WEEK 10: Exploratory Data Analysis

Name: Sainad Reddy Naini (Individual)
Email: nainisainad@gmail.com
Country: India
College: University of Bath (Graduate -UK)
Specialization: Data Science

1: PROBLEM DESCRIPTION

ABC Bank is preparing to launch a new term deposit product and aims to maximize the effectiveness of its marketing efforts. To achieve this, they want to develop a machine learning (ML) model that can predict whether a customer will subscribe to the term deposit based on historical data from previous marketing campaigns. This predictive model will help the bank identify potential customers who are more likely to purchase the product, allowing them to focus their marketing efforts more effectively and efficiently.

Note: Refer to the "Data Cleansing and Transformation" PDF and the previous week's materials (Week 9) for a rough idea of the data problems and how we addressed them. In this section, we will perform exploratory data analysis (EDA) to identify patterns and insights.

2: GITHUB REPO LINK

https://github.com/sainadreddy/Data_Glacier_Intership_2024

3: EDA PERFORMED ON THE DATA

We completed some exploratory data analysis (EDA) tasks last week, including filling null values, treating outliers, and handling duplicates during data cleaning and transformation. Now, we will delve into univariate and bivariate analysis of both numeric and categorical variables. Rather than including all the screenshots, you can refer to the code or await the presentation I'll prepare and upload next week. For now, let's review an overview of the analysis performed.

3.1: Categorical Features

I initially focused on categorical features and conducted univariate analysis using count plots, leading to several key insights. The dataset contains 10 categorical features, excluding the target variable. The majority of individuals hold administrative positions, with blue-collar workers and technicians being the next largest groups, while students are relatively few. Most people are married and literate. The data collection is skewed, with more records from May and fewer from December. Regarding financial features, there are more individuals without a house loan compared to those without a personal loan. Additionally, the "default" feature shows a significant imbalance, with many more "no" cases than "yes," which led to the decision to drop this feature from further analysis.

Subsequently, I conducted bivariate analysis with respect to the target variable and observed several patterns. Despite many administrative professionals, retirees display a higher interest in term deposits. Single individuals are more likely to subscribe to term deposits compared to married ones. Education level influences subscription rates: clients with less education, such as illiterates, are more inclined to sign up, followed by those with university degrees.

Financially, people with personal loans are less likely to subscribe to term deposits compared to those with house loans, though the presence of any loan does not significantly impact subscription rates. Contact methods also affect subscription likelihood; clients contacted via cell phone are more likely to sign up than those reached by telephone. Seasonal trends indicate increased interest in term deposits in December, March, October, and September, with the highest subscription rates on Thursdays and Tuesdays. Finally, a successful outcome from previous marketing efforts correlates with a higher likelihood of term deposits.

3.2: Numerical Features

Next, I analysed the numeric features using univariate analysis. The dataset includes 9 numeric features, excluding the target variable. While the age feature is slightly right-skewed, it is approximately normally distributed. Other numeric features display varying degrees of skewness, either right or left, which will be addressed if they impact performance, as previously discussed. Outliers have been managed, and necessary data cleaning has been performed as detailed earlier.

For bivariate analysis, I first examined a heatmap of correlations among all variables and found that none of the variables exhibit a strong positive or negative relationship with each other. I then explored the features individually with respect to the target variable and made several observations. Specifically, an increase in the number of contacts does not correlate with a higher number of term deposits, and most clients were not previously contacted. Additionally, individuals aged between 30 and 64 are less likely to subscribe to term deposits, despite being the most frequently contacted group.

4: FINAL RECOMMENDATION

Based on the EDA, several recommendations emerge: focus on retirees and single individuals for term deposit campaigns, prioritize cell phone outreach, and target peak periods in December, March, October, and September, with higher engagement on Thursdays and Tuesdays. Address numeric feature skewness as needed. Additionally, perform a more in-depth analysis to develop an effective machine learning model for predicting client subscriptions to term deposits.