**VIETNAM NATIONAL UNIVERSITY OF HO CHI MINH CITY**

**INTERNATIONAL UNIVERSITY**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

[Predicting-Term-Deposit-Subscriptions](https://github.com/23092003e/Predicting-Term-Deposit-Suscriptions)



|  |  |  |
| --- | --- | --- |
| **Name** | **ID** | **Contribution** |
|  |  |  |
|  |  |  |
| Hoàng Văn Long | ITDSIU21096 |  |

**TABLE OF CONTENTS**

1. **Introduction:**
   1. Overview
   2. Problem and solution
   3. Dataset description
2. **Data Collecting and Manipulation:**
   1. Read Data
   2. Data Manipulation
3. **Exploratory Data Analysis:**
   1. Assess the level of “customer attrition” in the dataset
   2. Variables distribution in customer attrition
4. **Data Preprocessing:**
   1. Variable Summary
   2. Correlation Matrix
   3. Visualizing data with principal components
5. **Introduction:**

**1.1. Overview:**

Dataset:[**Bank Marketing Dataset**](https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset/data)

* **Dataset Overview:** This dataset contains data on a bank's Term Deposit ([**Term Deposit**](https://timo.vn/tai-khoan-tiet-kiem/khi-nao-nen-su-dung-goal-save-va-term-deposit/)) product marketing strategy.
* **Source:** [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

**1.2. Problem and Solution:**

* **Classification Problem:** predict whether a customer will decide to enroll in this program “Deposit”/“No Deposit” (2-class classification problem) through the characteristics of that customer.
* **Overview of the data set analysis plan in this report:**
  + *Data Collection*
  + *Data Manipulation:* feature extraction, normalization
  + *Complementation:* “Null/None/NA” handling
  + *Statistics:* 
    - Statistics on data/label volume
    - Feature distribution, correlation, etc
    - Distribution charts
  + *Cleaning:* Data pre-processing: normalization, splitting the training test set, dropping irrelevant features, etc
  + *Feature extraction*: Encoding features into feature vectors and label vectors for computation
  + *Analysis/forecasting using machine learning algorithms:*
    - Selecting some techniques
    - Evaluating effectiveness
    - Using metrics appropriate to the problem

**1.3. Dataset Description:**

Description of information of some fields in the data set:

* **age**: Represents the age of the customer.
* **job**: Describes the person's occupation.
* **marital**: Indicates the marital status of the person (e.g., married, single, divorced).
* **education**: Represents the person's level of education (e.g., primary, secondary, tertiary).
* **default**: Indicates whether the person has a credit card (‘yes’, ‘no’, or ‘unknown’).
* **housing**: Indicates whether the person has a housing loan (‘yes’, ‘no’, or ‘unknown’).
* **loan**: Indicates whether the person has a personal loan (‘yes’, ‘no’, or ‘unknown’).
* **contact**: Describes the communication method used to contact the person (e.g., ‘cellular’, ‘telephone’).
* **day**: Indicates the day of the week of the last contact.
* **month**: Indicates the month of the last contact.
* **duration**: Duration of the last contact in seconds.
* **campaign**: The number of contacts performed during this campaign.
* **pdays**: Represents the number of days since the last contact with the person, or -1 if they were never contacted before.
* **previous**: The number of contacts performed before this campaign.
* **poutcome**: Describes the outcome of the previous marketing campaign.
* **deposit**: The target variable that indicates whether the person subscribed to a savings account (‘yes’ or ‘no’).

1. **Data Collecting and Manipulation:**

**2.1. Read Data:**

**Dataset:**

pic

**Overall Dataset:**

pic

**Check null and duplicated variables:**

**Pic**

**2.2. Data Manipulation:**

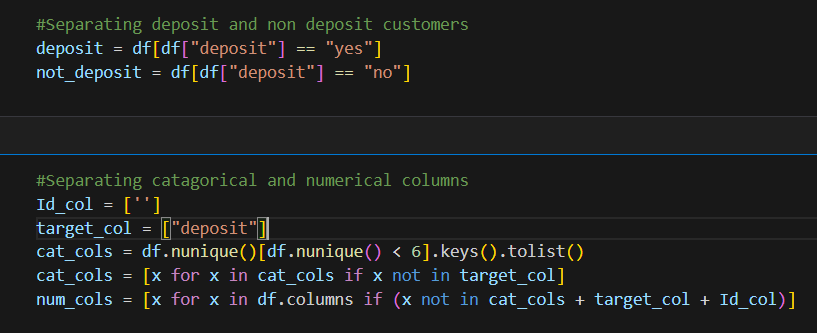
* Count value of “job” attribute:

