Week 7 submission

Group Name: EvolveData
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

Data Intake report (see annex B)

Data understanding

1. There are no obvious missing values but some particular 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. There are no January and February months in dataset. Assume it’s not an error
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. 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
3. Try to train model on dataset with and without duplicate rows
4. Nothing to do here
5. Cutoff tail by replacing outlying values (> Q3+1.5*IQR) with respective value
6. Drop column
7. Test model with and without this column
8. The same

Github repo link:

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	1
Total number of features	21
Base format of the file	.csv
Size of the data	5.8 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

1. As all the data is concentrated in 1 file the primary key will be the default index (row #)

2. There are no obvious missing values but some particular values mark absence of data (‘unknown’, 999)
3. 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:

- 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