**Milliman 2023 Data Science Summer Internship Take-Home Challenge**

**Overview**

The Minneapolis Health & Life Data Science team works on various projects, including building custom predictive models for our clients along with other tasks. Therefore, in order to gauge your skills, we would like you to complete a take-home data science challenge. You can complete the challenge in R or Python or Julia and are welcome to use any IDE or developer tools that you would like. Once you have finished the challenge we ask that you reply to this email with a single PDF or HTML document that displays your code and any supporting output or visualizations. Please note that we will not be responding to any questions at this stage of the interview process – we would like to see what you are able to do without direction.

**Task Description**

An Accountable Care Organization (ACO) has engaged us to assist in helping them develop a predictive modeling solution to reduce the prevalence of strokes in the population they manage. Their plan is to develop an outreach program where a team of health coaches will work with members to reduce their risk of stroke. However, they do not have enough resources to engage every member of their population. Therefore, they would like us to develop a model that can be used to match the health coaches with those individuals who have the highest risk of stroke. To solve this problem, we would like you to first build a generalized linear model (GLM). If time permits, we would then like you to try to build a more advanced model to compete with the benchmark performance of your GLM. We are interested to see both the difference in performance and your thoughts on any important advantages or disadvantages to either method.

**Data**

The data set for this challenge is attached to this email in the “stroke\_data.csv” file. Please note that this data set is purely hypothetical and is not actual client data. Fields should be self-explanatory, and you are welcome to handle them in any way that you’d like.

**Tips**

As previously mentioned, while we are giving you 11 days to complete this take-home challenge, we do not recommend spending more than a baseline of 6-10 hours on this project. We understand that this is not enough time to fully flesh out everything that you might want to do, so feel free to note what you would’ve attempted with more time available.

Keep in mind that high predictive performance is the goal – but try not to get carried away in developing the very best possible model. What we are trying to judge from this task, primarily, is the thought process that you use to tackle the exercise along with your problem solving, coding, and documentation skills.

**Dataset Explanation:**

The stroke dataset is a collection of data related to individuals who have had a stroke, as well as some who have not. It contains information on various factors that are known to be associated with stroke, including age, gender, marital status, hypertension, heart disease, occupation, place of residence, smoking status, and various health metrics. The dataset is commonly used in machine learning and statistical analysis to develop predictive models for stroke risk.

Here is a brief explanation of each column in the dataset:

id: A unique identifier for everyone in the dataset.

gender: The gender of the individual, recorded as "Male", "Female", or "Other".

age: The age of the individual in years.

married: Whether the individual is currently married or not, recorded as "Yes" or "No".

hypertension: Whether the individual has hypertension (high blood pressure), recorded as 1 if present and 0 if not.

heart\_disease: Whether the individual has a history of heart disease, recorded as 1 if present and 0 if not.

occupation: The occupation of the individual, recorded as "Goverment Job", "Private Job", "Self-employed", "Children", or "Never Worked".

residence: The type of residence of the individual, recorded as "Urban" or "Rural".

metric\_1: A health metric related to the level of glucose in the blood.

metric\_2: A health metric related to the level of body mass index (BMI).

metric\_3: A health metric related to the level of physical activity.

metric\_4: A health metric related to the level of total cholesterol in the blood.

metric\_5: A health metric related to the level of systolic blood pressure.

smoking\_status: The smoking status of the individual, recorded as "formerly smoked", "never smoked", "smokes", or "Unknown".

stroke: Whether the individual has had a stroke, recorded as 1 if yes and 0 if no.

Overall, this dataset provides a rich set of information about individuals who have had strokes, which can be used to develop models for predicting stroke risk and improving stroke prevention efforts.