Bank-Telemarketing-Campaign Prediction

# Introduction

Nowadays, marketing spending in the banking industry is massive, meaning that it is essential for banks to optimize marketing strategies and improve effectiveness. Understanding customers’ need leads to more effective marketing plans, smarter product designs and greater customer satisfaction.

### Main Objective:

### The main objective of this project is to increase the effectiveness of the bank's telemarketing campaign[¶](http://localhost:8888/notebooks/anaconda3/Jupyter%20Projects/Materials%20Provided(ineuron)/ML%20Projects-Mine/INTERNSHIP_INEURON-PROJECT/reference%20projects/bank%20campaign_internship/ml-project-bank-telemarketing-analysis.ipynb#Main-Objective:-increase-the-effectiveness-of-the-bank's-telemarketing-campaign)

* ***The classification goal is to predict if the client will subscribe a term deposit***

This project will enable the bank to develop a more granular understanding of its customer base, predict customers' response to its telemarketing campaign and establish a target customer profile for future marketing plans.

By analyzing customer features, such as demographics and transaction history, the bank will be able to predict customer saving behaviours and identify which type of customers is more likely to make term deposits. The bank can then focus its marketing efforts on those customers. This will not only allow the bank to secure deposits more effectively but also increase customer satisfaction by reducing undesirable advertisements for certain customers.

**Analysis objectives :**

* Find the best strategies to improve the next marketing campaign.
* How can the financial institution have a greater effectiveness for future marketing campaigns?
* In order to answer this, we have to analyze the last marketing campaign the bank performed and identify the patterns that will help us find conclusions in order to develop future strategies.

This project shall be delivered in two phases:

**Phase 1:** Modeling

Phase 2: Implementing Transformation Pipelines

**Phase3:** Integration of UI to all the functionalities.

This document also captures the different workflows involved to build the solution, exceptions in the workflows and any assumptions that have been considered.

Once agreed as the basis for the building of the project, the flowchart and assumptions will be used as a platform from which the solution will be designed.

**Note: All the code will be written in python version 3.7**

**Dataset**

This dataset is about the direct phone call marketing campaigns, which aim to promote term deposits among existing customers, by a Portuguese banking institution from May 2008 to November 2010. It is publicly available in the UCI Machine learning Repository, which can be retrieved from [http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#](http://archive.ics.uci.edu/ml/datasets/Bank+Marketing).

**Datafile**

* bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]

## Attribute Information:

***Demographic features:***

1 - **age**: (numeric)

2 - **job**: type of job (categorical)

3 - **marital** : marital status (categorical)

4 - **education**(categorical)

***Personal details:***

5 - **default** : has credit in default? (categorical)

6 - **housing** : has housing loan? (categorical)

7 - **loan**: has personal loan? (categorical)

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

8 - **contact**: contact communication type (categorical)

9 - **month** : last contact month of year (categorical)

10 - **day** : last contact day of the week (categorical)

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)

***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- **deposit** - has the client subscribed a term deposit? (categorical)

## Work flow in short

1. Getting the system ready and loading the data
2. Understanding the data
3. Exploratory Data Analysis (EDA)

* Univariate Analysis
* Bivariate Analysis

1. Outlier treatment
2. Feature Importance
3. Feature Engineering
4. SMOTE Analysis for unbalanced dataset
5. Transformation Pipelines

* Data Cleaning Pipeline  
  Log Transformation Pipeline
* Scaling Pipeline
* Encoding Pipeline
* SMOTE Pipeline
* Full Datapreprocessing Pipeline

1. Lazy Predict Modelling
2. Model Building: Part 1

* RandomForest
* XGBoost
* SVM
* KNN

1. Model Building: Part 2

* XGBoost Hyperparameter Tuning

1. Saving Model as pickle file
2. Final Pipeline including Model
3. Saving Pipeline as pickle file
4. Pipeline testing

**Understanding the data**

Data has 41188 rows and 20 inputs, ordered by date (from May 2008 to November 2010),

**Data Cleaning Pipeline**

* **Dropping unwanted columns**

'emp.var.rate','cons.price.idx','cons.conf.idx',

'euribor3m','nr.employed','contact','poutcome'

* **Column Rename:**

{'default':'credit\_default','housing':'housing\_loan',

'loan':'personal\_loan','day':'last\_contacted\_day',

'month':'last\_contacted\_month','duration':'last\_call\_duration',

'campaign':'contacts\_during\_camapign','pdays':'days\_passed',

'previous':'contacts\_before\_campaign','y': 'deposit'}

* **Converting some Numerical columns to categorical columns**
* **days\_passed** was converted to categories[**recent,never\_contacted**]
* **job** –[**salaried,self\_employed,retired,unemployed,students,unknown**]
* **education** – [**primary,secondary,tertiary,illetrate,unknown**]
* **last\_contacted\_month** – [**jan-april,may-aug,sep-dec**]
* **contacts\_before\_campaign** – [>10 ,0]
* converted **call\_duration** from seconds to minutes

**EDA**

**Univariate Analysis**

* **Target variable(deposit)**

-Relative frequency of classes are 89% and 11%

- It is clearly an unbalanced dataset

* **Categorical variables**
* Demography
  + Majority of the customers are salaried and married and belongs to the secondary and tertiary educational categories
* Around 50% of the customers have housing loan whereas only around 15% of them have personal loan
* Bank has contacted most of the customers between May and August
* Customers were contacted mainliy in week days
* Credit defaulters are majority
* **Numerical columns**
* last\_call\_duration and contacts\_during\_campaign are right skewed
* Majority of customers contacted are between 30 to 50 years

**Bivariate Aalysis**

**Categorical variables vs target variable**

* Customers with credit default do not have deposits
* It is more likely for a person without a housing loan and personal loan to start a deposit.

**Numerical variables vs target variable**

* calls of higher duration have resulted in higher number of deposits.
* customers with higher account balance are likely to open a term deposit account

**Normalization using Log Transformation(Pipeline)**

* As the numerical features ,ie, age,call\_duration and calls\_during\_campaign are right skew we do log transform to normalize them

**Feature Importance**

* Finding feature importance using ExtraTreeRegressor-This basically helps to find the important features ,ie ,those features which are highly correlated with the target variable
* Found **last\_call\_duration** to be one of the most important features

**SMOTE Analysis**

* Relative frequency of classes are 89% and 11%It is clearly an **unbalanced** dataset
* So Smote Analysis was performed to make it balanced.

**Modeling :Part 1**

* Tried 4 models to find the best performing one
* RandomForest
* XGBoost
* SVM
* KNN
* Found XGBoost to be performing the best accuracy and F1score

**Modeling : Part 2**

**XGBoost Hyperparameter Tuning**

* To improve the performance of the model we do hyperparameter tuning using **Randomized Search Cross Validation**
* This could improve the model accuracy and F1 score to

**Final Pipeline**

* Combining all the pipelines using column Transformers from sk-learn

**Saving the model and pipeline**

* Saving both the model and the final pipeline to pickle file for future use

**Testing the pipeline**

* Finally testig the final pipeline by providing an input data