****

* Remove unknown values and special characters from the “job” column

pic

* Separating:
* Deposit and non-deposit customers
* Categorical and numerical columns



1. **Exploratory Data Analysis (EDA)**

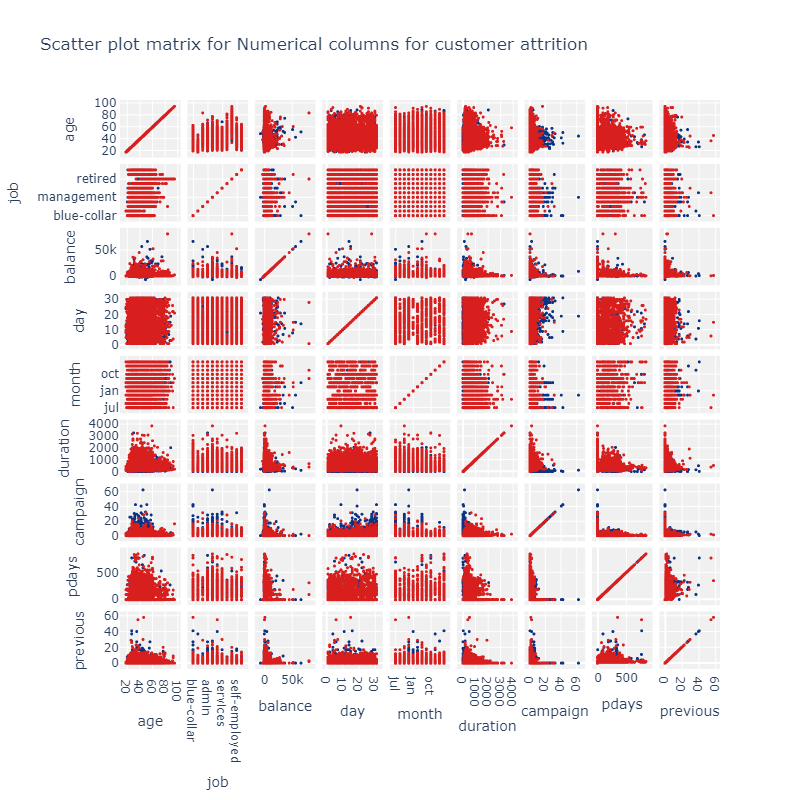
**3.1. Assess the level of “customer attrition” in the dataset**

* What is ‘customer attrition“?

⇒ **Customer attrition is defined as the loss of customers by a business for whatever reason.**

**3.2. Variables distribution in customer attrition**

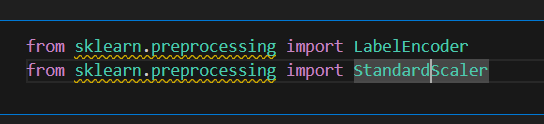
* **Age:**
* **Job:**
* **Balance:**
* **Day:**
* **Duration:**
* **Campaign:**
* **p-day:**
* **Previous:**
* Scatter plot matrix for Numerical columns:



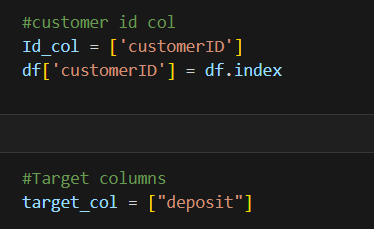
**=> The scatter matrix shows that duration has the strongest correlation with the outcome, while variables like campaign and p-days exhibit no clear relationships.**

