WEEK 9 DELIVERABLES

NAME: VIVIAN KERUBO MOSOMI

Email: kerubomosmi7@gmail.com

Country: Kenya

College: Jomo Kenyatta University of Agriculture and Technology (JKUAT)

Specialization: Data Science

Problem Description

ABC Bank wants to predict whether a particular customer will buy their term deposit product based on their past interactions with the bank or other financial institutions. This will allow the bank to focus its marketing efforts on customers who are more likely to purchase the product.

GitHub Repository Link

https://github.com/Vee2002/DataGlacier_Internship/tree/vc/Data%20Glacier

Data Cleansing and Transformation on the data

i) Renaming the columns
Some columns were poorly named. Example:
- emp.var.rate which I renamed to employee_variation_rate
- cons.price.idx which I renamed to consumer_price_index
- cons.conf.idx which I renamed to consumer_confidence_index
- euribor3m which I renamed to euribor_three_month_rate
- nr.employed renamed to number_of_employees

ii) Assessing values for each column
- Job column had values that were wrongly named. Some had full stops, and hyphens. For example, admin., blue-collar and self-employed.
  I renamed admin. to admin, blue-collar to blue_collar, self-employed to self_employed.
- Education column also had wrongly named categories which had full stops. Including basic.9y, basic.4y, basic.6y, university. degree and professional. course. I renamed this to basic_9y, basic_4y, basic_6y, university_degree and professional_course
- Pdays column which represents the number of days since the customer was last contacted contains a value 999 which is a special value as all other values range between 0 to 30. This value represents people who were never contacted but working with such a value seems illogical. And many records contain the value 999.
  I replaced the value 999 with -1 to represent people who were never contacted cause zero was included and it would have created contradictions.
- Number of employee's column had values like 5191.0, 5195.8, 5176.3 which are illogical. Employee counts are inherently integer values because we cannot have 3.1 employees, so rounding the values to the nearest whole number makes the data more meaningful and interpretable.

Other transformations that I'll perform on the data later include scaling, one-hot encoding to covert categorical variables to numerical cause machine learning models cannot take categorical data, and log or square root transformation for skewing after I've done the Exploratory Data Analysis.