**Week 7 submission**

**Project: Bank Marketing Campaign**

Group Name: Evolve Data

Name: Dmitry Sharukhin

Email: sharuhinda@gmail.com

Country: Russia

College/Company: Finval GC

Specialization: Data Science

**Problem description**

ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which 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).

Bank wants to use ML model to shortlist customer whose chances of buying the product are more so that their marketing channel (tele marketing, SMS/email marketing etc.) can focus only to those customers.

This will save resource and their time (which is directly involved in the cost (resource billing)).

The task is to create binary classifier to forecast the probability of customer’s agreement to open term deposit

**Business understanding**

To get an answer about new product bank is going to make a series of calls. Each call has its cost and Bank wants to make only reasonable number of calls to limit corresponding costs.

The results of such calls are considered final. Situations when potential customer changes his mind (for whatever reason) after the call are not considered.

Any client can be contacted several times if it raises chances on positive result.

Some social and economic indicators might be useful (features 16-20) as they reflect general situation that influences the propensity to use banking services.

**Project lifecycle along with deadline (see annex A)**

**Data Intake report (see annex B)**

**Data understanding**

1. There are no obvious missing values but some values mark absence of data (‘unknown’, 999)
2. Raw dataset basically is about half categorical (9 out of 20 features) and half numerical (11 of 20 features), but some features can be converted to Boolean type or one-hot encoded
3. Duplicated rows contain data about different observations (different clients)
4. Assumptions about features are given in Data Intake Report (see below)
5. Features ‘month’ and ‘day\_of\_week’ are given to check if seasonality is presented in target distribution

**What type of data you have got for analysis**

Section: 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' (means divorced or widowed), 'married', 'single', 'unknown')

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')

Section: 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.

Section: current and previous campaigns 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')

Section: social and economic indicators

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)

**What are the problems in the data (number of NA values, outliers, skewed, etc.)**

1. Dataset is highly imbalanced with only 11,3% positive values
2. There are no obvious missing values but some values mark absence of data (‘unknown’, 999)
3. There are 24 duplicated rows in initial dataset. Assume they are not errors.
4. According to description data are related to from May 2008 – November 2010 period. January and February are not represented in dataset nor in 2009, neither in 2010 year. Assume it’s not an error and the campaign was not run during these periods.
5. `campaign` feature has huge right tail with single observations across it
6. `pdays` contains hidden NAN value (999) in 96% of records, thus it is a low-variance feature
7. `previous` feature contains zeros in 86% of records, thus it is also a candidate for low-variance feature. But it should be explored further
8. `poutcome` feature contains 'nonexistent' value in 86% of records, thus it is also candidate for low-variance feature. Should be explored further

**What approaches you are trying to apply on your data set to overcome problems like NA value, outlier, etc. and why?**

1. To overcome imbalance problem training set should be oversampled using any applicable strategy (for example, SMOTE). This will let model to generalize better
2. (1) Some missing values may be imputed based on correlated columns (for example, job and education can be used to impute values to each other using “most frequent” strategy). Have to test correlations further to find another relations OR (2) use some different imputing technique such as ‘most frequent’ strategy, KNN or IterativeImputer class from scikit-learn library
3. Try to train model on dataset with and without duplicate rows if possible
4. Nothing to do here
5. (1) Cutoff tail by replacing outlying values (> Q3+1.5\*IQR) with respective value OR (2) perform binning values
6. (1) Drop column OR (2) perform binning values
7. (1) Drop column OR (2) perform binning values
8. (1) Drop column OR (2) perform binning values

All the mentioned steps require effectiveness comparing. For this purpose, it is necessary to build data preprocessing pipeline. This will ease variants testing and will give the ability to apply the same pipeline to test dataset.

