Phase 1

Exploratory Data Analysis

Learning Objectives

LO1. Get familiar with a given dataset

LO2. Produce useful visualizations

LO3. Practice cleaning and formatting data for a data science task

Scenario

You are a data scientist at a small bank that is trying to get more customers to subscribe to term deposits. In an effort to increase the number of subscriptions, the bank has been running a telemarketing campaign. You have been asked to analyze the data collected so far in the campaign and use it to inform which customers the bank should call next week to maximize the number of subscriptions.

Introduction and Deliverables

In this phase, you will analyze the provided Bank Marketing dataset. The data was collected from direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. The classification goal is to predict if a client will subscribe to a term deposit (this is denoted by a "yes" or a "no" in the last column of the dataset). The bank wants clients to subscribe to their term deposit.

Week	Task	Deliverable
1	Download and clean the dataset, understand the features, produce visualizations of the data, identify key features, construct new features	Jupyter notebook
2	Write-up of project overview and Phase I and review with AI Professional mentor	Design document

Table 1.1: Deliverables in Phase I. Progress review weeks are highlighted.

Tasks

Instructions:

1. `[Ind]` Download the dataset from the AI2C GitHub here. You can clone the entire repo using:
   `$ git clone https://github.com/AFC-AI2C/ai-tech-capstone`
   (a) Read the `README.md` file describing the dataset you just downloaded.
   (b) For this phase, you will be using the `train.csv` data.
2. `[Ind]` Open a Jupyter notebook and use Pandas to read in the dataset. Print the head of the dataset to confirm that is was read in correctly.
   (a) If you need to install Pandas, read this documentation.
3. `[Ind]` Produce summary statistics for the dataset (reference: here)

   (a) Analyze the summary statistics. How many columns are in the summary statistics DataFrame? Why are there fewer columns in the summary statistics DataFrame than in the training dataset?

   (b) What were the ages of the youngest and oldest people called?

   (c) What was the average duration of the calls?

4. `[Ind]` Produce histograms for the `age`, `job`, `marital`, and `education` features (reference: here).

   (a) To start, try to use this Pandas histogram command. This should work for the `age` column, but not for the `job`, `marital`, and `education` columns. Why is this? Figure out another method for producing the histograms for these columns.

   (b) Based on visually inspecting the histograms, what is roughly the ratio of married customers to single customers?

   (c) What is the most common job among the customers?

5. `[Ind]` What percentage of customers subscribed to the term deposit?

6. `[Ind]` Print the record at index 10000.

7. `[Grp]` Answer the following questions in your design document (write the code used to obtain these answers in your Jupyter notebook):

   (a) The labels are the rightmost column of the dataset. What is the data type of the labels (ex. integer, float, boolean)?

   (b) How many records are in the dataset?

   (c) How many features are in the dataset?

   (d) Name 5 features that you think might be most useful in your analysis. What are the data types of the features you identified?

   (e) Are there any missing entries? If so, how are you going to handle the missing entries?

   (f) What is the age, job, marital status, and education level of the record at index 10000?

   (g) Look at the `contact` feature. How many calls were made by telephone? How many calls were made by cellular?

   (h) What month corresponds to the highest success rate for getting clients to subscribe? How did you determine this?

   (i) What day of the week corresponds to the highest success rate for getting clients to subscribe? How did you determine this?

8. `[Ind]` Drop the following columns: `duration`, `contact`, `month`, and `day_of_week`.

9. `[Ind]` How many single people younger than 40 were called? (reference: here)

   (a) Of these customers, how many of them subscribed to the term deposit?

   (b) Does this demographic subscribe to the term deposit more or less than the average for all of the customers?

10. `[Ind]` How many customers are single, older than 40, and have a university degree?

   (a) Of these customers, how many of them subscribed to the term deposit?

   (b) Does this demographic subscribe to the term deposit more or less than the average for all of the customers?

11. `[Ind]` Convert categorical variables to one-hot encodings.

12. `[Grp]` Answer these administration questions in your design document:

   (a) Who is in your team?

   (b) Who is your AI Professional mentor?

   (c) What day of the week is your "capstone day"?

   (d) Where are you going to work on your capstone?

   (e) What skills are you most comfortable with after the AI Technician course?

   (f) What skills are you least comfortable with after the AI Technician course?

13. `[Grp]` Answer these reflection questions in your design document:

   (a) Do you believe you accomplished the learning objectives?

   (b) What tasks were most helpful for accomplishing these learning objectives?

(c) What challenged you during the phase? How did you solve this? If nothing challenged you, what could be added to this phase to make it more interesting/challenging?

14. `[Grp]` Write these sections in your design document:

    (a) Administration - see Question 12

    (b) Introduction - what problem are you working on?

    (c) EDA Findings - what did you learn during the EDA? This is where you will put the answers to the questions listed above. This section should also include figures that support your analysis.

    (d) Phase Reflection - see Question 13