Capstone Project Proposal: Predicting Term Deposit Subscriptions for a Portuguese Bank

---

Problem Identification

Problem Statement Formation

The primary objective of this project is to predict whether a client will subscribe to a term deposit based on a bank's marketing data. The current marketing strategy involves repeated phone contacts with clients, resulting in high costs and time investment. The goal is to develop a predictive model that identifies clients more likely to subscribe, enabling the bank to target high-potential leads, reduce campaign costs, and improve conversion rates.

Context

The dataset captures data from a Portuguese bank's marketing campaigns conducted between May 2008 and November 2010, focused on promoting term deposits through phone calls. Each record includes client demographic details, previous interactions, and subscription outcomes. Given the high volume of outbound calls and associated costs, an optimized targeting model is valuable for enhancing efficiency and ROI in marketing efforts.

Criteria for Success

The success of this project will be measured by:

- The accuracy and precision of the model in predicting "yes" outcomes for term deposits.
- Reduction in the number of contacts needed to achieve a given level of conversions, as estimated through model testing.
- Achieving a minimum target model accuracy of 80% on the test dataset.
- Demonstrating a reduction in campaign costs (hypothetically estimated based on contact reduction).

Scope of Solution Space

This project will focus specifically on:

1. Building and evaluating a predictive model to classify potential subscribers.
2. Exploring various classification algorithms to improve predictive performance.

3. Providing actionable insights based on model outcomes for targeted marketing strategies.

Non-goals:

- Detailed financial analysis of campaign costs or the impact of external economic factors on client behavior.
- Developing real-time or live deployment (the focus is on generating insights and recommendations).

Constraints

- **Data Limitations**: The dataset includes only historical campaign data, so it may not fully capture current client behaviors or external factors.
- **Imbalanced Classes**: There may be fewer "yes" responses, which could require special handling techniques.
- **Computational Resources**: Given the dataset size, computationally heavy models may require additional processing time.
- **Time Constraints**: Limited to a project timeline of two months.

Stakeholders

- **Primary Stakeholders**: The bank's marketing department, which will use the insights to optimize campaigns.
- **Secondary Stakeholders**: The financial team, as improved targeting could reduce costs and increase profitability.

Data Sources

Data Description

The dataset, sourced from a Portuguese banking institution, comprises:

- **Records**: 41188 entries from May 2008 to November 2010.
- **Features**: 20 input variables, including:
    - Demographic data (age, job, marital status, etc.)
    - Socioeconomic factors (education, housing status, etc.)

- Interaction data (campaign contact details, previous interactions)
- Outcome variable: Whether the client subscribed to the term deposit (yes/no).

Data Acquisition and Preparation

The data is provided as a CSV file. The preparation will include:

- Cleaning and handling any missing values.
- Encoding categorical variables as required.
- Addressing any class imbalance issues (using techniques like SMOTE or class weighting).
- Normalizing or scaling features if needed, based on model requirements.

Solution Approach

Problem-Solving Strategy

The project will involve the following steps:

1. **Exploratory Data Analysis (EDA)**: Assess data distributions, identify patterns, and check for any preprocessing needs.
2. **Feature Engineering**: Transform features, encode categorical variables, and consider interaction features if beneficial.
3. **Model Selection and Baselines**:
   - Start with simpler models like Logistic Regression and Decision Trees as baselines.
   - Progress to more complex models such as Random Forests, Gradient Boosting, and potentially Neural Networks for enhanced accuracy.
4. **Model Evaluation and Tuning**:
   - Use metrics such as accuracy, precision, recall, F1-score, and AUC-ROC to evaluate models.
   - Cross-validation and hyperparameter tuning to optimize model performance.

5. **Interpretation and Insights**:
    - Identify key features that influence predictions.
    - Provide insights for targeting potential customers.
6. **Testing and Validation**: Test the model on hold-out data and validate results, checking generalizability.

Deliverables

- **GitHub Repository**:
    - Containing all code, data processing scripts, EDA, and model training notebooks.
- **Project Report**:
    - Detailed documentation covering the problem statement, methodology, results, and recommendations.
- **Presentation Slide Deck**:
    - A summary of key findings, including visuals and insights for stakeholder presentation.