Roche Data Science - Home Study Case

You are working as a Data Scientist and recently you were asked by Digital Transformation Managing Partner Simon to join an innovative project that can actually help saving many lives.

This pharmaceutical company focuses on developing innovative treatments for stroke, which is one of the leading causes of death in modern society. The company treats a significant number of patients who have suffered from strokes, with an average of 11,000 patients treated annually.

To enhance their research and development efforts, the pharmaceutical company has been diligently collecting data on all its stroke patients over the past few years. This extensive dataset includes information such as patient demographics, medical history, treatment protocols, medication effectiveness, and post-treatment outcomes. The primary objective of this use case is to leverage this accumulated data to gain valuable insights into stroke treatment and improve patient outcomes. By analyzing the collected data, the pharmaceutical company aims to identify patterns, correlations, and potential treatment strategies that can enhance the efficacy of their stroke medications and interventions.

The analysis of this dataset can involve various methodologies, including statistical analysis, machine learning algorithms, and data visualization techniques. By applying these approaches, the pharmaceutical company can uncover crucial information about the effectiveness of different treatment protocols, identify risk factors associated with poor outcomes, and discover novel approaches for stroke prevention and management.

Ultimately, the use case seeks to drive evidence-based decision-making within the pharmaceutical company, helping them develop more targeted and effective treatments for stroke patients. The insights gained from the data analysis can guide clinical trials, refine treatment guidelines, and contribute to the advancement of stroke research and development.

Anonymized dataset contains a record of 43401 patients and each of them is described by a following set of variables:

ID – Unique Identification Number
Gender – categorical variable (Male, Female, Other)
Age_In_Days – indicates patient's age in days
Hypertension – binominal variable (1 – patient has hypertension, 0 – patient without hypertension)
Heart_Disease – binominal variable (1 – patient has heart disease, 0 – patient without heart disease)
Ever_Married – binominal variable (Yes – patient is (was ever) married, No – patient has never been married)
Type_Of_Work – categorical variable related to different working status (patient is self-employed, works in a private firm, has a government job, never worked or is still a child)
Residence – binominal variable (Urban – patient currently lives in urban area, Rural – patient currently lives in rural area)
Avg_Glucose – patient's average glucose level for the past 3 months
BMI – patient's current BMI score
Smoking_Status – categorical variable that indicates patient smoking habits
Stroke – binominal target variable (1 – patient after stroke, 0 – patient never had stroke)

Since time is pressing and in this case time means saved lives you were given ***three days to prepare the following:***

- Build a solution in the programming language of your preference (Python or R) that predicts if the patients will suffer from stroke or not. Thanks to that the hospital can identify patients with high risk of stroke and sent them for proper treatment
- Try out at least a few different Machine Learning Algorithms, evaluate them, compare and choose the best model by providing a proper justification
- Create a presentation in (Google Slides, PDF or PowerPoint) where you will walk your main stakeholder - Director of VU University Medical Center Jurian Kuyvenhoven - through your solution and provide a recommendation as well as the next steps that you would take given more time