Week 8 Deliverable

Group Name: Suyog Nemade

Team Member's Details:

Name: Suyog Nemade
Email: suyognemade005@gmail.com
Country: United States
College/Company: University of Colorado
Specialization: Data Science

Problem Description:

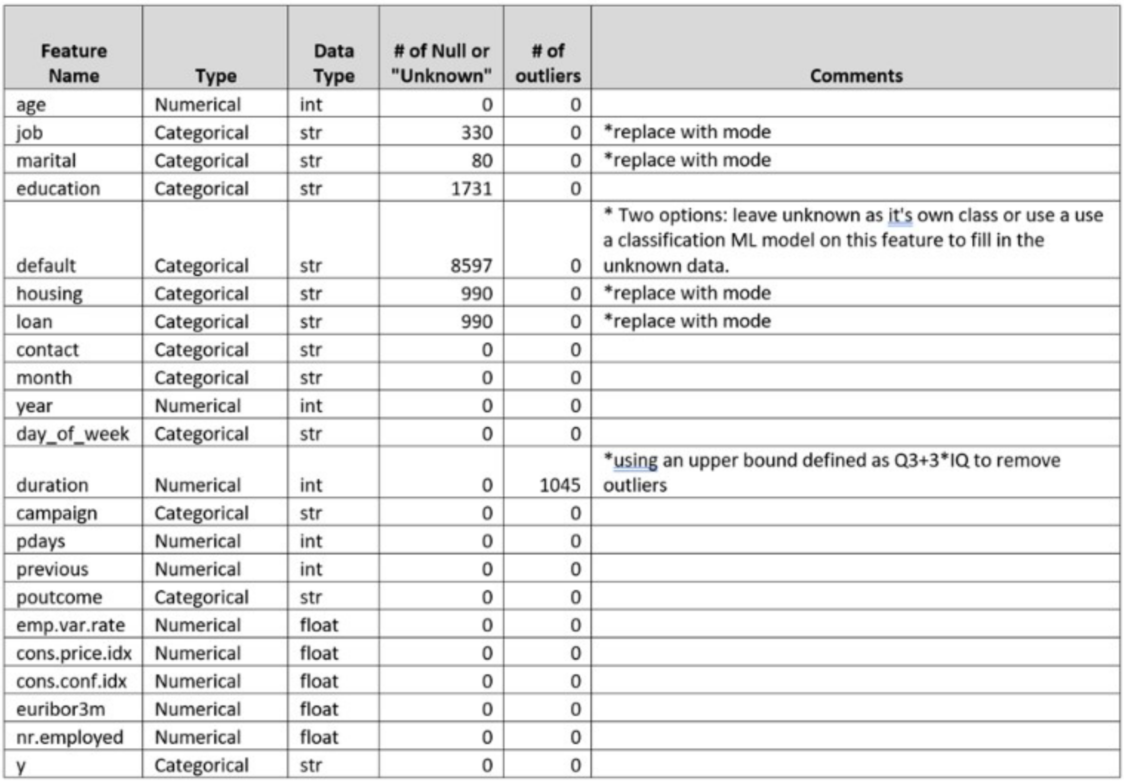
ABC Bank wants to predict whether a customer will subscribe to their term deposit product based on past interactions. They aim to develop a machine learning model to identify customers who are more likely to purchase the product. In other words, they want to shortlist customers whose chances of buying the product is more.

Data Understanding:

The dataset we are going to use 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.

Type of Data for Analysis:

The data set contains a mixture of categorical and numerical data.



Feature Name	Type	Data Type	# of Null or "Unknown"	# of outliers	Comments
age	Numerical	int	0	0	
job	Categorical	str	330	0	*replace with mode
marital	Categorical	str	80	0	*replace with mode
education	Categorical	str	1731	0	
default	Categorical	str	8597	0	* Two options: leave unknown as it's own class or use a use a classification ML model on this feature to fill in the unknown data.
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	
year	Numerical	int	0	0	
day_of_week	Categorical	str	0	0	
duration	Numerical	int	0	1045	*using an upper bound defined as Q3+3*IQ to remove outliers
campaign	Categorical	str	0	0	
pdays	Numerical	int	0	0	
previous	Numerical	int	0	0	
poutcome	Categorical	str	0	0	
emp.var.rate	Numerical	float	0	0	
cons.price.idx	Numerical	float	0	0	
cons.conf.idx	Numerical	float	0	0	
euribor3m	Numerical	float	0	0	
nr.employed	Numerical	float	0	0	
y	Categorical	str	0	0	

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

Six category features—job, education, marital status, default, home, and loan—have missing data. Outlier data may be found in one numerical attribute, "duration." In particular, the greatest value for "duration" is 4918, indicating the presence of outliers, whereas the mean is about 258. Additionally, the target variable for the predictive classification model skews 90% to the "N" case, indicating that the dataset is often unbalanced.

Approaches to Overcome These Problems:

Various approaches will be employed to address missing (NA) values, dependent on the severity of each column and its impact on the overall dataset. deleting the missing data for characteristics ("marital" and "job") that contain less "unknown" data values. replacing the missing data with "loan" and "housing" with the most popular category. Additionally, an ML classification model is used to fill in the missing values for the "default" and "education" variables. As mentioned before, for the outlier numerical data, we may use an upper outer fence set at 3IQ (upper fence = Q3 + 3*IQR), where IQR is the interquartile range. As a result, 97% of the original data will be preserved. We may help account for the imbalance related to the target variable in the model by choosing the right assessment metric. To ascertain which models produce the highest True Positive and False Negative prediction results for this data set, the AUROC curve will most likely be used. Given the size of the dataset, we may also consider under-sampling from the majority scenario. Instead of randomly assigning folds, we may divide the data during training so that the uncommon occurrences are always kept and only randomly separated from the majority case. We may also alter the rare:majority case ratio in the training set to over-represent the exceptional event for the model.