DATA CANNOT BE SHARED AND MUST BE DELETED AFTER FINISHING THE COURSE**. Please read the data disclosure agreement at the end of the coursework description.

You have been provided with two datasets:

- Household spending per dissemination area.
- Demographics data per dissemination area.

The documentation of each dataset can be seen in the Documentation folder of each dataset.

With this information, write a professional report with your analysis of the data, answering the following questions.

1. **(40%) Clustering and Dimensionality Reduction.** The first part of the coursework will focus on identifying the characteristics of Canadian households, excluding their pension behaviour. For this, do not include, in your clustering and dimensionality reduction models, the target of the regression model in Part 2.
    a. (10%) Clean the data using your knowledge of the course. For context, you will create in later questions a model that predicts a household's proportion of income spent on total personal insurance premiums and retirement/pension contributions and apply a clustering algorithm over the full data.
    b. (10%) Create a K-Means clustering of the data, identifying the optimal number of clusters using both the silhouette and the elbow method. Do they agree?
    c. (10%) Now we will apply a linear dimensionality reductions technique to the data.
        i. Apply PCA to your data.
        ii. Plot the data of the first two PCs in a scatterplot and colour the points as per the cluster labels you calculated in the previous step. What can you say about your data? Are your clusters clearly defined in the output? Interpret the first three components of your PCA output.
        iii. For the first three components, calculate the average value of each component, for each cluster, so your data should look like a table with cluster number and average component value. Give a name to the clusters from this output and justify your choice.
    d. (10%) Now, use UMAP to reduce the data to two dimensions. Justify your choice of parameters by searching for the optimal value as you deem reasonable. Again, colour the data to differentiate each cluster that you named. What do you see? Is your UMAP a better or worse interpretation than PCA?

2. (50%) **Regression.** Now we will create models for a household's proportion of income spent on total personal insurance premiums and retirement/pension contributions.
    a. (15%) Train a regularized elastic net linear regression from your data.
        i. Create your target variable from the variables in the dataset. Do not use those components on the training database.
        ii. Apply any data transformation / variable creation you deem necessary to obtain a good result.
        iii. Discuss the grid that you chose to search for the parameters and the output that you obtained.
        iv. For your test set, create a scatterplot of the original response and the predicted response. Report the MSE and R2 on the test set and calculate a bootstrapped confidence interval of the output.
        v. Interpret the coefficients of the top five most important variables in the regression.
    b. (20%) Now we will train an XGB model.
        i. Follow the same process for the previous question to prepare the data, if necessary.
        ii. Discuss your choice of parameter grid and how you searched for the optimal parameters.
        iii. Calculate the same scatterplot and performance measures as before. Compare the CIs of the model versus the linear model. Is it better? Worse?
    c. (15%) Using the output of the XGB model and your test dataset, interpret the results using the Shap values arising from your model. What are the most important variables? Interpret the behaviour of the top five most important variables using the most appropriate plot for this. How do these results compare to the coefficient analysis in the linear model? Do you think this problem is non-linear?

The remaining 10% of the coursework corresponds to the format and language evaluations.

Conditions of the coursework

Software: You must use Python to run the numerical calculations over the dataset. A copy of your Jupyter notebook must be attached to the coursework in the submission (submit both the ipynb file and your report in PDF format). The notebook text MUST be attached in ipynb format, otherwise **a 25% discount will apply**.

Word Limit: 2500 words +/-10% either side of the word count is deemed to be acceptable. Any text that exceeds an additional 10% will not attract any marks. The relevant word count *includes* items such as cover page, executive summary, title page, table of contents, tables, figures, in-text citations and section headings if used. The relevant word count *excludes* your list of references and any appendices at the end of your coursework submission.

You should always include the word count (from Microsoft Word, not Turnitin), at the end of your coursework submission, before your list of references.

Title/Cover Page: You must include a title/cover page that includes: your group's Student IDs and names, Course Code, Assignment Title, Word Count.

Submission Deadline: April 11th, 23:59.

Turnitin Submission: The assignment MUST be submitted electronically via Brightspace OWL. All required papers may be subject to submission for textual similarity review to the commercial plagiarism detection software under license to the University for the detection of plagiarism. All papers submitted for such checking will be included as source documents in the reference database to detect plagiarism of papers subsequently submitted to the system. Use of the service is subject to the licensing agreement, currently between The University of Western Ontario and Turnitin.com (http://www.turnitin.com).

IMPORTANT DATA USAGE NOTICE

Environics Analytics (EA) (https://environicsanalytics.com/about) has kindly provided this dataset for use within Data Science 3000B - Introduction to Machine Learning. The dataset provided by Environics Analytics for this coursework is confidential and protected under a Non-Disclosure Agreement (NDA) signed between EA and UWO. By accessing this data, you acknowledge and agree to the following conditions:

1. The data is watermarked and can only be used for coursework purposes within Data Science 3000B - Introduction to Machine Learning.
2. Sharing, distributing, or using this data outside of DS 3000B coursework is strictly prohibited.
3. All copies of the data must be permanently deleted upon completion of the course.
4. Any unauthorized use, sharing, or retention of this data may result in legal action.

By proceeding with this coursework, you acknowledge these terms and agree to comply with all data usage restrictions.