Data Glacier Project
Data Science: Bank Marketing (Campaign)

Team members:

Group	Names	E-mail	Country
DataWizz	Akhil Abraham	akhilabrahamuk@gmail.com	UK
	Pravallika Sheshabhatter	sbpravallika3@gmail.com	UK
	Shiva Ramezani	shvramezani@gmail.com	US

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 have a higher probability of purchasing 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.

Business Understanding:

The business objective of this project is to increase the effectiveness and efficiency of the bank's marketing campaigns for its term deposit product. By utilizing a machine learning model, the bank aims to identify customers who are more likely to subscribe to the product, enabling targeted marketing efforts. The project seeks to optimize the allocation of marketing resources and improve the overall success rate of the marketing campaigns.

Project Lifecycle and Deadline:

The project lifecycle typically consists of the following phases:

- Project Planning: This phase involves understanding the project requirements, defining goals and objectives, identifying stakeholders, allocating resources, and establishing a project timeline. The deadline for completing this phase depends on the complexity of the project but should be completed within a reasonable timeframe prior to July 30.

- Data Collection and Preparation: In this phase, the required data for training the machine learning model is collected. The data may be obtained from the bank's internal databases or external sources. Data cleaning, preprocessing, and feature engineering are performed to prepare the dataset for analysis. The duration for this phase depends on the size and complexity of the dataset and the quality of the initial data.

- Exploratory Data Analysis: This phase involves exploring the dataset to gain insights, understand the relationships between variables, and identify patterns. It includes visualizations, statistical analysis, and data profiling. The duration for this phase depends on the complexity of the dataset and the depth of analysis required.

- Model Building: In this phase, various machine learning models, such as logistic regression, ensemble methods, or boosting algorithms, are developed using the prepared dataset. The models are trained, evaluated, and optimized to achieve the best performance. The duration for this phase depends on the complexity of the models being explored and the amount of experimentation required.

- Model Selection and Performance Reporting: Once multiple models have been built, they are evaluated and compared using appropriate evaluation metrics such as accuracy, precision, recall, or F1-score. The best-performing model(s) are selected for further analysis. Performance reporting involves generating comprehensive reports on model performance, including metrics and visualizations. The duration for this phase depends on the number of models evaluated and the complexity of the performance reporting.

- Model Deployment: In this phase, the selected model is deployed into a production environment where it can be used to make predictions on new, unseen data. The model is integrated into the bank's systems, and necessary considerations for scalability, reliability, and security are addressed. The duration of this phase depends on the complexity of the deployment process and any infrastructure requirements.

- Converting ML Metrics into Business Metrics and Result Explanation: Once the model is deployed, the focus shifts toward analyzing the model's predictions and converting the machine learning metrics into meaningful business metrics. This involves understanding the impact of the model on the bank's marketing campaigns, assessing the return on investment, and explaining the results to the business stakeholders in a clear and understandable manner. The duration for this phase depends on the depth of analysis and the complexity of translating ML metrics into business metrics.

- Presentation Preparation: This final phase involves preparing a presentation that summarizes the project, its findings, the developed model, and its impact on the business.

Phase	**Stage**	**Deadline**
Phase 1	Data Understanding	26/06/2023
Phase 2	Exploratory Data Analysis and Data Preparation	02/06/2023
Phase 3	Data Modelling and Evaluation	09/07/2023
Phase 4	Model Selection	16/07/2023
Phase 5	Model Deployment	23/07/2023
Phase 6	Presenting and Reporting	30/07/2023

Data Intake Report

Name: Bank Marketing (Campaign)
Report date: 06/19/2023
Internship Batch: LISUM21
Version: 1
Group Name: DataWizz
Data intake by: Shiva Ramezani, Akhil Abraham, Pravallika Sheshabhatter
Data intake reviewer: N/A
Data storage location: https://archive.ics.uci.edu/dataset/222/bank+marketing
GitHub Link: https://github.com/ShivaRamezani/Capstone

Tabular data details: bank.csv

Total number of observations	4522
Total number of files	N/A
Total number of features	17
Base format of the file	.csv
Size of the data	453 KB

Tabular data details: bank-full.csv

Total number of observations	45212
Total number of files	N/A
Total number of features	17
Base format of the file	.csv
Size of the data	4.5 MB

Tabular data details: bank-additional.csv

Total number of observations	4120
Total number of files	N/A
Total number of features	21
Base format of the file	.csv
Size of the data	571 KB

Tabular data details: bank-additional-full.csv

Total number of observations	41189
Total number of files	N/A
Total number of features	21
Base format of the file	.csv
Size of the data	5.6 MB

Proposed Approach:

- **Mention the approach of duduk validation (identification)**

Identify Key Fields: Determine the key fields that uniquely identify a customer. In this dataset, it could be a combination of attributes like name, contact number, or any other relevant information.

Data Sorting: Sort the dataset based on the key fields identified in the previous step. This step ensures that similar records are placed together.

Record Comparison: Compare consecutive records based on the key fields to identify potential duplicates. If the key fields match, it indicates a potential duplicate record.

Duplicate Identification: Flag or mark the potential duplicate records for further analysis or processing.

Data Quality Analysis: Perform an in-depth analysis of the potential duplicates to determine their validity. Some common techniques for deduplication include:

a. Data Sampling: Randomly select a subset of potential duplicates and manually review them to confirm if they are indeed duplicates.

b. Automated Methods: Utilize automated techniques like fuzzy matching, string similarity algorithms, or record linkage algorithms to compare and match potential duplicates.

c. Domain Knowledge: Leverage domain knowledge and business rules to identify duplicate records based on specific criteria.

Duplicate Handling: Decide on the appropriate action to handle the duplicates. Options include removing duplicates, merging duplicate records, or keeping only the most recent or most complete record.

- **Mention your assumptions (if you assume any other thing for data quality analysis**

Missing Values: Assume that missing values exist in the dataset and devise strategies to handle them appropriately. This can involve techniques like imputation, deletion, or treating missing values as a separate category.

Outliers: Assume the presence of outliers in numerical variables and determine their impact on the analysis. Decide whether to remove outliers or transform variables to mitigate their influence.

Data Consistency: Assume that inconsistencies might exist within the dataset, such as conflicting values or data format discrepancies. Address such inconsistencies through data cleansing and standardization techniques.

Data Integrity: Assume that the dataset is reliable and represents accurate information. If there are concerns about data integrity, explore methods to verify the data's accuracy and rectify any inconsistencies.

Data Balance: Consider the possibility of class imbalance in the target variable ('y') and evaluate techniques to handle it during model building, such as oversampling, undersampling, or generating synthetic samples.