Technical Assessment - Data Scientist Position at Nzuri Strategy

I. Medical Expenditure Analysis Report

Use the provided **NMES (National Medical Expenditure Survey)** data set (folder contains columns dictionary) for persons 40 and older to build a regression model that estimates the risk of having annual expenditure greater than $20,000. Your model should include indicators of lung cancer (lc5) and heart disease (chd5) and other demographic and socio-economic variables you find are useful predictors. The goal of the analysis is to estimate the rate (risk) of large (>$20,000) positive expenditures among individuals.

Questions:

- **For your final model, estimate its coefficients and check the model for consistency with the observations by comparing the observed rates within several bins of predicted rates. Check for extremely influential observations in your final model.**
- **Use your final model to calculate the sensitivity and specificity for classifying a person as having a large expenditure (or not) as a function of classification threshold. Estimate the area under the ROC curve. Compare your final model with one that only has lc5, chd5, age and gender (main effects only) using area under the ROC curve. Compare your area under the curve with and without cross-validation.**

  Does cross-validation make a difference in this case; why or why not?

- **Summarize your final model and its ability to predict large expenditures in a paragraph as if for a public health journal (Model explainability is expected here)**

Make sure to explain the choice of your final model and provide step by step model selection method. Include visualizations that supports your model selection and those that support the final model performance.

Investigate the data thoroughly and provide checks that a statistician would conduct.

For the final report, you may provide a RMarkdown file or JupyterLab file or LaTex , all rendered in PDF format.

Create a folder named Assessment_PartI_{InsertYourLastName}, include the **Analysis report** (with detail explanations of each step, visualizations, answering the highlighted Questions above) file and the **Code** file (R or Python).

II. Building an XGBOOST model using House Prices Data

Build an XGBoost model using the provided House Prices Data, columns are explained in the data folder, the data is already split into train an test sets.

Include hyperparameter tuning and model explainability.

Provide code of your analysis (**R or Python**) and one page report of the model explainability plots with interpretation of these plots in your own words.

Create a folder named Assessment_PartII_{InsertYourLastName}, include the **Code** file (R or Python) and one page in PDF or Word called **Plots** (including the model explainability plots with interpretation of these plots in your own words)