**Introduction**

**Title**: Decoding Term Deposit Subscriptions: Advanced Modelling and Hypothesis Validation in Banking Analytics

**Objective**: To investigate the factors influencing term deposit subscriptions in the banking sector using advanced statistical modeling and hypothesis validation techniques.

## Table of Contents

- Further Steps
- Learnings and Future Aspiration

5. References
6. Appendix

## 1. Introduction and Background

### 1.1 Overview and Problem Statement

The banking industry relies significantly on long-term deposits for income. This study aims to understand the variables that impact term deposit subscriptions by examining diverse client attributes, marketing strategies, and economic factors. The goal is to enhance marketing initiatives and product sales through a deeper comprehension of these variables.

### 1.2 Literature Review

The literature review covers various studies that utilized machine learning algorithms for predicting term deposit subscriptions. The models and their accuracy from multiple research publications are summarized in a detailed table within the document.

### 1.3 Hypotheses Assumed

1. Personal Loan vs. Subscribed
   - H0: No relationship between Personal Loan and Subscribed
   - H1: Relationship exists between Personal Loan and Subscribed
2. Pdays vs. Subscribed
   - H0: No relationship between Pdays and Subscribed
   - H1: Relationship exists between Pdays and Subscribed

3. Occupation vs. Subscribed
    - H0: No relationship between Occupation and Subscribed
    - H1: Relationship exists between Occupation and Subscribed
4. Credit Default vs. Subscribed
    - H0: No relationship between Credit Default and Subscribed
    - H1: Relationship exists between Credit Default and Subscribed
5. Campaign vs. Subscribed
    - H0: No relationship between Campaign and Subscribed
    - H1: Relationship exists between Campaign and Subscribed

## 2. Methodology

### 2.1 Analytical Approach and Tasks

The process involves data pre-processing, hypothesis creation and testing, data visualization, statistical association measurement, regression analysis, and model evaluation, following the CRISP-DM methodology.

### 2.2 Data Exploration and Data Quality Assessment

The dataset comprises 21 variables, including 11 categorical and 10 numerical indicators. The descriptive statistics for the variables are provided.

### 2.3 Variable Selection

A total of 15 variables were selected based on hypothesis, research papers, and logical reasoning.

### 2.4 Data Quality Issues

Data quality issues identified include outliers and missing values in variables like age, marital status, month, and day of the week.

### 2.5 Addressing Data Quality Issues

Techniques like IQR for outlier removal and mode imputation for missing values were used to address data quality issues.

### 2.6 Hypothesis Testing

Hypothesis testing involved evaluating the null and alternative hypotheses for various variables using statistical tests like Pearson's Chi-squared test and Welch Two Sample t-test.

### 2.7 Regression Model Techniques

Multiple logistic regression was used to assess binary outcomes with models incorporating a linear combination of predictors.

### 2.8 Model Building

Three models were built using the forward selection method:

- Model 1: Based on hypothesis-derived variables
- Model 2: Based on hypothesis and literature review-derived variables
- Model 3: Incorporating all variables from hypothesis, literature review, and logical reasoning

## 3. Results and Discussion

### 3.1 Presentation of Key Outputs

- **Model 1**: 89.47% accuracy with moderate fit (Pseudo-$R^2$ values)
- **Model 2**: 89.58% accuracy with slightly better fit (Pseudo-$R^2$ values)
- **Model 3**: 90.78% accuracy with the best fit (Pseudo-$R^2$ values)

### 3.2 Presentation of Key Outputs of All Models

The output metrics for all three models are summarized in a detailed table.

### 3.3 Plot of Key Outputs of Best Model (Model 3)

Diagnostic plots and ROC curves for Model 3 are provided, demonstrating strong model performance.

### 3.4 Model Assumptions

The assumptions of multicollinearity, assessed using VIF, and other regression diagnostics are discussed.

## 4. Reflective Commentary

### 4.1 Further Steps

Future improvements include addressing data quality issues, integrating advanced algorithms, and enhancing model interpretability.

### 4.2 Learnings and Future Aspiration

The experience has enhanced proficiency in statistical modeling, leading to aspirations in advanced supervised learning and machine learning strategies.

## 5. References

A comprehensive list of references used throughout the study is provided.