Analytics Startup Plan

Synopsis: ***This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

Project	Direct Market Campaign Analysis
Requestor	Portuguese Banking Institution
Date of Request	13 July, 2022
Target Quarter for Delivery	8 August, 2022
Epic Link(s)	NA; Not an agile group
Business Impact	Predict the subscription of term deposit by a client
Presented By	Hinna Mary Steephen

1.0 Business Opportunity Brief

Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment.

The purpose of this project is to determine a marketing strategy to conduct the direct phone call campaign that reaches out to the clients for the subscription of the term deposit.

Different variables are analyzed to determine the factors the most contribute most to the success of the campaign so that the campaign can be conducted most effectively to achieve maximum subscriptions with minimum cost.

This analysis will help the campaign to allocate resources to a target set of clients who are most likely to subscribe to the term deposit.

The specific ask:

Clearly articulate the specific task you will be conducting to help achieve the opportunity

Predicting the likelihood of a client subscribing to the term deposit of the bank when reached out through telemarketing.

1.1 Supporting Insights

Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally?

Portuguese banks have a sophisticated financial system that is comparable to those in Europe and other developed nations. The majority of banks provide online banking services, and there are ATMs scattered over the nation. Almost any type of transaction is possible, including deposits (cash and checks), interbank transfers, and payments for services like mobile phones and the internet.

Top 5 Banking competitors in the Portuguese market as of 2020

Bank	Total Asset	Interest Rate
Caixa Geral de Depósitos	EUR 96.29 billion	5.7%
Banco Comercial Portugues	EUR 88.57 billion	9.25%
Banco Santander Totta S.A.	EUR 54.40 billion	5.0%
Banco Popular Portugal	EUR 723 million	4%
Banco Internacional do Funchal	EUR 523 million	3.5%

Considering how much money is being spent on marketing in the banking sector, banks must increase the efficiency of their marketing plans. Recognizing client needs, results in more successful marketing strategies, clever product ideas, and higher levels of customer happiness.

Goals that are poorly defined or unrealistic are one of the main problems telemarketers confront. To deem a campaign successful, the telemarketing team must ascertain the anticipated cost, the target clientele, and the conversion rate. To make the campaign successful, they must analyse and determine how much time is required as well.

1.2 Project Gains

Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? *This section is particularly key for prioritization against company goals and KPI's.*

This project will help the marketing team understand the persona of the clients they are contacting and will help the telemarketer predict if the client is a potential subscriber. With the help of this analysis, the company will be able to focus their resources on clients determined by the analysis where there is potential for subscription of the term deposit.

This project will help the company in all 4 quantifiable measurements i.e., revenue, quality, time and cost. Since the analysis will help the marketing team target the client pool. They can focus majority of their time and resource of potential customers this will in turn improve the quality of the campaign and increase the revenue as more subscription can be achieved.

If we do not conduct this analysis and determine the potential clients, we will be wasting our time and resources trying to contact all the clients or the wrong pool of clients and most of the clients the telemarketers reach out to may not be interested in the term deposit leading to a low conversion rate.

We may also miss out of potential clients because we could not prioritize the call or exhausted our resource before reaching to potential client contact. This will lead to high expense and low revenue

Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.

2.0 Analytics Objective

List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.

Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed.

Data Description

The dataset consists of45211 records with 17 variables and 1 binary target variable.

Dependent Variable: y(yes or no) ; yes - subscribed to term deposit

Independent variables: 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

Job, marital, education, contact, day_of_month, month, default, housing, loan, poutcome are categorical variable and others are numeric.

Modelling

I will be building classification models such as logistic regression, K-NN model, Random Forest, Naïve Bayes to predict the outcome

The dependent variable is a binary variable with output yes or no. Hence, we are using classification models.

The model with the highest AUC will give the best prediction of the campaign result.

2.1 Other related questions and Assumptions:

List any assumptions that may affect the analysis

The following assumptions are made for the study:

- the data provides is a sample data set and is representative.
- It is reliable, original and comprehensive
- It is ethical data

2.2 Success measures/metrics

What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.

AUC definition: AUC stands for area under the curve. The curve mentioned is the ROC(Receiver Operating Characteristic) curve. The binary classifier's performance is represented by the ROC curve. It's a 2-dimensional curve with y axis displaying the True Positive Rate (TPR) and its x axis displaying the False Positive Rate (FPR).

Technical Success metric: We will be using the AUC (area under the curve) value determine which model gives the optimum result. The higher the AUC value the better the model. Once the best model is determined we can use the best model to predict the result.

Business Success metric: With the past campaign, the bank was able to achieve 11% conversion rate i.e., Of the 45211 clients contacted, 5289 clients subscribed to the term deposit.