**Github repo link:** https://github.com/sharuhinda/bank\_marketing\_campaign

Annex A

**Project lifecycle**

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Section / Tasks | Deadline | Completion mark |
| **Initiation and planning stage** | | | |
| **1.** | **Week 7** | **Dec 19, 2022** |  |
|  | **Report form: github repo link (1) PDF document** |  |  |
| 1.1. | Problem description |  |  |
| 1.2. | Business understanding |  |  |
| 1.3. | Project lifecycle along with deadline |  |  |
| 1.4. | Data Intake Report |  |  |
|  |  |  |  |
| **Execution stage** | | | |
| **2.** | **Week 8** | **Dec 26, 2022** |  |
|  | **Report form: github repo link (1) PDF document** |  |  |
|  | Problem description |  |  |
|  | Data understanding |  |  |
|  | Data types |  |  |
|  | Data problems (missing values, outliers, skeweness, etc.) |  |  |
|  | What approaches you are trying to apply on your dataset to overcome problems and why? |  |  |
|  |  |  |  |
| **3.** | **Week 9** | **Jan 02, 2023** |  |
|  | **Report form: github repo link (1) PDF document, (2) IPYNB notebook, (3) peers review comments** |  |  |
| 3.1. | Problem description |  |  |
| 3.2. | Data cleansing and transformations done on the data | *Dec 28, 2022* |  |
| 3.3. | Each member should code and review peers work. (Review comment should be present in the github repo) |  |  |
|  | ***NOTES:***  *(1) Each team member should work on different data cleansing approach*  *(2) Try at least 2 techniques to clean the data for NA values: (mean/median/mode/Model based approach to handle NA value/WOE)*  *(3) Try different techniques to identify and handle outliers as well*  *(4) You are allowed to merge the code of each individual and work together to get good result*  *(5) If team decide to not merge the code, then code of each team member should be placed at provided URL (single repository for whole team)* |  |  |
|  |  |  |  |
| **4.** | **Week 10** | **Jan 09, 2023** |  |
|  | **Report form: github repo link (1) PDF document, (2) IPYNB notebook with EDA** |  |  |
| 4.1. | Problem description |  |  |
| 4.2. | EDA performed on the data |  |  |
| 4.3. | Final Recommendations |  |  |
|  |  |  |  |
| **5.** | **Week 11** | **Jan 16, 2023** |  |
|  | **Report form: github repo link (1) PDF document** |  |  |
|  | Problem description |  |  |
|  | EDA presentation for business users |  |  |
|  | Last slide of EDA should be dedicated to technical user which should contain recommended models for this dataset |  |  |
|  |  |  |  |
| **6.** | **Week 12** | **Jan 23, 2023** |  |
|  | **Report form: github repo link** |  |  |
| 6.1. | Select your base model and then explore 1 model of each family (Linear models, Ensemble model, Boosting model, other models if you have time (like stacking)) |  |  |
|  | ***NOTES:***  *(1) Selected model should fit in your business requirement. For example: if your business does not want black box model then select only those models which can be used to explain the prediction*  *(2) You are allowed to merge the code of each individual and work together to get good result*  *(3) If team decide to not merge the code, then upload the code of each team member and other deliverables in the single repo and share the URL of that repo* |  |  |
|  |  |  |  |
| **Closure stage** | | | |
| **7.** | **Week 13** | **Jan 30, 2023** |  |
|  | **Report form: github repo link (1) Report, (2) Power point presentation** |  |  |
| 7.1. | As it was group assignment hence go far a call with your team and discuss the solution of each member and select that solution which is best and is per the requirement |  |  |
|  | ***NOTE:***  *(1) You are allowed to merge the code of each individual and work together to get good result* |  |  |

Note: All PDF reports should contain:

* Team member's details : Group Name (give a name to your group)
* Name
* Email
* Country
* College/Company
* Specialization: Data Science

Annex B

Data Intake Report

Name: Bank Marketing (Campaign)

Report date: Dec 18, 2022

Internship Batch: LISUM15

Version: 1.0

Data intake by: Dmitry Sharukhin

Data intake reviewer:

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

**Tabular data details:**

|  |  |
| --- | --- |
| **Total number of observations** | 41 188 |
| **Total number of files** | 3 |
| **Total number of features** | 21 |
| **Base format of the file** | .csv |
| **Size of the data** | 6.3 MB |

**bank-additional-full.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 41 188 |
| **Total number of files** | 1 |
| **Total number of features** | age (int), no missing values  job (str), no missing values  marital (str), no missing values  education (str), no missing values  default (str), no missing values  housing (str), no missing values  loan (str), no missing values  contact (str), no missing values  month (str), no missing values  day\_of\_week (str), no missing values  duration (int), no missing values  campaign (int), no missing values  pdays (int), no missing values  previous (int), no missing values  poutcome (str), no missing values  emp.var.rate (float), no missing values  cons.price.idx (float), no missing values  cons.conf.idx (float), no missing values  euribor3m (float), no missing values  nr.employed (float), no missing values  y (str), no missing values – target feature  **Total: 21 features** |
| **Base format of the file** | .csv (‘;’-separated) |
| **Size of the data** | 5,8 MB |

**bank-additional.csv**

|  |  |
| --- | --- |
| **Total number of observations** | 4 119 |
| **Total number of files** | 1 |
| **Total number of features** | age (int), no missing values  job (str), no missing values  marital (str), no missing values  education (str), no missing values  default (str), no missing values  housing (str), no missing values  loan (str), no missing values  contact (str), no missing values  month (str), no missing values  day\_of\_week (str), no missing values  duration (int), no missing values  campaign (int), no missing values  pdays (int), no missing values  previous (int), no missing values  poutcome (str), no missing values  emp.var.rate (float), no missing values  cons.price.idx (float), no missing values  cons.conf.idx (float), no missing values  euribor3m (float), no missing values  nr.employed (float), no missing values  y (str), no missing values – target feature  **Total: 21 features** |
| **Base format of the file** | .csv (‘;’-separated) |
| **Size of the data** | 584 KB |

**bank-additional-names.txt**

|  |  |
| --- | --- |
| **Total number of observations** | - |
| **Total number of files** | 1 |
| **Total number of features** | Description of features |
| **Base format of the file** | .txt |
| **Size of the data** | 5 KB |

1. The most of the data is concentrated in file bank-additional-full.csv. Therefore primary key will be the default index (row #).
2. Bank-additional.csv contains 10% sample from full version of the dataset. It’s not indended to be used in project.
3. There are no obvious missing values but some values mark absence of data (‘unknown’, 999)
4. Assume that ‘default’ feature means the presence of the loan in default in any of the banks on the moment of contact. The same applies to ‘housing’ and ‘loan’ features

**Proposed Approach:**

* Use only bank-additional-full.csv as data source. Use bank-additional-names.txt to clarify meanings of features.
* There’s no separate test dataset so we will have to split given data to train and test datasets before performing EDA
* EDA should be performed only on train dataset