

## Week 8

### Team Member's Details:

- **Group Name:** Data Science Bank Marketers
- **Members:**
  - **Amr Hacoglu** – [amr.hacoglu@gmail.com](mailto:amr.hacoglu@gmail.com) - Turkey - University of Karabuk - Data Science
  - **Ha My Pham** – [mpham25@wooster.edu](mailto:mpham25@wooster.edu) - US – College of Wooster – Data Science

**Problem Description:** ABC Bank aims to develop a machine learning model to predict whether a customer will subscribe to a term deposit product. This model will help the bank focus its marketing efforts on customers with a higher likelihood of purchasing the product, thereby optimizing resource allocation and reducing marketing costs.

### Data Understanding:

#### Type of Data:

- **Format:** PDF file
- **Number of Records:** 41,188
- **Number of Features:** 21

#### Features:

1. **age:** Numeric
2. **job:** Categorical (e.g., 'admin.', 'blue-collar', etc.)
3. **marital:** Categorical (e.g., 'divorced', 'married', etc.)
4. **education:** Categorical (e.g., 'basic.4y', 'high.school', etc.)
5. **default:** Categorical ('no', 'yes', 'unknown')
6. **housing:** Categorical ('no', 'yes', 'unknown')
7. **loan:** Categorical ('no', 'yes', 'unknown')
8. **contact:** Categorical ('cellular', 'telephone')
9. **month:** Categorical (e.g., 'jan', 'feb', etc.)
10. **day\_of\_week:** Categorical (e.g., 'mon', 'tue', etc.)
11. **duration:** Numeric (duration of the last contact)
12. **campaign:** Numeric (number of contacts during the campaign)
13. **pdays:** Numeric (days since last contact; 999 if not previously contacted)

- 14. **previous:** Numeric (number of contacts before the campaign)
- 15. **poutcome:** Categorical ('failure', 'nonexistent', 'success')
- 16. **emp.var.rate:** Numeric (employment variation rate)
- 17. **cons.price.idx:** Numeric (consumer price index)
- 18. **cons.conf.idx:** Numeric (consumer confidence index)
- 19. **euribor3m:** Numeric (euribor 3-month rate)
- 20. **nr.employed:** Numeric (number of employees)
- 21. **y:** Binary target ('yes', 'no')

### Problems in the Data:

- **Missing Values:**
  - 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'poutcome'
  - Strategy: Impute missing values using the most frequent category or 'unknown'.
- **Outliers:**
  - Numeric features like 'age', 'duration', 'campaign', 'pdays', 'previous', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed'
  - Strategy: Identify and handle using IQR method or capping.
- **Imbalance:**
  - The target variable 'y' is imbalanced.
  - Strategy: Use SMOTE, class weighting, or undersampling techniques.

### Approaches to Overcome Data Problems:

- **Handling Missing Values:**
  - **Imputation:** Fill missing categorical values with the most frequent value or 'unknown'.
  - **Rationale:** Maintains data integrity without biasing the dataset.
- **Handling Outliers:**
  - **IQR Method:** Remove outliers beyond  $1.5 \times \text{IQR}$ .
  - **Capping:** Cap extreme values to a specified percentile.
  - **Rationale:** Reduces noise and improves model robustness.
- **Handling Imbalanced Data:**
  - **SMOTE:** Synthetic Minority Over-sampling Technique to balance class distribution.
  - **Class Weighting:** Adjust model to give more importance to minority class.
  - **Undersampling:** Randomly reduce majority class samples.

- **Rationale:** Ensures balanced learning and better model performance.

**GitHub Repo Link:** [Data Glacier Internship Repository](#)