1. **Data Preprocessing:**

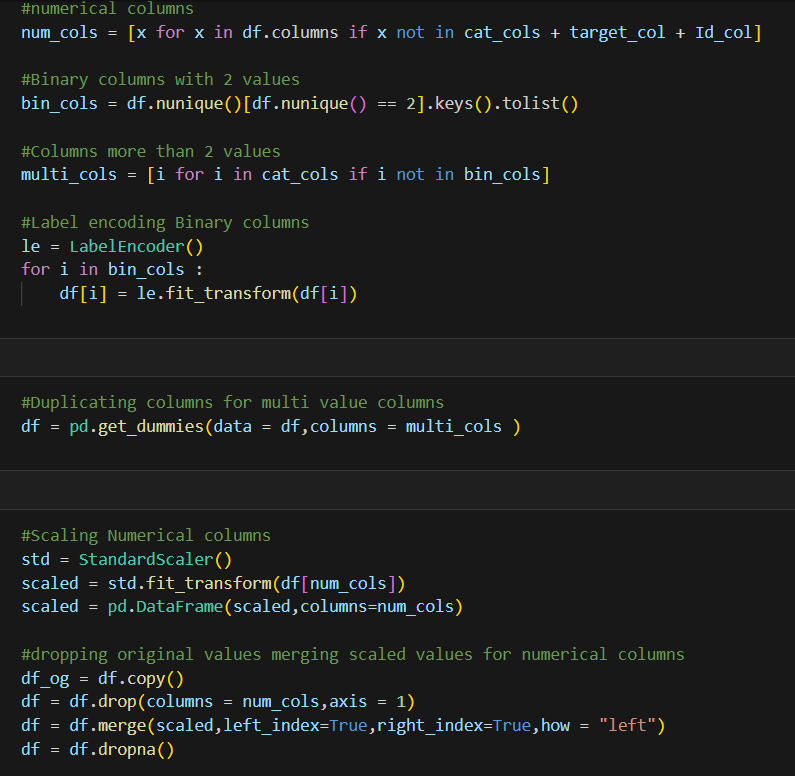
* In this step, I used sk-learn to process data:



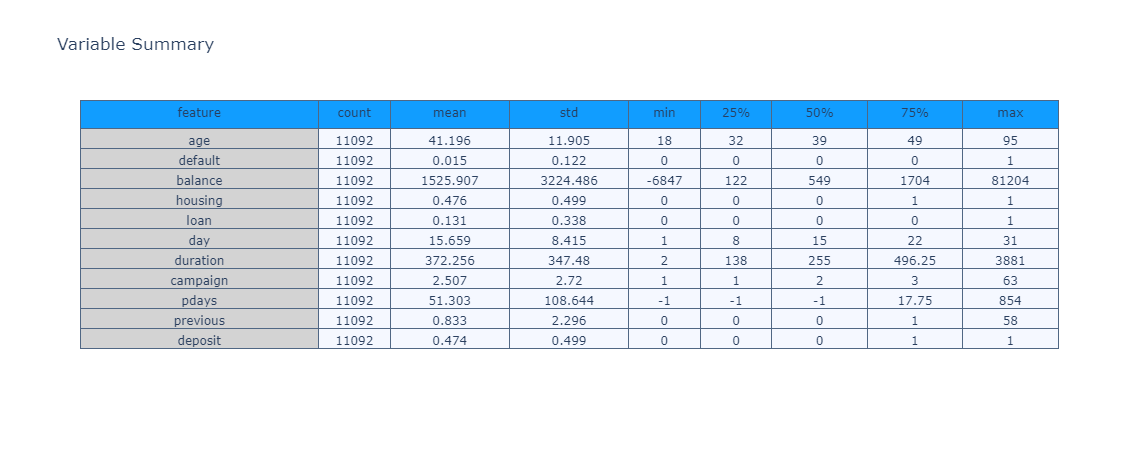
* Customer ID and target column:



