Project Report : Predicting Bank Term Deposit Subscription

❖ Introduction

➢ Background

- In the competitive banking sector, marketing campaigns play a crucial role in promoting financial products such as term deposits. However, these campaigns often require significant resources and operational costs. To improve efficiency and reduce unnecessary expenses, banks aim to identify potential customers who are more likely to subscribe to a term deposit.

➢ Objective

- The main objective of this project is to build a predictive model that can determine whether a client will subscribe to a term deposit based on historical marketing data. This prediction enables banks to focus their marketing efforts on the most promising customers, thereby increasing the success rate and optimizing resource allocation.

➢ Dataset Description

- The dataset used in this project is the **Bank Marketing Dataset**, originally made publicly available by [Moro et al., 2014]. It consists of **41,188 records** and **20 input features** along with a binary target variable (`y`), which indicates whether a customer subscribed to a term deposit (`yes` or `no`). The features include client demographic details, previous campaign information, and economic indicators.
- Dataset Link : https://archive.ics.uci.edu/dataset/222/bank+marketing

```
➢ ### Attribute information:
➢
➢    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 for modeling):
➢
➢   21 - y - has the client subscribed a term deposit? (binary:
   "yes","no")
➢
```

- Problem Statement
  - This is a **binary classification problem** where the target variable `y` takes two possible values:
    - `yes` → client subscribed to a term deposit
    - `no` → client did not subscribe
  - The challenge lies in the **imbalanced nature of the dataset**, where the number of customers who subscribed (`yes`) is significantly lower than those who did not (`no`).

- Term Deposit
  - A Term deposit is a deposit that a bank or a financial instituruion offers with a fixed rate (often better than just opening deposit account) in which your money will be returned back at a specific maturity time. For more information with regards to Term Deposits please click on this link from Investopedia: https://www.investopedia.com/terms/t/termdeposit.asp

- **Results**

  - **Without Predictive Model**:

    - Bank calls **all 5,911 customers** (test dataset).
    - Total marketing cost = **$29,555** (at $5 per call).
    - Net profit = **$10,407**.
- **With Predictive Model**:
    - Bank calls **only 1,041 predicted customers** (likely to subscribe).
    - Total marketing cost = **$5,205**.
    - Net profit = **$20,953**.

Impact:

- Marketing cost reduced by **82%** (from $29,555 to $5,205).
- Net profit increased by **101%** (from $10,407 to $20,953).

Note: here cost and revenue numbers are assumed according to banking domain to **Define Custom business matrix** and find optimal threshold according to it.

If we not want to do this, we can also find optimal threshold using highest F1 score or other matrix according to problem, but in real business problems we need to define custom business matrix and decide threshold according to domain for maximize profits. So i had taken approximate numbers for this problem which are as beloved.

```
### Define business metrics ###

# 1. marketing_cost per call
# Affected by TP and FP --> we make call to both , 1.actual true and
predicted true , 2. Actual false but predicted true)
call_cost = 5.0  # (including labor & infrastructure)

# 2. this cost are affect after customer Acquisition (agree for
subscription)
#    (only Affected by TP-actual true and predicted true )
# Onboarding & KYC verification cost
onboarding_cost = 5
#Account management costs(back-office ops, compliance)
acc_cost = 15
# other overheads (like customer support):
other_overheads = 10
total_customer_cost = onboarding_cost + acc_cost + other_overheads  #
Total acquisition cost

# 3. Calculating revenue per customer (only Affected by TP-actual true and
predicted true)
# Average deposit amount = 4000
# Interest spread (Interest Earned by Bank on Loans or Investments -
Interest Paid to Customer on Deposit) = 2.2% annually
```

```
# Customer holds deposit for 1 year
subscription_revenue = 88 # (4000 * 0.022)
```

Custom business matrix

Profits =

(Actual subscribers from all call made * revenue from each subscriber) – (Actual subscribers from all call made * each customer cost) – (total code made * cost per call)

❖ Key Insights from dataset (EDA)

1) Most people are 30 to 60 year old. People age less than 30 year have slightly higher conversion ratio then 30 to 60. People age greater than 60 year have highest conversion ratio among all (nearly 0.43-> 43 percent conversion).
2) Most people contacted <= 4 times during the campaign. Repeated contact (>4) is associated with reduced conversion rates. So aggressive campaigning May Hurt conversion rates.
3) Most people not contact before the campaign. Who contacted before current campaign their chance of conversion is increase (0.09 - not contact before, 0.62 - Who contacted before).
4) Economic indicators play decisive role in subscription .
   a. When emp.var.rate is low -> conversion are high. (strong indicator)
   b. When cons.price.idx is low -> conversion are high.
   c. Not much deceive but if cons.conf.idx is very low or very high -> conversion are high.
   d. When euribor3m is low -> conversion rate are high. (strong indicator)
   e. When nr.employed is low -> conversion are high.(strong indicator)
5) Student and retired people are having more conversion rate.
6) Single individuals are slightly more conversion rate then married or divorced.
7) Higher education has better conversion.
8) Housing and personal loan have minimal effect on conversion.
9) Having no credit default leads to much higher chance of saying yes.
10) People contacted via cellular were 3x more likely to say yes then people contacted via telephone.
11) Months mar , dec , oct , sep have high success rates and may had very poor.
12) Day of week not much effective.
13) Previous campaign success is best predictor. If customer previously responded positively, they are 65% likely to convert again.

❖ approaches used to build predictive model and solve this problem

To address this problem, the following approaches were performed:

Approach 1

Name: O1_model

a. In this model we used standard scalar and one hot encoder in preprocessing pipeline.
b. Find best model using ROC AUC score from all model dictionary.
c. Done hyper parameter tuning on best model.
d. Done model training on best parameters.
e. Defined custom matrix to find optimal threshold.
f. Model evaluation (classification report).
g. SHAP values to explain features importance.

Approach 2

Name: O2_smote_model

Done all step as approach 1 but also included resampling using SMOTETomek in CV loop to avoid data leakage.

Approach 3

Name: O3_fs_model

Done all step as approach 1 but also included feature selection in cv loop to avoid feature selection bias.

Approach 4

Name: O4_model

Done all step as approach 1 but also included resampling using SMOTETomek and feature selection in CV loop.

Approach 5

Name: binning_model

Done best approach from above four with Binnig of some columns (e.g. age)

Approach 6

Name : cluster_model

Done segmentation using clustering .then used best approach from first four approaches and makes predictions models for each segment.

Approach 7

Name : nn_model

Done model training using neural networks .

- ✓ **Best performing model (after hyperparameter tuning) : LightGBM**
- ✓ **Best performing approach : approach 1(3,5,7 give nearly same results)**

- Also done web app for predictions using best approach model.

Github : https://github.com/mayur-kachhad/Bank-Marketing