Group name: Laaroussi Saadeddine

Name: Laaroussi Saadeddine

Email: laar.saad.eddine@gmail.com

Country: Morocco

Specialization: Data science

Project: Bank Marketing (Campaign)

Batch code: LISUM09

Table of contents

Problem description

- ABC Bank wants to sell its term deposit product to customers.
- Before launching the product, they want to develop a model that will help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Business understanding

- ABC Bank wants to use ML (machine learning) model to shortlist customers whose chances of buying the product are higher.
- They want their marketing channel (tele marketing, SMS/email marketing etc) to focus only on those customers whose chances of buying the product are higher.
- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.
- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Project lifecycle along with deadline

Deadline: 30 July

11 June – 18 June: problem description, business understanding, deadline determination, project lifecycle, data intake report.

18 June – 25 June: data exploration for types and problems in data like NA values

25 June – 2 July: Data cleansing and transformation

2 July – 9 July: EDA of data and recommendation

9 July - 16 July: EDA presentation for business users

16 July – 23 July: Model Selection and Model Building

23 July – 30 July: Final Project Report and Code

Data Intake Report

Name: Laâroussi Saâdeddine

Report date: 13-06-2022

Internship Batch: LISUM09

Version: 0.1

Data intake by: Laâroussi Saâdeddine

Data intake reviewer: Laâroussi Saâdeddine

Data storage location: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

Tabular data details:

Bank

Total number of observations	4521
Total number of files	1
Total number of features	17
Base format of the file	Csv
Size of the data	451 Ko

Bank-full

Total number of observations	45211
Total number of files	1
Total number of features	17
Base format of the file	Csv
Size of the data	4 503 Ko

Proposed Approach:

- Cleaning data by checking null values and duplicate values
- Adding columns
- Describing the data and finding correlation between numerical features to search for possible outliers
- Removing outliers
- Analyzing the data:
- Selecting model and making predictions.
- Giving a recommendation in which company to invest

Data understanding

- Features and values

Features	Types	Description	Values	Null ?	Outliers ?
Age	Int64	Age of the person	Between 19 and 95	No	No
Job	Object	Job of the person	['admin.' 'blue-collar' 'entrepreneur' 'housemaid' 'management' 'retired' 'self-employed' 'services' 'student' 'technician' 'unemployed' 'unknown']	No	No
Marital	Object	Marital situation	['divorced' 'married' 'single']	No	No
Education	Object	Education	['primary' 'secondary' 'tertiary' 'unknown']	No	No
Default	Object	Has a default credit	['no' 'yes']	No	No
Balance	Int64	Amout of balance	Between -8019 and 102127	No	Yes
Housing	Object	Has a house	['no' 'yes']	No	No
Loan	Object	Took a loan	['no' 'yes']	No	No
Contact	Object	Was contacted with	['cellular' 'telephone' 'unknown']	No	No
Day	Int64	Number of day in a month	From 1 to 31	No	No
Month	Object	Months of a year	['apr' 'aug' 'dec' 'feb' 'jan' 'jul' 'jun' 'mar' 'may' 'nov' 'oct' 'sep']	No	No
Duration	Int64	Last contact duration, in seconds	Between 0 and 4918	No	Yes
Compaign	Int64	Number of contacts performed for this campaign for this client	Between 1 and 63	No	Yes
Pdays	Int64	Number of days before last contact	Between -1 and 871	No	Yes
Previous	Int64	Number of contacts performed before this compaign for this client	Between 0 and 275	No	Yes
Poutcome	Object	Outcome of the previous marketing campaign	['failure' 'other' 'success' 'unknown']	No	No
Y	Object	Has subscribed or not	['no' 'yes']	No	No

First analysis (Outliers, Skewness, NA values)

Numeric features that might contain outliers are :

- Age
- Balance
- Day
- Duration
- Compaign
- Pdays
- Previous

All features aside from Age are skewed to the right.

There are no NA values in data, however from the data and the skewness, we can see that the null values in many features are either -1 or 0, or 'unknown' for the object values.

Outliers are present in data as well. For example for the feature previous max value 275 while the value before 275 is 58.

Removing outliers with IQR (interquartile range) will help fix the skewness and overall data.

Data cleansing and transformation done on data:

No Null data is in data. However, some values are used as null.

From the data we can assume that :

- Pdays null value is -1
- And previous null value is 0
- Both of these values are connected since the values that have a -1 is pdays have a 0 in previous. their total number is 36 954

Using the describe function to separate numeric value into different categories:

- Age -> age group:
  - <33
  - 33-39
  - 39-48
  - >48
- Balance -> balance group:
  - <=72
  - 72-448
  - 448-1428
  - >1428
- Duration -> duration time:
  - <=103
  - 103-180
  - 180-319
  - >319
- Campaign -> campaign #:
  - 1

- 2
- >3

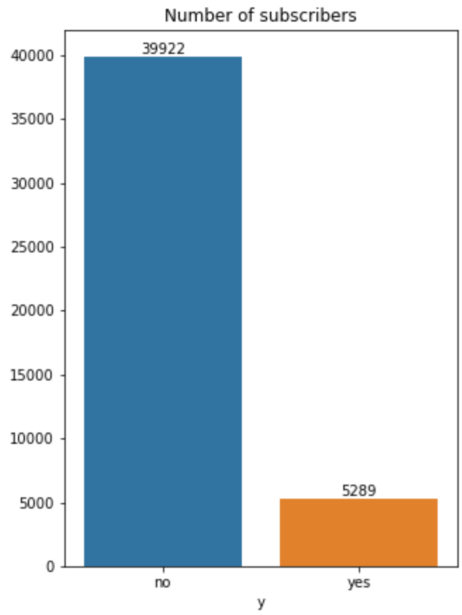
- Pdays and previous -> contacted:
  - No
  - Yes

