**Data Glacier Project**

**Data Science: Bank Marketing (Campaign)**

# **Team members:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Group** | **Names** | **E-mail** | **Country** | **College** |
| DataWizz | Akhil Abraham | akhilabrahamuk@gmail.com | UK | Queen’s University Belfast |
| Pravallika Sheshabhatter | sbpravallika3@gmail.com | UK | University of Hertfordshire |
| Shiva Ramezani | shvramezani@gmail.com | US | Cal State Northridge |

**Problem Description:**

The problem at hand is that ABC Bank wants to sell its term deposit product to customers. However, before launching the product, they want to develop a model that can predict whether a particular customer is likely to buy the product based on their past interactions with the bank or other financial institutions. The goal is to assist the bank's marketing efforts by targeting customers who are more likely to purchase the product. By focusing their marketing channels, such as telemarketing, SMS, or email marketing, on these potential customers, the bank can save resources, time, and costs associated with marketing to customers who are less likely to subscribe to the term deposit.

**Data understanding &What type of data you have got for analysis:**

The data provided for analysis consists of information related to direct marketing campaigns of a Portuguese banking institution. The dataset contains both input variables and the desired target variable.

The input variables include demographic information about bank clients (such as age, job, marital status, education), financial information (credit default, housing loan, personal loan), details about the last contact made during the campaign (communication type, month, day of the week, duration), as well as other attributes related to the campaign and social/economic context.

The desired target variable is denoted as 'y' and represents whether or not the client subscribed to a term deposit ('yes' or 'no').

**What are the problems in the data (number of NA values, outliers, skewed, etc.)**

In data preparation, several issues need to be addressed before developing the model. These include missing values, outliers, skewed data, and imbalanced datasets. Missing values can impact the analysis and modeling process, while outliers can disrupt the statistical properties of the dataset and influence the model's performance. Skewed data, with uneven distribution, can affect the model's assumptions and predictions. Additionally, imbalanced datasets, where one class significantly outnumbers the other, can lead to biased models favoring the majority class. Techniques like oversampling, under-sampling, or synthetic data generation can be used to balance the dataset.

Data quality checks were conducted on the bank-full and additional-full datasets, confirming the null values' absence. The data imbalance was identified, requiring corrective techniques to address the bias. Outliers were detected using the z-score method and box plots, while skewness was assessed through distribution plots. These analyses provided valuable insights for subsequent data preprocessing and model development.

**What approaches you are trying to apply to your data set to overcome problems like NA value, outlier, etc., and why?**

To handle null values in a dataset, common approaches include removal (excluding null-containing rows or columns), imputation (replacing null values with estimated or predicted values), forward/backward fill (using the last or next observed value), etc. However, there were no NULL values in both big datasets.

While examining the data, it was observed that certain variables contained missing values labeled as "Unknown" and "Nonexistent." Further analysis will be necessary to determine how to address these missing values. Furthermore, it was observed that outliers exist in the "age" and "duration" variables. We could categorize the “**age**” variables, in order to derive valuable insights and it allows for a more structured analysis.

To treat skewness in data, options include logarithmic or square root transformations, power transformations like Box-Cox, binning, and data normalization. The only feature that might need to treat its skewness is Age.

To remove outliers in a dataset, follow these steps: identify variables of interest, calculate statistical measures, set a threshold for outliers, identify and handle data points outside the threshold, repeat for each variable, and consider the impact on the dataset and analysis. We will be removing the outlier as the number of outliers was approximately 2%.

To address imbalanced datasets, techniques include oversampling, under-sampling, class weighting, ensemble methods, anomaly detection, cost-sensitive learning, and data augmentation. Each method aims to balance class distribution and improve model performance. In this case, we will over-sample the data because if we try the under-sampling, the model wouldn’t act well.