**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Statistical Techniques for DA |
| **Assessment Title:** | CA1 |
| **Lecturer Name:** | Marina Iantorno |
| **Student Full Name:** | Riccardo Possieri |
| **Student Number:** | sba23439 |
| **Assessment Due Date:** | 12th November 2023 at 23:55 hs |
| **Date of Submission:** | 12/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

In this first assignment I am going to talk about the importance of the customer acquisition. As per assignment, I will use the dataset provided by the professor “Bank\_Mkt\_Campaign.csv”. The dataset is related with direct marketing campaigns, and I am going to predict if the client will subscribe a term deposit with the variable “y”.

In bank world, bank term deposit is a popular product seeking to maximize the returns of customers. In this assignment I will understand the factors that influence the decision making of customers because that is the crucial part for banks to effectively target their marketing campaigns. In the following points there is the explanation of each of the five questions of the CA1.

**Exploratory Data Analysis**

One of the most important part of these projects, is understanding the dataset with all its variables. This is the EDA or Exploratory Data Analysis where I proceeded to load as “data\_df”. The dataset includes demographic variables such as age, job, marital status, and education, as well as campaign-related variables such as contact method, previous deposits, and campaign duration. As per question number 1, I computed the descriptive statistics for all the numerical variables with some pandas functions such as:  
  
.head()  
.describe()

I also used the library ProfileReport from ydata\_profiling, which I didn’t add to my code because all the chooses and codes I wrote, were taken without using that HTML file but it was helpful for me at the very beginning due to understand the variables of the data.

**Descriptive Statistics and Central Tendencies and Variability**

Descriptive statistics provided me a summary of the data and helped me to identify patterns and trends. Some of the numerical variables which can be considered relevants are:

- Age: can directly influence the outcome of the campaign, as older people are more likely to subscribe to a term deposit.  
- Balance: can directly influence the outcome of the campaign, as people with higher balance are more likely to subscribe to a term deposit.

From the statistics I could see that the people to which the deposit have been offered are on average 41 indicating that the target audience is middle-aged, and the std wasn’t high meaning most of the data will be close to the mean (30-50). I could also saw that some of the people are younger or older by checking the min 19 and the max 87. The median is 39, which is close to the mean, meaning that the distribution is not skewed. For the balance I could see instead the mean balance, which is 31,784.87, suggesting that the campaign is targeting customers with moderate to high financial resources. Also, I could see a very wide spreader distribution since its more likely to have very few rich people and a lot of people with low balance.

All the analysis done up in the code could be finished with some plots analysing the distribution of the variables. (question 2)

**Descriptive statistics for categorical Variables**

Some of the most important categorical variables were:

- Job: could directly influence the outcome of the campaign, as some jobs are more likely to subscribe to a term deposit.  
- Marital: could directly influence the outcome of the campaign, as some marital status are more likely to subscribe to a term deposit.  
- Education: could directly influence the outcome of the campaign, as some education level are more likely to subscribe to a term deposit.

At this point I calculated some frequencies to answer to the first and third questions: the most frequent job titles in the dataset are management and blue-collar. Over 60% of customers were married, suggesting that marital status may be a relevant factor in campaign success. The most common education level is secondary, followed by tertiary and primary education.

**Visualizations**

In the second question was required to create two visualisations. They help us explore the data in a more intuitive way. Histograms and boxplots were used to visualize the distribution of numerical variables like age and balance. Histograms revealed that age is not skewed, while balance is highly right-skewed, with a long tail towards higher values. Boxplots showed that there are a few outliers for both age and balance.

**Relationships between Variables**

To understand the relationships between variables, I conducted cross-tabulations and correlation analyses in the question 3. Cross-tabulations revealed that certain job titles, marital statuses, and education levels were associated with higher subscription rates for term deposits. Correlation analyses showed that age has a weak positive correlation with balance, indicating that older customers tend to have higher balances.

**Binomial Distribution for Probabilities**

The binomial distribution is a discrete probability distribution that describes the number of successes in a sequence of independent trials. In order to answer to the forth question I took a subset of the data as taken 50% of the rows randomly selected. I computed the probability of success for the subset of the data with two new functions:

“no\_yes”  
“no\_no”

Consequently, I did the formula that could be interpreted as the probability of getting a "yes" outcome given that the outcome either "no" or "yes”. Since the outcome was binary, I used the binomial distribution to calculate the probability of success (i.e. the probability of a customer subscribing to the term deposit) for a given number of trials. So, I could derive meaningful statistics for the campaign. I checked some examples such as the probability of getting exactly 1 conversion in 10 trials and the probability of getting more or equal to 5 conversions in 10 trials.