EDA performed on data:

Various analysis was done on data to find what type of customer to target:

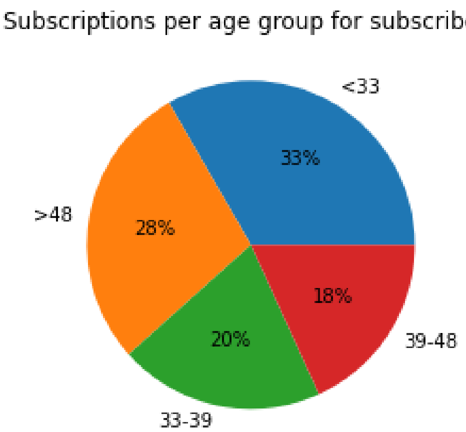
Number of subscribers

- 11% of the customers in the data chose to subscribe.



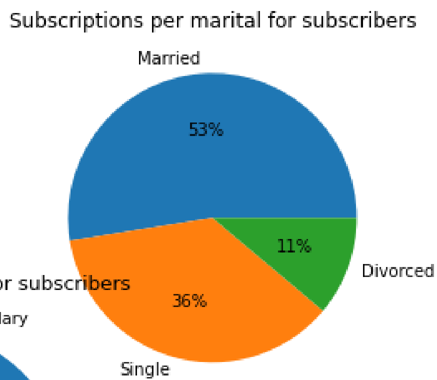
Subscriptions per age group

- 33% of the customers that chose to subscribe are under 33.
- 28% of them are over 48.
- 20% are between 33-39
- 18% are between 39-48



Subscribers per marital status

- 52% of subscribers are married.
- 36% of them are single.
- 11% are divorced.

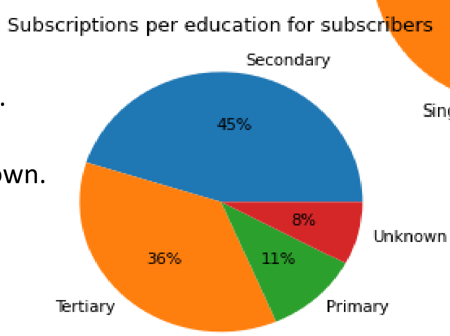


Subscribers per default credit

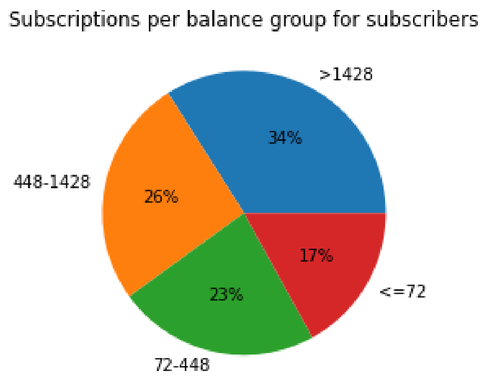
- 99% of subscribers do not have default credit.

Subscribers per education

- 46% of subscribers have secondary education.
- 37% of them have tertiary education.
- 11% have primary education
- 8% of subscribers education is unknown.



Subscribers per balance group



- 34% of subscribers have a balance above 1428.
- 26% of them have a balance between 448-1428.
- 23% have a balance between 72-448.
- 17% have a balance lower or equal than 72.

Subscribers per housing

- 63% of subscribers do not have a house.
- 37 of them have a house.

Subscribers per loan status

- 90% of subscribers do not have a loan status.
- 10% of them have one.

Subscribers per contact mean

- 82% of subscribers were contacted via cellular.
- 10% of them were contact via unknown means.
- 8% were contacted via telephone.

Subscribers per duration time spent

- 63% of subscribers were contacted for a duration over 319 seconds.
- 24% of them for a duration between 180 and 319 seconds.
- 11% of them for a duration between 113 and 180 seconds.
- 2% of them for a duration less than 113 seconds.

Months with most subscribers

- Months with most subscribers are : may, aug and july.

Days with most subscribers

- Day with most subscribers are : 30,12,13,15.

Subscribers per campaign number

- 49% of subscribers were contacted 1 time only during this campaign.
- 26% of them were contacted 2 times.
- 25% were contacted more than 3 times.

Subscribers per contact

- 64% of subscribers were never contacted in any previous campaigns
- 36% were contacted in a previous campaign.

Subscribers per outcome

- 64% of subscribers outcome was unknown.
- 18% of the outcome was considered a success.
- 11% of the outcome was considered a failure.
- 7% of the outcome was classed as other.

Final recommendations

The bank should consider advertising to :

- People that are **under 33**.
- **Married** people.
- Customers that **do not have a default credit**.
- Customers with **at least a secondary education**.
- Customers with **a balance higher than 1428**.
- Customers that **do not own a house**.
- Customers **without a loan**.

The bank should consider contacting their customers via **cellular** and spend **at least 319 seconds** contacting them.

The bank should consider advertising during **the months of May, August, and July**. Either during **the end of the months or the middle of the months**.

The bank should mainly focus on **contacting customers one time** and should **prioritize customers that have never participated in a campaign**.

Model recommendations

- Since the outcome of the model is a yes or no, this can be seen as a classification problem.
- For classification problems, most known methods that can be used are K means, KNN (K nearest neighbor), SVM (Support Vector Machine) or Random forest.
- Some methods can be used for regression and classification problems such as decision trees or neural networks which can also work for this problem.

Models studied

- Decision tree classifier with 92% precision
- Logistic regression with 72% precision
- Random forest classifier with 94% precision
- Cat boost classifier with 94% precision

As predicted the models that use a classifier have a higher precision than the models that use a regressor