Week 10

Exploratory Data Analysis

Name: Xiyuan Wu

Email: xwu136@gmail.com

Country: US

College/Company: UCR

Specialization: Data Science

Table of Content

1.1 Problem Description

The data is related to a Portuguese banking institution's direct marketing campaigns (phone calls). The classification goal is to predict if the client will subscribe to a term deposit (variable y).

1.2 Business Understanding

The project's main objective is to predict customer subscriptions to time deposits based on a direct marketing campaign conducted by a Portuguese banking institution via telephone. This goal translates into optimizing marketing resources, enhancing customer engagement strategies, and ultimately increasing the effectiveness of these campaigns. By leveraging historical data for a binary classification challenge, the project aims to effectively identify potential subscribers, improving campaign ROI and customer experience through targeted and informed promotion.

1.3 Project Lifecycle

Week	Deadline	Task
Week 7	Apr 19, 2024	Project Preparation, Data Intake Report
Week 8	Apr 26, 2024	Data Processing
Week 9	May 2, 2024	Data Processing (Advanced)
Week 10	May 9, 2024	Data Analysis, EDA
Week 11	May 16, 2024	Build Model Preparation
Week 12	May 23, 2024	Explore Different Model
Week 13	May 29, 2024	Presentation for data result & Model Evaluation, Code

1.4 Data Intake Report

In GitHub Repo.

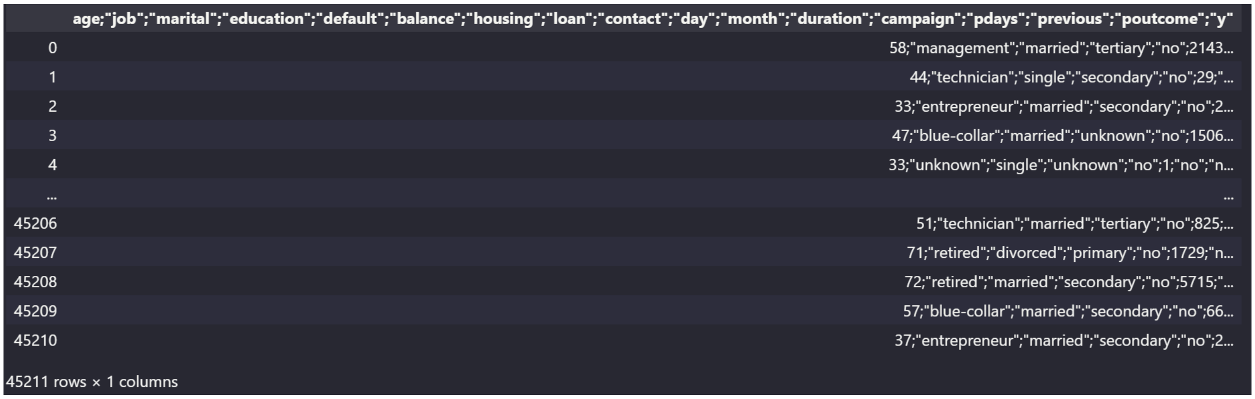
1.5 Data Understanding

1.5.1 Columns

- age: The client's age (numeric).
- job: The type of job (categorical).
- marital: Marital status (categorical).
- education: Education level (categorical).
- default: Has credit in default? (binary: "yes","no").
- balance: Average yearly balance, in euros (numeric).
- housing: Has housing loan? (binary: "yes","no").
- loan: Has personal loan? (binary: "yes","no").
- contact: Contact communication type (categorical).
- day: Last contact day of the month (numeric).
- month: Last contact month of year (categorical).
- duration: Last contact duration, in seconds (numeric).
- campaign: Number of contacts performed during this campaign and for this client (numeric).
- pdays: Number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted).
- previous: Number of contacts performed before this campaign and for this client (numeric).
- poutcome: Outcome of the previous marketing campaign (categorical).
- y: Has the client subscribed to a term deposit? (binary: "yes","no").

1.5.2 Dataset Issue

The dataset appears structured as a single column with semicolon-separated values, suggesting that the CSV reader does not automatically parse it into separate columns.



	age;"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"duration";"campaign";"pdays";"previous";"poutcome";"y"
0	58;"management";"married";"tertiary";"no";2143...
1	44;"technician";"single";"secondary";"no";29;"...
2	33;"entrepreneur";"married";"secondary";"no";2...
3	47;"blue-collar";"married";"unknown";"no";1506...
4	33;"unknown";"single";"unknown";"no";1;"no";"n...
...	...
45206	51;"technician";"married";"tertiary";"no";825;...
45207	71;"retired";"divorced";"primary";"no";1729;"n...
45208	72;"retired";"married";"secondary";"no";5715;"...
45209	57;"blue-collar";"married";"secondary";"no";66...
45210	37;"entrepreneur";"married";"secondary";"no";2...

45211 rows × 1 columns

Our first thing will be to handle them.

1.5.3 Addition Info

Since we have not started the analysis yet, we can't tell if data contain outliers, missing values, etc.

But to handle missing values,

- if we have too many missing values, try to replace them with common values for category value to median for continuous value
- if the missing value only contains a small portion, we can just delete it

For outliers: this depends on different datasets, and we will process this step when we get there.

Overall, this information provides a comprehensive overview of the data that can be used to predict whether a customer will subscribe to a term deposit. These variables include demographic information (e.g. age, job, marital status, education) and campaign-specific data (e.g. contact type, campaign details, previous contacts). The target variable of our predictive model is "y", which represents whether the customer subscribes to a time deposit.

1.6 Exploratory Data Analysis

This week, I've thoroughly analyzed a banking dataset, delving into three primary segments: client demographics, recent campaign interactions, and other relevant attributes. The analysis culminated in a heatmap to examine potential correlations between variables. Despite this extensive exploration, no definitive relationship between pairs of variables emerged that would warrant a solid recommendation at this time. However, the upcoming modeling phase should shed light on the key features influencing a client's decision to subscribe to a term deposit, providing us with actionable insights. And I should be able to offer a final Recommendation in the end.