Data Science - Bank Marketing Campaign

1. Group Information

Group Name: Datalux

Group Members: 3

Name	Email	Country	Uni/Company
Huu Thien Nguyen	nguyenhuuthien27296@gmail.com	Sweden	Skövde University
Asmaa Alqurashi	asmaa.idk@gmail.com	Saudi Arabia	Taif University
Deepak Rawat	deepakrawat68@gmail.com	Ireland	Dublin Business School

Specialisation: Data Science

Submitted to: Data Glacier canvas platform

Internship Batch: LISUM09

2. Problem description

2.1. Introduction

A common approach for increasing business is to run marketing and selling campaigns. Companies use direct marketing to reach certain categories of clients to achieve a specific goal. Customer distant interactions may be centralised in a contact centre, making campaign administration easier. Technology allows us to reimagine marketing by optimising customer lifetime value through analysing accessible data and customer KPIs, allowing us to develop longer and closer relationships in line with company needs.

The ABC Bank wants to market its term deposit product to clients in this project. Before doing so, they want to construct a machine learning model that will assist them in determining whether a particular consumer would buy their product based on the customer's previous interactions with the bank or other financial institution.

To solve the problem mentioned above, the bank wants to use machine learning modelling to identify the customers who are more reluctant to buy their services so that their marketing channels will only focus on these customers, which in turn will save the time and resources and finally leads to optimised cost for this campaign.

3. Data understanding

3.1. Business Problem:

ABC Bank wants to sell its term deposit product to customers. Before launching the product, they want to develop a model which helps them in understanding whether a particular customer will buy their product or not (based on the customer's past interaction with the bank or other Financial Institution). Bank wants to use the ML model to shortlist customers whose chances of buying the product is more so that their marketing channel (telemarketing, SMS/email marketing etc.) can focus only on those customers whose chances of purchasing the product is more.

3.2. The Data:

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

Size: 41188 records, 20 explanatory variables, and one binary response variable

3.3. Columns Description:

- Customer Information:

1 - age (numeric): customers' age which ranges between 17 and 98 in this dataset

2 - job (categorical): customers' job

  - admin
  - blue-collar
  - technician
  - services
  - management
  - retired
  - entrepreneur
  - self-employed
  - housemaid

- unemployed
- student
- unknown

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

- married
- single
- divorced
- unknown

4 - education (categorical)

- university.degree
- high.school
- basic.9y
- professional.course
- basic.4y
- basic.6y
- unknown
- illiterate

5 - default (categorical): Does the customer has credit in default?

- yes
- no
- unknown

6 - housing (categorical): Does the customer have a housing loan? (categorical)

- yes
- no
- unknown

7 - loan (categorical): Does the customer have a personal loan? (categorical)

- yes

    - no
    - unknown

- Communication Information:

8 - contact (categorical): communication type

    - cellular
    - telephone

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

10 - day_of_week (categorical): previous contact day of the week (in working days)

11 - duration (numeric): last contact duration, in seconds

- Campaign Information:

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

13 - pdays (numeric): number of days that passed by after the client was last contacted from a previous campaign

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

15 - poutcome (categorical): outcome of the last campaign marketing

    - nonexistent
    - failure
    - success

- social and economic context attributes:

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

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

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

19 - euribor3m (numeric): Euro Interbank Offered 3-month rate - daily indicator

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

- **Target**:

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

4. Data analysis

- What type of data you have got for analysis?
- What are the problems in the data (number of NA values, outliers, skewed etc.)
- What approaches are you trying to apply to your data set to overcome problems like NA value, outlier, etc., and why?

4.1. General data analysis:

- Dataset does not have a NaN value
- Dataset has 12 duplicate rows
- 41188 rows and 21 columns
- 10 numeric columns and 11 categorical columns

4.2. Univariate analysis:

Thien

1 - age (numeric): customers' age

- Many outliers
    - Ranges between 17 and 98
    - Check outliers (Q1, Q3) with y
    - Can feature engineer 'age_range' (binning value)
    - Check std, mean, descriptive analysis (Boxplot)

2 - job (categorical): customers' job

- Check 12 distinct values
- Check imbalance between groups

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

- Check four distinct values
- Imbalance data
    - Remove or binning value

4 - education (categorical)

- Check eight educations
- Imbalance data
    - Remove or binning value

5 - default (categorical): Does the customer have credit in default?

