**Project Charter**

***Project Group 9***

***Zeju Li, Zezhen Liu, Zonghong Yu***

***Wednesday, September 7, 2023***

# **Overview**

The banking industry, historically rich in data, has consistently leveraged vast amounts of information to understand, predict, and enhance customer behavior (Zhuang, Yao, & Liu, 2018). Traditional models primarily focus on individual banking transactions - deposits, withdrawals, loans - forming a granular understanding of a customer's financial habits. However, in the evolving world of finance, there's a growing recognition: that personal financial decisions aren't made in a vacuum. They're often deeply intertwined with the broader socio-economic landscape.

The "Bank Marketing (with social/economic context)" dataset from the UCI repository exemplifies this shift in perspective. Beyond the conventional attributes like age, job, and education, it incorporates socio-economic factors, providing a more holistic view of potential bank clients. The added dimensions promise to unveil deeper layers of their decision-making processes, especially in areas like telemarketing, which relies heavily on understanding client motivations.

Central to our exploration is a pivotal question: ***Can a client's decision to subscribe to a term deposit be better predicted by harmonizing both individual banking data and overarching economic indicators?*** This paper endeavors to unravel this, hypothesizing that a comprehensive analysis, blending these two realms, will yield a more refined predictive accuracy. Our aim is to bridge the gap between micro-level banking analytics and macro-level economic insights, presenting a unified, more insightful lens to view and predict banking behaviors.

## Project Background and Description

***What is so interesting about this problem?***

The banking sector, despite its long-standing existence, remains one of the most dynamic industries, adapting to new technologies and methodologies. This intersection of banking behavior with larger economic indicators is fascinating to the team as it offers a multifaceted view of customer decisions. We also find it interesting for it challenges the traditional boundaries of banking data analytics by suggesting that perhaps, a client's decision to subscribe to a term deposit is not just influenced by their personal banking history but also by the economic environment around them.

***What are the proposed benefits of the solution (why do we need to solve this)?***

The implications of an enhanced predictive model are manifold. For banks, this could mean a more targeted and efficient telemarketing strategy, leading to cost savings and increased revenue. For policymakers and economic researchers, understanding these correlations can offer insights into how national economic indicators trickle down and influence individual financial decisions.

***What problem type are we evaluating (classification? regression, or something else?)***

At its core, we are tackling a classification problem. *The objective is binary: predicting if a client will or will not subscribe to a term deposit.* However, the factors influencing this decision, as we hypothesize, could be multivariate, encompassing personal banking details and broader economic indicators.

## 1.2 Project Scope

The overarching goal is to navigate through the "Bank Marketing (with social/economic context)" dataset, understand its intricacies, and formulate predictive models that can capitalize on the enriched data. The project will employ a combination of generative and non-generative methods, further enhancing them with gradient boosting and bagging techniques. A significant focus will be on hyperparameter tuning, generative methods, and some other models, comparing the models to select the most effective one. The deliverable is not just a model but an understanding of the interactions between various attributes and their collective influence on a client's decision.

***What methods are we evaluating? (generative, nongenerative, trees? hyperparameters?)***

The exploration will start with generative and non-generative methods as foundational analytical tools. As we delve deeper, techniques like gradient boosting and bagging will be employed. These methods will be tuned, with various parameters like n\_estimator and learning rate being adjusted to optimize performance. Decision trees might also be explored for their interpretability and to understand feature importance.

***What is the deliverable?***

The project aims for multiple deliverables:

* A comprehensive report detailing the dataset analysis.
* A trained machine learning model with documented accuracy metrics.
* Comparative analyses of different methods offer insights into the optimal approach for this dataset.

***What is out of scope?***

Given the constraints of the dataset and the problem, deep learning or neural network approaches might not be the best fit and are considered out of scope. Using the 'duration' attribute for real-time predictions is also out of scope due to its post-call nature.

***What are the constraints?***

The dataset, while rich, comes with its limitations. Categorical attributes with "unknown" labels can introduce ambiguity. The "duration" attribute, which holds significant predictive power, presents a paradox: while it's crucial for prediction, in a real-world scenario, its value isn't known until after a call, rendering it unusable for proactive predictions. Also, the dataset is a snapshot from May 2008 to November 2010, which might not reflect current trends or economic conditions.

***What are the success criteria?***

* A model that can predict with high accuracy.
* Demonstrable evidence that the integration of social and economic indicators enhances prediction accuracy.
* A thorough understanding of which attributes hold the most predictive power, offering insights into client behavior.



