**Project Proposal**

**Project title:**

**Bank Marketing Tactic for a Term Deposit**

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Data Set Information:

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

**Problem Statement and business use case**

* Portuguese banking institution is trying to determine through their previous years data that how many people will subscribe to a term deposit scheme launched by them.
* Now in order to tackle this, the company has decided to form a strategy which will be able to provide better and innovative services to their customers along with making sure that people utilize their product leading to revenue generation.

**Data Science Workflow**

1. **Measures as recorded**

Input variables:

**# Bank client data:**

1 - age (numeric)

2 - job: type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")

3 - marital: marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)

4 - education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")

5 - default: has credit in default? (Categorical: "no","yes","unknown")

6 - housing: has housing loan? (Categorical: "no","yes","unknown")

7 - loan: has personal loan? (Categorical: "no","yes","unknown")

**# Related with the last contact of the current campaign:**

8 - contact: contact communication type (categorical: "cellular”, “telephone")

9 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

10 - day\_of\_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

**# Other attributes:**

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

**# Social and economic context attributes**

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: "yes","no")

1. **Basic assumptions.**

* Data shared with us is of existing and previous client data base.
* We want to build a Logistic regression model to predict how many people will be subscribing to the term deposit scheme.
* The classification goal is to predict if the client will subscribe a term deposit.

1. **Solutions:**

* Identify which factors are important in predicting whether or not a person will subscribe to the term loan. Using EDA.
* To build a model to predict the output variable based on these attributes.

1. **Alternative solutions, is to work on some exploratory data analysis.**

* At what points along the duration and age is most of the data accumulated.
* For how many people the marketing campaign was successful previously alongside how many people are considering subscribing to a term deposit.
* For how many people the marketing campaign was a failure previously alongside how many people are considering subscribing to a term deposit.
* Up to what duration, maximum people in this dataset prefer giving their time to such marketing campaigns contact point.
* what was the age bracket most of the people in the dataset belonged to?
* How much contact was made with the client alongside what proportion of them are subscribing to and not subscribing to the term deposit.
* How many of each category professionals are subscribing to the term deposits.
* How many of each category of marital statuses are subscribing to the term deposits.
* How many of each category of educational backgrounds are subscribing to the term deposits.
* How many people have made defaults in their last loans, how many have an existing housing loan and a personal loan; how many of each of these people are going to subscribe for the term loans.
* How many contacts were made in the previous campaigns for these clients and how many of them are going to opt for the term deposit scheme in this campaign.

1. **What can we do to fix it?**
   * Awareness campaigns.
   * More offers.
   * Defining target market.
   * Better service than competitors.
   * 24/7 online support.
   * Better knowledge of public demands.

**Data Collection**

* Data set is in a csv form.
* Data set consists of previous and present years data.
* Data shape = 41188 rows and 21 columns
* It has size of 5.56MB
* Data columns are: 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'duration', 'campaign', 'pdays','previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'

**Data preprocessing preparation and Feature Engineering.**

* Identifying data quality.
* Check for data inconsistency, data types and no encoding issues.
* Handle missing values.
* Treating categorical values
* Treating outliers
* Data scaling
* Feature engineering.

**Machine Learning and model selection**

* Logistic Regression Model