Which Clients are more likely to subscribe to Term Deposit?

It is consisted of 41,188 customer data on direct marketing campaigns (phone calls) of a Portuguese banking institution. The dataset provided can be divided into categories:

Client: age, job, marital, education, default status, housing, and loan

Campaign: last contact type, last contact month of year, last contact day of the week, and last contact duration

Others: number of contacts performed in current campaign, number of days that passed by after the client was last contacted, number of contacts performed before this campaign, outcome of previous campaign, and whether a client has subscribed a term deposit

Business Problem:

The Portuguese Banking Institution is observing a decline in their revenue, thus looking for ways to learn from the historical data and improve their existing system. The Portuguese Bank had run a telemarketing campaign in the past, making sales calls for a term-deposit product. The term Deposit refers to allowing bank to hold onto a deposit for a specific amount of time, so banks can invest in higher gain financial products to make a profit.

Target:

A classification approach to predict which clients are more likely to subscribe for term deposits.

Analysing Data:

The given data shape is (**41,188,21).**

The response or "y" has binary values yes/no. Only 11.7% of the total clients that were contacted subscribed for the term-deposit. Since the data is highly unbalanced, thus it requires scaling. Various techniques like up-sampling, down-sampling, SMOTE, ROSE can be implemented.

Some values in the dataset are "unknown", thus referred as missing values. Most of them is of variable "default", which tells us whether the client has previous credit or not.

None of the feature is correlated with the target. But a variable like duration is very essential as longer the call duration, more likely the customer will subscribe.

The age distribution is skewed. The clients contacted mostly belong to age-group (30-40), thus their target is middle aged people.

Amongst the job of clients, an admin is accounts for 24% of the total clients. While, blue collar workers come after them and the percentage is 21%. Students are the least around 2%.

As per the data, 30% clients have a university degree. While 22% have a high school degree.

Equal calls are made on all days.

Features like“duration” and “nr.employed” have the highest standard deviation, which indicates the last contact duration and quarterly number of employees are the key points that need be focused on in this.

Solution:

The categorical features like job, education, marital status, etc are encoded using one-hot encoding. The label encoding could also be used, but in this case ml model may consider the values 1,2,3 etc in a ranking order. Thus one-hot encoding is implemented, converting categorical features into 0’s and 1’s.

The missing values are replaced by the mode of the respective columns.

The data is balanced using up-sampling, increasing the size of minority class (“y” = yes).
The bank marketing data is split into 70% training and 30% test set. The training data is fitted into a **xgboost** model, which works well on sparse data.
The **accuracy** achieved **is 89.27%.**
While the **auc (area under curve)** is **89.49%.**