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)

Github repo link: https://github.com/sharuhinda/bank_marketing_campaign

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

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