The Data Detectives

Team member's details :

Name: Vincent Weng
Email: vweng@umich.edu
Country: United States
College/Company: University of Michigan
Specialization: Data Science, NLP

Name: Kefan Yu
Email: ky285@georgetown.edu
Country: United States
College/Company: Georgetown University
Specialization: Data Science, Data Analyst

Name: Connor Walker
Email: walkerconnor14@gmail.com
Country: England
College/Company: N/A
Specialization: Data Science

Name: Virginia Mullins
Email: vlmullin@gmail.com
Country: Lebanon
College/Company: N/A
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.

What type of data you have got 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	Drop missing values
marital	Categorical	str	80	0	Drop missing values
education	Categorical	str	1731	0	* Two options: leave unknown as its own class or use a use a classification ML model on this feature to fill in the unknown data.
default	Categorical	str	8597	0	
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	

What are the 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 ~90% to the “N” case.

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

In handling missing (NA) values, we will 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.

For the outlier numerical data, as mentioned above, we can 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, when splitting the data during training, instead of randomizing the folds, we can ensure that the rare cases are kept each time and only randomly split from the majority case. We can also manipulate the ratio of rare:majority cases in the training data to over-represent the rare case for the model.