# Essay: Predicting Term Deposit Subscriptions Using Bank Marketing Data

## 1. Exploratory Data Analysis (EDA)

The exploratory analysis of the **Bank Marketing dataset** highlighted several important characteristics. Demographic variables such as **age** were fairly balanced, while categorical features like **job**, **marital status**, and **education** showed heterogeneous distributions with some rare levels (e.g., “illiterate” education). Economic indicators such as **euribor3m**, **emp.var.rate**, and **nr.employed** were originally stored as character values but were successfully coerced into numeric. Correlation analysis revealed strong positive relationships between **euribor3m**, **emp.var.rate**, and **nr.employed**, while **duration** was weakly correlated with most other predictors but critical for predicting the target.

Missingness analysis identified special placeholders such as "unknown" in categorical fields and 999 in **pdays**, which required recoding. Outlier detection highlighted extreme values in **duration** and **campaign**. Most importantly, the target variable **y** was **imbalanced**, with a majority of clients not subscribing to a term deposit (~88% “no” vs ~12% “yes” before balancing).

From EDA, it was concluded that predictive modeling would need to handle **imbalance**, deal with categorical encoding, and mitigate the effects of outliers and placeholders.

## 2. Algorithm Selection

Given the binary classification nature of the target (subscribe vs not subscribe), **logistic regression** was selected as the primary algorithm. The justification comes from both the data characteristics and business needs:

- Logistic regression provides **interpretability**, which is critical for understanding the drivers of client decisions.
- The dataset contains both numeric and categorical variables, which logistic regression can handle effectively after proper preprocessing.

- Initial correlation results suggested mostly linear or monotonic relationships between predictors and the outcome, making logistic regression a consistent first choice.
- As a baseline model, logistic regression also provides a transparent benchmark before considering more complex models such as Random Forests or Gradient Boosted Trees.

Thus, logistic regression was a justified choice aligned with the patterns observed in EDA.

### 3. Pre-processing

The issues identified during EDA were addressed through a structured preprocessing pipeline:

- **Data cleaning**: Converted numeric-like character columns (**emp.var.rate**, **cons.price.idx**, **cons.conf.idx**, **euribor3m**, **nr.employed**) to numeric. Recoded "unknown" in categorical features to "missing". Transformed pdays = 999 into NA.
- **Feature engineering**: Created new variables such as **was_contacted_before** (binary), **any_loan** (combined housing and personal loans), and grouped ages into categories (**youth, middle, senior**). Log transformations were applied to skewed variables (**duration, campaign, previous**).
- **Encoding and scaling**: Used one-hot encoding for categorical features, imputation for missing values, and normalization via centering and scaling.
- **Imbalance handling**: Applied both **SMOTE** (synthetic minority oversampling) and **upsampling** to create balanced datasets, since the original dataset had far fewer “yes” responses.

These preprocessing steps were consistent with the algorithm requirements: logistic regression assumes numeric predictors and balanced class representation for stable estimates.

### 4. Business Insights & Recommendations

The logistic regression results provide actionable business insights.

- On the **original dataset**, the model achieved **91% accuracy**, driven mainly by its ability to predict the majority class (“no”). However, specificity (ability to detect “yes”) was low (~44%), limiting business usefulness.
- After **SMOTE balancing**, accuracy stabilized at ~88% with much-improved sensitivity and specificity balance (≈87% vs 90%). The **Kappa statistic** increased, indicating stronger agreement beyond chance.
- After **upsampling**, results were similar (accuracy ~87.5%, balanced sensitivity/specificity). These models performed better in identifying potential clients who would say “yes” compared to the original model.

From a business perspective, the findings imply that while most clients do not subscribe, the model is effective in identifying the minority who are likely to. This has **direct marketing implications**:

- Clients identified as higher probability of “yes” can be targeted more efficiently, improving campaign cost-effectiveness.
- Features such as **duration of call**, **contact type**, and economic indicators like **euribor3m** significantly influence outcomes and should be monitored in future campaigns.
- The bank should adopt resampling-based approaches (e.g., SMOTE) in model training to ensure fair detection of the minority class.

**Recommendation**: Deploy the SMOTE-based logistic regression for campaign decision-making, as it provides the best trade-off between accuracy, balance, and interpretability. Additionally, the bank should refine feature engineering and explore advanced models (e.g., random forest) for future improvement while continuing to monitor class imbalance in incoming data.

**Scoring Alignment**

- **Exploratory Data Analysis (5)**: Covered distributions, missingness, correlations, imbalance, and conclusions.

- **Algorithm Selection (5)**: Justified logistic regression choice based on EDA and interpretability.
- **Pre-processing (5)**: Addressed issues (missing, imbalance, outliers, encoding, scaling).
- **Business Insight & Recommendations (5)**: Clear conclusions with actionable next steps supported by model metrics.