## 

## 1. Generative Methods-Based Analysis of the "Bank Marketing (with social/economic context)" Dataset

## 

## 1.1. Traditional Exploration and Analysis: Focusing on predicting the likelihood of a client subscribing to a term deposit.

## 1.1.1. Preliminary visual exploration and analysis of the data set (exploratory data analysis): Dive deep into the dataset using Exploratory Data Analysis (EDA) to visualize attributes like age distribution, education levels, and economic indicators. Histograms, scatter plots, or heatmap might play a pivotal role in this phase.

## 1.1.2. Propose a possible outcome of the analysis: Based on preliminary insights, we might hypothesize, for instance, that clients within a certain age range, coupled with specific economic conditions, are more likely to subscribe to a term deposit.

## 1.2. Gradient Boost Evaluation: Implement the gradient boosting algorithm on the dataset. Parameters will be varied for optimization, with n\_estimator values set as [10, 25, 50, 75, 100, 125, 150] and learning rates ranging from 0.1 to 1.0.

## 1.2.1. Result Comparison: After model training, a comparative analysis of the results across varied n\_estimator and learning rate settings will be undertaken. The optimal parameter set will be identified and its implications for predicting a client's decision will be discussed.

## 1.3. Bagging Evaluation: Subsequent to gradient boosting, the bagging algorithm will be employed on the dataset. Similar to before, parameters will be varied, aiming to discern the most efficacious settings.

## 2. Repeat 1.1-1.3 for the Nongenerative methods.

## 3. Hyperparameter tuning and analysis will be performed based on the results of 1 and 2.

## 

## 1.3 Deliverables

1. ***General Deliverables*** 
   1. ***Common Approaches, Tools, and Sharable Implementations:*** 
      1. **Data definition and normalization practices:**The dataset from UCI, much like sklearn datasets, offers a detailed description of each feature. It comprises attributes like 'age', 'job', 'education', and socio-economic indicators. While 'age' might denote the age of the client, 'job' could represent the type of job the client is engaged in. Each attribute provides a facet of the client's profile and is instrumental in predicting if they'd subscribe to a term deposit. However, raw data often requires preprocessing. Features with numerical values might span different ranges. For instance, 'age' ranges from 18 to 98, while economic indicators might be in percentages. Normalization techniques, such as Min-Max scaling, will ensure that these features are on a similar scale, aiding in better model training and prediction.
      2. ***Feature definition and training models:*** Features like 'education', 'job', and 'housing' provide insights into the socio-economic standing of the client. Moreover, broader economic indicators might reflect the overall economic health and its potential impact on individual financial decisions. After preprocessing, these features will serve as inputs to our machine-learning models. Models like Decision Trees, Logistic Regression, and Support Vector Machines might be employed initially.
      3. ***Processing pipelines and optimization methods:*** The analysis will follow a structured pipeline: data preprocessing, feature engineering, model training, and evaluation. Preliminary models provide a baseline. However, ensemble methods, such as Gradient Boosting and Bagging, will be integrated to enhance prediction accuracy. These techniques bring the benefit of pooling knowledge from multiple models, ensuring robust predictions.
2. ***Generative Methods-Based Analysis of the dataset*** 
   1. Traditional exploration and analysis to predict the “dependent” variable.
      1. ***Preliminary visual exploration and analysis of the data set (exploratory data analysis):*** Using visual tools like histograms, scatter plots, and heatmaps, we'll decipher patterns, correlations, and outliers. This visual exploration aids in understanding feature distributions and their relationships.
      2. ***Propose a possible outcome of the analysis:*** From the preliminary analysis, we might discover that variables such as 'education' and certain economic indicators have strong correlations with the likelihood of a client subscribing to a term deposit.
   2. ***Use gradient boost to evaluate the dataset.*** Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
      1. ***Compare the results***: ​​Upon training, models will be evaluated, and a comparative analysis will shed light on the best parameter combinations. This will be pivotal in understanding the most influential features and their impact on the outcome.
   3. ***Use bagging*** to evaluate the dataset. Alter the parameters for the boosting procedure (for n\_estimator use [10, 25, 50, 75, 100, 125, 150]; for learning rate use [0.1, till 1.0]).
3. Repeat 1.1-1.3 for the Nongenerative methods.
4. Hyperparameter tuning and analysis will be performed based on the results of 1 and 2.

Reference:

Zhuang, Q. R., Yao, Y. W., & Liu, O. (2018). Application of data mining in term deposit marketing. In *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Vol. 2).

Moro,S., Rita,P., and Cortez,P.. (2012). Bank Marketing. UCI Machine Learning Repository. <https://doi.org/10.24432/C5K306>.