* Process Categorical and Numerical columns



**4.1. Variable Summary:**

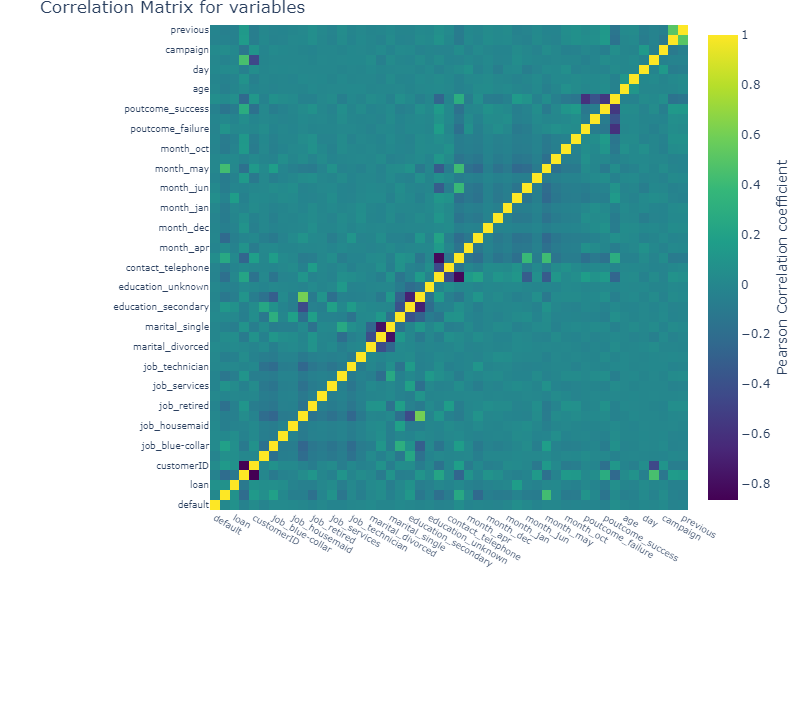
****

The dataset consists of customer attributes and behaviors, showing notable variability in financial balances, call durations, and previous campaign contacts. Key observations include:

* Age: Skewed toward younger adults, with a mean of 41.
* Balances: High variability, ranging from significant debt to large positive balances.
* Call duration (duration): Strong indicator with a wide range (up to 3881 seconds).
* Campaign contacts: Most customers were contacted very few times.
* Outcome (deposit): Balanced between success and failure (~47% subscribed).

⇒ **This suggests that variables like duration, balance, and previous contacts may be critical predictors of the target outcome.**

**4.2. Correlation Matrix:**

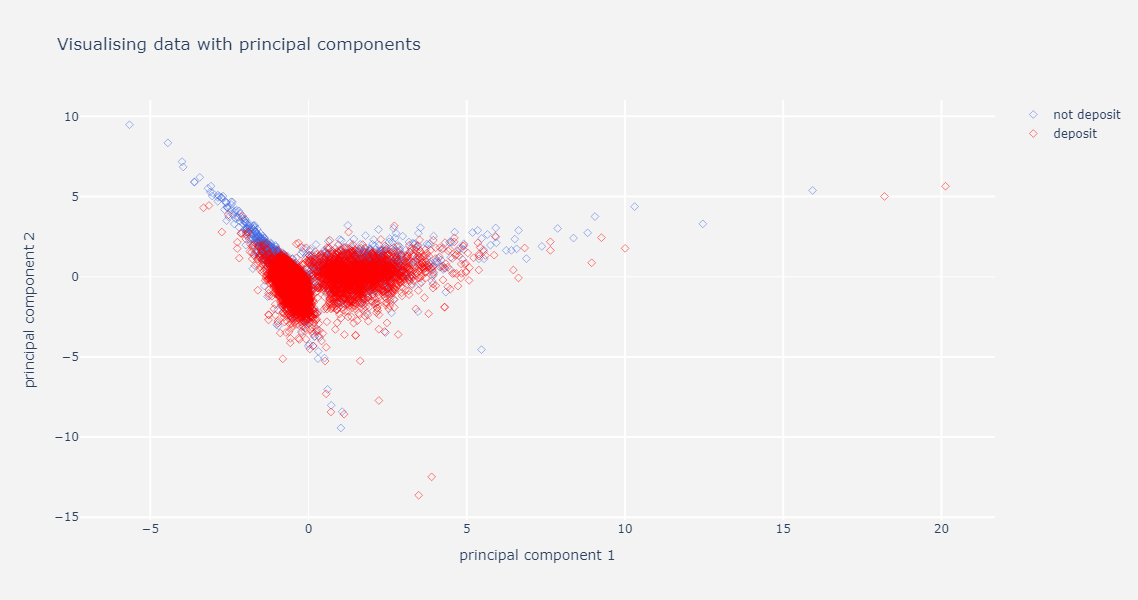


The correlation matrix highlights the relationships between variables:

* **Strongest Correlations**: duration shows a significant positive correlation with the outcome variables (poutcome\_success and deposit), confirming its importance as a predictor.
* **Weak Correlations:** Most variables, such as age, campaign, and pdays, exhibit low correlations with the outcome or with other features, indicating weak linear relationships.
* **Multicollinearity**: Some categorical variables (e.g., marital, education) have moderate correlations within their categories, suggesting overlapping information.
* **Binary Features:** Variables like loan, default, and housing show almost no correlation with the target variable (deposit), implying limited predictive power.

**⇒ In summary, duration remains the most influential variable, while most other features demonstrate low direct correlations with the target outcome.**

**4.3. Visualising data with principal components:**



**=> We can see that the first principal component (PC1) is a strong discriminator between the two categories. It appears that higher values of PC1 are associated with "not deposit" cases, while lower values are associated with "deposit" cases.**