Week 8 Deliverables

Name: Riwaj Neupane
Email: neupaneriwaj64@gmail.com
Country: Nepal
College/Company: NA
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 helps 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).

Data Description:

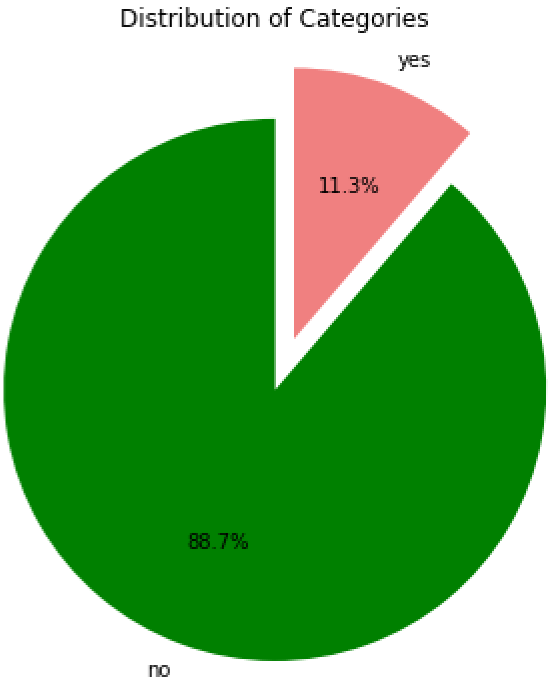
The dataset going to be used for the analysis is called “bank-additional-full.csv”, which contains 41188 observations and 21 features, encompassing features related to clients’ basic information such as age, job, marital status, education, credit in default, housing, and loan; details about contact such as contact communication type, last contact month, last contact day, last contact duration, number of contacts, etc., and information about marketing campaigns like outcome, employment variation rate, consumer price index, consumer confidence index, euribor 3 month rate, and number of employees. We also have the target variable y, which is the answer for the yes-no question “has the client subscribed a term deposit?”, and it will be used in future prediction.

Feature Name	Type	Data Type	Number of unknowns	Number of Outliers	Comments
age	Numerical	int	0	381	
job	Categorical	str	330	0	Replace with mode
martial	Categorical	str	80	0	Replace with mode
education	Categorical	str	1731	0	
default	Categorical	str	8597	0	Leave unknown as its own type
housing	Categorical	str	990	0	Replace with mode

loan	Categorical	str	990	0	Replace with mode
contact	Categorical	str	0	0	
month	Categorical	str	0	0	
day_of_week	Categorical	str	0	0	
duration	Numerical	int	0	861	Replace with upper bound defined as Q3+IQR
campaign	Numerical	int	0	0	
previous	Numerical	int	0	0	
poutcome	Categorical	str	0	0	
emp.var.rate	Numerical	float64	0	0	
cons.price.idx	Numerical	float64	0	0	
cons.conf.idx	Numerical	float64	0	0	
euribor3m	Numerical	float64	0	0	
nr.employed	Numerical	float64	0	0	
y	Categorical	str	0	0	

Problems in the Data (number of NA values, outliers , skewed etc):

There are 6 categorical features with missing data (job, education, marital, default, housing, & loan). There is one numerical feature (“duration”) that contains outlier data. Specifically, we have the mean for “duration” is around 258, but the maximum value is 4918, which indicates the existence of outliers. And in general, the dataset is imbalanced, as the target variable for the predictive classification model skews highly to the “N” case.



Approaches to Overcome These Problems:

In handling missing (NA) values, I can employ a variety of techniques tailored to the severity of each column and its overall impact on the dataset. Dropping the missing data for those features with lower numbers of “unknown” data points (“marital” & “job”). Replacing the missing data with the most frequent category for “housing” and “loan”. And using a ML classification model to fill the missing values for the “default” and “education” features. Techniques such as KNN Imputer to fill the missing data. For the outlier numerical data, as mentioned above, use an upper outer fence defined at 3IQ (upper fence = Q3 + 3*IQR), where IQR is defined as interquartile range, allowing us to retain 97% of the original data.

For the imbalance related to the target variable, we can help account for this imbalance in the model by choosing the correct evaluation metric. For this data set, that most likely will mean using the AUROC curve to help identify which models provide the best results for True Positive and False Negative predictions. Additionally, since the size of the dataset is large enough, we could consider under-sampling from the majority case. Or also do oversampling for matching