Job Title:	Data Scientist	**Reports to:**	Head of Data Science-gnuGrid CRB.
Unit:	Data Science	**Department:**	Technology

A brief description of the data:

1. The data consist of 1000 records of loan data. In this dataset, each entry represents a person who takes a credit by a bank. Each person is classified as good or bad credit risks according to the set of attributes.

2. A data dictionary is provided together with the data to help the data scientist understand the data attributes. This is provided under sheet 2 of the excel worksheet.

3. The original data set had both categorical and numerical features. However, the dataset has been transformed to numerical variables. The data dictionary will assist the analyst understand the meaning of the transformed variables.

4. A default is defined as a failure to repay a loan installment for more than 90 days.This is already provided as the Response variable among the variables.

Instructions

1. You have a week to undertake analysis and prepare a Report for the Head of Data science at gnu Grid CRB.
2. You have been provided with data on an Excel Worksheet. Open the file and save it in your name.
3. The report can be provided as a PowerPoint presentation, python notebook or R notebook.
4. **The deadline for submission is Monday 14th August,2023 at 5pm.**

EXERCISE

Perform the following exercises.

1. Data validation is the process of checking that the data is clean, correct and useful. Describe how you would perform data validation on this data.

2. Identify the most predictive variables.
3. Using the data extract provided in excel, build a model that will predict default. Document the steps and any assumptions that you make including any new variables that you create from the variables provided. Additional credit will be given for informative and relevant visuals e.g., tables, graphs, etc. The model required is a simple application scorecard for new customers.