TEAM MEMBER'S DETAILS

Group Name: Banking Insights Squad
Group: Members: Canmert Demir & Joseph Pang
Names: Canmert Demir-Bank Marketing (Campaign) -- Group Project
Email: canmertdemir2@gmail.com
Country: Turkey
College/Company: Msc Bartin University-Applied Mathematics / Data Glacier
Specialization: Data Science
Name: Joseph Pang - Bank Marketing (Campaign) -- Group Project
Email: joseph302156@gmail.com
Country: United States
College/Company: University of California, Berkeley/ Data Glacier
Specialization: Data Science
Github Repository: https://github.com/Canmertdemir/DataGlacierWeek8

PROBLEM DEFINITION

ABC Bank, a prominent financial institution, aims to introduce a novel fixed-term deposit product tailored to potential clients seeking secure and profitable investment opportunities. Leveraging its extensive reservoir of customer data and historical banking interactions, the bank is embarking on a strategic initiative to delve into intricate patterns within customers' past banking behaviors using data analytics and machine learning techniques.

This comprehensive approach involves analyzing diverse financial interactions such as deposit patterns, loan histories, investment inclinations, and overall financial preferences. The primary goal is not solely the introduction of a new financial product but rather personalized engagement with potential clients.

By developing a sophisticated machine learning model derived from substantial historical data, the bank intends to design precise marketing strategies aligned with specific customer segments

identified through comprehensive data analysis. The primary objectives are to introduce the fixed-term deposit product in alignment with customer preferences and to optimize marketing efforts.

By targeting individuals inclined towards secure and profitable investments, ABC Bank aims not only to increase sales but also to enhance customer satisfaction through customized financial offerings. This initiative reflects the bank's dedication to promoting financial literacy and delivering innovative products tailored to meet diverse investment needs.

With a strong emphasis on customer-centricity and data-driven decision-making, ABC Bank aims to maintain its legacy of providing financial solutions tailored to individual aspirations and financial objectives.

DATA UNDERSTANDING

The dataset comprises four main components: `bank`, `bank_additional`, `bank_full`, and `bank_additional_full`. These datasets differ in features and observations. The primary focus is analyzing and modeling the `bank_additional` dataset. Initially, we aim to comprehend the dataset and acknowledge its imbalance, recognizing its potential impact on subsequent analytical steps.

Imbalanced datasets present significant challenges for machine learning models. When one class is notably less represented than others in the target variable, models trained on such data often show biases towards the more prevalent class. This leads to poor predictions for the less represented class and impacts the model's ability to generalize effectively. Metrics like accuracy become unreliable as models might achieve high accuracy by favoring the majority class, neglecting valuable insights from the minority class.

To mitigate these challenges, various strategies can be applied. Techniques like oversampling the minority class (e.g., using SMOTE) or undersampling the majority class can balance the dataset. Choosing alternative metrics such as F1-score, precision-recall curve, or AUC-ROC can offer better insights into model performance. Preferential selection of algorithms known for handling imbalanced data, like Random Forests or Gradient Boosting, along with employing

cost-sensitive learning or generating synthetic samples for the minority class, are effective strategies. By leveraging these approaches and staying mindful of imbalance, machine learning models can better navigate imbalanced datasets and provide more equitable predictions across all classes in the data.

Data Exploration and Preprocessing:

Begin by reviewing each dataset's shape, statistical summary, variable types, and initial and final observations using the `quick_look()` function.

- Manage Imbalance: Acknowledge the imbalance in the dataset and address it appropriately.
- Target Variable: Convert the `y` variable into boolean values suitable for modeling purposes.

Correlation Analysis:

Identify correlations among numerical features using the `cor_analiz_cardinals()` function.

- Take note of significant correlations, particularly those associated with the target variable (`y_bool`).

Variable Categorization:

Use the `grab_col_names()` function to categorize variables into different types:

- Categorical Variables: Encompassing both categorical and numerical categorical variables.
- Numerical Variables: Pure numerical variables excluding categorical representation.
- Categorical but Cardinal Variables: Categorical variables exhibiting high cardinality.
- Numerical but Categorical Variables: Numerical variables that behave similarly to categorical ones.

Analysis of Variable Distributions and Missing Values:

This analysis focuses on studying the attributes "**cons.conf.idx**", "**cons.price.idx**" ,"**emp.var.rate**", "**euribor3m**", "**nr.employed**", "**duration**", "**pdays**", "**previous**", and "**y_bool**."

Characteristics of Distribution:

1. **cons.conf.idx:** Displays a distribution skewed to the right.
2. **cons.price.id:** Indicates a distribution skewed to the left.
3. **emp.var.rate:** Shows a distribution skewed to the left.
4. **euribor3m:** Exhibits a distribution skewed to the left.
5. **nr.employed:** Presents a distribution skewed to the left.

Missing Data:

One missing value is observed in the "**duration**" variable.

"**pdays**" has two missing values.

The "**previous**" variable contains 3523 missing values.

"**y_bool**" records 3668 missing values.

Proper consideration of the distribution attributes and handling missing data using suitable methodologies could significantly impact accurate outcomes during analysis and modeling procedures.

Report on Variable Distribution and Imputation:

The assessment indicates that several variables exhibit skewed distributions, necessitating the replacement of zero rows within these variables with their respective medians. Given the dataset's relatively modest size, a replacement method has been chosen for its feasibility. This approach is specifically tailored to accommodate the dataset's dimensions. Replacing the zero values with medians is intended to rectify the skewed distribution observed within these variables effectively.

Outlier Analysis Report

The examination conducted on the variables "**cons.conf.idx**", "**cons.price.idx**", "**emp.var.rate**" ,"**euribor3m**", and "**nr.employed**" revealed no outliers beyond the predetermined threshold

values. This finding denotes that these variables exhibit behavior consistent with the analysis, displaying no significantly divergent patterns from other observations. Furthermore, dataset does not contain NA values.

The absence of outliers in these variables is particularly valuable as outliers can introduce misleading aspects in both analysis and modeling processes. Based on this analysis, it's evident that these five variables align within the expected boundaries. This substantiates the overall consistency of these features within your dataset.

This outcome underscores the reliability of the dataset in terms of these specific attributes, indicating that the data is robust and aligns with anticipated behaviors.