**Normal Distribution for Age Variable**

The normal distribution is a continuous probability distribution that describes bell-shaped data. I have found that the age variable in the dataset is approximately normally distributed, meaning that most customers are middle-aged, as I described previously. In order to interpret the variable age, I used a cumulative distribution of a normal distribution, doing:

norm.cdf(70, mean, std): This gave me the cumulative probability that the random variable was less than 70.  
norm.cdf(60, mean, std): This gave me the cumulative probability that the random variable was less than 60.

Subtracting these two values gave me the probability that the random variable was between 60 and 70. In other words, it calculated the probability that the age was greater than 60 but less than 70 with a result of 0.034298987805852055. I finally plotted the distribution of the variable age and checked if it was normally distributed. The result was somewhat normally distributed, but not perfectly. Using a normal distribution to model the age variable could still be a good approximation.

**Conclusion**

In this project I understood the importance of statistical analysis and for my very first time I used tools to visualise and calculate probabilities. Investigating trends and relationships has been an important research tool to draw valid conclusions. Since the very first step, I specified my hypotheses and made decisions such as chose the most important features for me to do an appropriate job research. After collecting all the data from the dataset, I could organize the data with descriptive statistics. In quantitative research, the first step of statistical analysis is to describe each variable with important parameters like mean, median, or a relation between more variables.

Is important to understand the 3 different points of descriptive statistics:

1. Distribution
2. Measures of central tendency (mean, mode, median)
3. Measures of variability (standard deviation, variance, range, interquartile range)

Age and Balance have been chosen as the most potential relevant numerical variables for their impact on campaign outcomes. By knowing what factors affect customer decisions, banks can make campaigns that suit different customer groups, send personalized marketing messages, and use predictive modelling to find potential subscribers.

*Key Findings*

Another important step of this project has been to understand and decide how exactly I would have measured my chosen variables. I understood what kind of data they contained and divided them into nominal and ordinal for all those categorical data, and interval scale or ratio scale for all those quantitative data.   
Demographic factors: Job title, marital status, and education level are all significantly related to the subscribing to a term deposit.   
Probability of success: To model the probability of success I used the binomial distribution for a given number of trials.   
Normal distribution: The age variable in the dataset is approximately normally distributed which it shows that most customers are middle-aged.

*Visualization*

To complement the descriptive statistics, I created a visualization of relevant numeric variables: age and balance. The histograms and bloxplot showed the distribution patterns, follow a relatively normal distribution with few outliers, while balance showed a right skewed distribution with a significant number of outliers. In the third question I used bar plots which were generated to illustrate the relative frequencies of positive and negative outcomes for each category within these variables. These visualizations provided a clear comparison of how different categories influence the success or failure of the campaign, offering insights for targeted marketing efforts.

*Binomial distribution and Normal distribution*

In the fourth and fifth questions I randomly selected half of the customer data and created a hypothetical scenario that reflects a real-life marketing campaign. I simulated various outcomes using binomial distribution and finally, I looked at how old the people in the data were. I used a normal distribution to figure out how likely it was that someone in the data would be in a certain age group, like 60 to 70 years old or younger than 50 years old.

Each step, from understanding the dataset (data\_df) to showing data spread and calculating probabilities, helps us gain a complete understanding of the customer data. These insights facilitate us to create custom strategies and improve campaign performance based on a deep understanding of the data patterns. As the banking industry becomes increasingly data-driven, such analyses play a critical role in ensuring the effectiveness and efficiency of marketing initiatives.

**References:**

https://www.kaggle.com/code/aleksandradeis/bank-marketing-analysis

https://seaborn.pydata.org/generated/seaborn.FacetGrid.html

Müller, A.C. and Guido, S. (2018) Introduction to machine learning with python: A guide for data scientists. Sebastopol: O’Reilly Media.

McKinney, W. (2022) Python for Data Analysis: Data wrangling with pandas, NumPy, and Jupyter. Beijing: O’Reilly.

CCT College Dublin: Available at: https://moodle.cct.ie

Riccardo