- Check three distinct values
- Imbalance data
    - Remove or binning value

6 - housing (categorical): Does the customer have a housing loan? (categorical)

- Check three distinct values
- Imbalance data
    - Remove or binning value

7 - loan (categorical): Does the customer have a personal loan? (categorical)

- Check three distinct values
- Imbalance data
    - Remove or binning value

Asmaa

8 - contact (categorical): communication type

- Two distinct values.
- Check imbalance
- Encode the categorical feature (one-hot encoder)

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

- Ten distinct values.
- Check imbalance
- Encode the categorical feature (one-hot encoder)

10 - day_of_week (categorical): last contact day of the week (in working days)

- Five distinct values.
- Check imbalance
- Encode the categorical feature (one-hot encoder)

11 - duration (numeric): last contact duration, in seconds

- Check two approaches

    - As mentioned in the description, the duration greatly impacts the outcome, so it will be better to drop the duration column since it's obtained "After" the call to targeted clients.

    - feature engineer the duration and see the correlation between the outcome and the time of the call if it's above or below average.

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

- Scale the values

13 - pdays (numeric): number of days that passed by after the client was last contacted from a previous campaign

- The value would be -1 if the client were not contacted previously
- Scale the values

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

- Scale the values

Deepak

15 - poutcome (categorical): outcome of the last campaign marketing

- Check three distinct values
- Imbalanced data
- Encode categorical values

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

- Values ranging from -3.4 to 1.4
- Outliers present
- Left skewed
- Feature scaling needed

- Ten distinct numeric values

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

- Values ranging from 92.2 to 94.7
- Most values are close to 93
- Feature scaling

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

- Values ranging from -50.8 to -26.9
- Outliers present
- All negative values
- Feature scaling

19 - euribor3m (numeric): Euro Interbank Offered 3-month rate - daily indicator

- Values ranging from -0.634 to 5.045
- Left skewed
- Feature scaling

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

- Values ranging from 4963.6 to 5228.1
- Feature Scaling

Duplicate rows
12 Duplicate rows

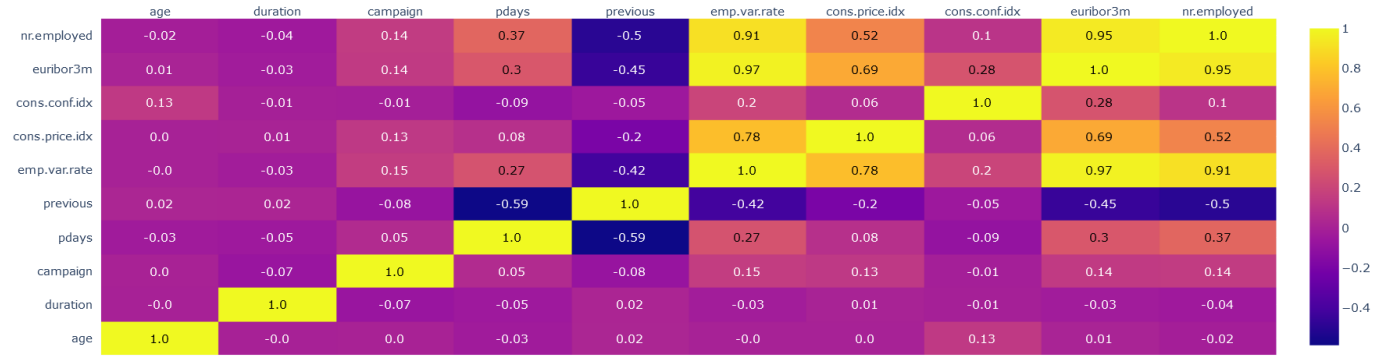
- **Target**:

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

- Count of 'yes' = 4640, count of 'no' = 36548
- Imbalance output
    - Undersampling - It will reduce dataset rows to 9200 approx.
    - Oversampling - This is recommended to use the entire data set
        - Random oversampling
        - SMOTE upsampling
    - K-fold

4.3. Multivariate analysis:

4.3.1. Numeric vs Numeric - Deepak



From the above correlation heat map for numerical variables, it is concluded that there is a high correlation between ‘emp.var.rate’, ‘nr.employed’, ‘euribor3m’ and ‘cons.price.idx’. We can filter the features which have a high correlation for modelling purposes.

4.3.2. Numeric vs Categorical - Asmaa

The categorical features will be analysed paired with the numerical features individually to get more insight into the data. For example, the age and job of the clients and how it affects the decision of making a term deposit.
We have nine numeric features and 11 categorical features, and the analysis will show the correlation with the target.

4.3.3. Categorical vs Categorical - Thien

Most features are categorical types, which would be difficult to analyse manually. Therefore, the automated step using function code will be applied for this process. The categorical features will be compared pair-wise with the help of a crosstab heatmap visualisation. Finally, the results will be examined to prepare for feature engineering and data cleaning.

5. GitHub link

The link for GitHub: https://github.com/AndrewNguyen27296/DataGlacier