After the implementation of the analysis if the conversion rate improves from 11% 25%, then the campaign will be considered to be successful.

2.3 Methodology and Approach

Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.

Type of Analysis: Classification
Models: Logistic regression, Random Forest, Naïve Bayes, K-NN Classifier
Methodology:

Data cleaning:

- **Transform data:** change column names, data types wherever necessary
- **Missing values:** There are many representations of missing data. Such as NULL values, NA, empty string, filled as unknown or even numeric values such as 0, 88 or 99. We need analysis and see if the value is the actual data or a missing value. For ex. 0 could be missing value or the value itself.
- **Duplicate:** check whether there are duplicates and remove duplicate values.
- **Outliers:** look for outliers using data visualization as outliers can skew the results.

Exploratory Analysis:

- Look for relationship between different variable using bar charts, scatter plot for example the relation between age and balance.
- Generate correlation matrix understand how different variable are correlated with each other. If any variable is highly correlated then decision must be made whether to keep the variable in study or to remove.
- Draw insights from the analysis that could be used to provide recommendation to the stakeholders

Preprocessing:

- Create dummy variables – convert categorical variable to dummy variable
- handle missing values – fill the missing values using mean, median, mode or remove the record. Depending on the data set and its implication
- Scaling data: if different variables are measures in different units, then we need to normalize the data.
- Splitting data to train and test: 70% train; 30% test

Modelling:

- Once the data is process and ready. We can build different classification model using the training data.

- For this project we are building the following models:
    - Logistic regression
    - Random Forest
    - Naïve Bayes
    - K-NN Classifier
- After building the models we are going to test the model using the validation data.
- We are going to generate the confusion matrix and calculate the accuracy score of each model and determine the model with the highest value. That model will give us the best prediction.

Output: The output will be a set of insights, rules and strategic recommendations we draw from this analysis that will help us to increase the effectiveness of the telemarketing campaign.

3.0 Population, Variable Selection, considerations

Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today.

Audience/population selection: NA

Observation window: May 2008 – November 2010

Inclusions: Age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y

Exclusions: TBD

Data Sources: UCI Machine Learning Repository: Bank Marketing Data Set. (n.d.). Retrieved July 14, 2022, from http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Audience Level: Marketing team

Variable Selection: TBD

Derived Variables: TBD

Assumptions and data limitations: TBD

4.0 Dependencies and Risks

Identification of key factors that may influence the outcome of the project and likelihood of it happening:

Risk	Likelihood (based on historical data)	Delay (based on historical data)	Impact

Recession	*The first quarter of 2022 the GDP growth rate of USA was -1.5%.* *If there is a negative growth rate in 2nd quarter, USA is in recession and its to impact countries all over the world*	*If the GDP growth rate increases*	Inflation, unemployment. Unable to pay of loans.

5.0 Deliverable Timelines

List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process.

Item	Major Events / Milestones	Description	Scope	Days	Date
1.	Kick-off / Formal Request	Schedule meeting with the Advisor	*low*	*2*	*12/07/2022*
2.	Assessment / Triage	Collect the dataset. Do basic analysis and research Determine the impact of the project Understand business requirement	med	3	*14/07/2022*
3.	Analysis Plan	Based on the research conducted create an analysis plan	med	1	*15/07/2022*
4.	Data cleaning	Determine if there are: • Missing values • Duplicates • Data types of the variable • Check for outliers	med	3	*18/07/2022*
5.	Exploratory analysis	• Look for relationship between different variable using bar charts, scatter plot • Generate correlation matrix • Draw insights from the analysis	*high*	*4*	*22/07/2022*

6.	Preprocessing	• Transform variables if needed (column names or data types) • Create dummy variables • Filling missing values • Scaling data • Splitting data to train and test	*high*	*6*	*28/07/2022*
7.	Modelling	Build models using: • Logistic regression • Random Forest • Naïve Bayes • K-NN Classifier Compare the accuracy of the models and determine best modes	*med*	*8*	*5/08/2022*
7.	Documentation	Create a comprehensive report detailing the though process and steps taken to complete the analysis	*low*	*7*	*12/08/2022*
7.	Internal team Presentation	Create a presentation to showcase the findings to the stake holder along with the recommendation	*med*	*10*	*22/08/2022*
8.	Go/No Go	The stakeholders meeting to determine the life of the project	*low*	*3*	*25/08/2022*
9.	Story Board 2	If any change in the requirement new story board created	*low*	2	*27/08/2022*
10.	Pilot				
11.	Delivery & sign-off	Project launch	*low*	*2*	*1/09/2022*