# **Bank Marketing Campaign – Predicting "Yes" Responses**

## **Project Overview**

This project analyzes the Bank Marketing dataset to predict which customers will respond "yes" to a direct marketing campaign for term deposits. In real-world marketing, not all eligible customers will respond positively, resulting in a highly imbalanced dataset where the majority of responses are "no" and only a small fraction are "yes".

## **Business Context**

The campaign's objective is to maximize the number of true "yes" responses (i.e., customers who actually subscribe). In this context, recall for the "yes" class is the most important metric:

* Recall ("yes") = Of all actual "yes" customers, how many did the model successfully identify?
* High recall helps ensure that as many potential subscribers as possible are targeted, even if this means accepting more false positives (lower precision).

## **Data**

The dataset (bank.csv) contains customer demographic, financial, and campaign-related information, including engineered features such as:

* log\_pdays, log\_campaign, log\_previous (log-transformed for normalization)
* Job and education dummies
* Relationship and loan status indicators

## **Modeling Approach**

Several machine learning models were trained and evaluated:

* Logistic Regression (with and without SMOTE)
* Decision Tree
* Support Vector Machine (SVM)
* K-Nearest Neighbors (KNN)
* All models were evaluated using stratified train/test splits and feature scaling where appropriate.
* SMOTE was used to address class imbalance, especially for models that do not support class weights.

## 

## **Evaluation Metrics**

* Recall ("yes"): Main metric for model selection (maximize true positives).
* Precision ("yes"): Secondary metric; lower precision is tolerated in favor of higher recall.
* F1-score ("yes"): Balance between recall and precision.
* Computation Time: Measured to compare model efficiency.

## **Results Summary**

| **Model** | **Recall ("yes")** | **Precision ("yes")** | **F1-score ("yes")** | **Accuracy** | **Computation Time (s)** | **Notes** |
| --- | --- | --- | --- | --- | --- | --- |
| Logistic Regression | 0.61 | 0.15 | 0.24 | 0.56 | Fast (~1s) | Best recall, very efficient |
| Decision Tree | 0.58 | 0.16 | 0.24 | 0.59 | Very fast (<1s) | Good recall, interpretable |
| SVM | 0.58 | 0.16 | 0.24 | 0.59 | Moderate (~2-5s) | Similar recall, slower than LR/DT |
| KNN | 0.62 | 0.23 | 0.33 | 0.81 | Slowest (varies) | Slightly better F1, but slower |

## **Best Model for Recall**

* Logistic Regression with SMOTE achieved the highest recall for the "yes" class, making it the best choice for the campaign's goal of maximizing true positives.
* Computation time for Logistic Regression was also among the lowest, making it suitable for large-scale or real-time applications.

## 

## **Trade-offs**

* Models with higher recall tended to have lower precision, meaning more false positives (non-subscribers predicted as "yes"). This is acceptable for this marketing use case, as the cost of missing a potential subscriber is higher than the cost of contacting an uninterested customer.

## **Recommendations**

* Use the Logistic Regression model with SMOTE for campaign targeting to maximize "yes" responses.
* Consider further feature engineering and threshold tuning to fine-tune the balance between recall and precision if business constraints change.
* Monitor model performance regularly as customer behavior and market conditions evolve.

## **Author**

Ambreen Zavaer