Week 8 Deliverable
Group Name: Neural Nomad

Name: Nitish Jena
Email: nitishkumar.jena.2023@student.ism.de
Country: Germany
College: ISM Hochschule
Specialization: Data Science

Problem Description:
ABC Bank seeks to develop a predictive model to determine the likelihood of customers subscribing to their term deposit product based on historical interactions. The objective is to leverage machine learning techniques to identify potential customers who exhibit a higher propensity to purchase the product. This will enable the bank to efficiently target and prioritize customers with a greater likelihood of engagement, thereby optimizing marketing efforts and improving conversion rates.

Data Understanding:
The dataset used for this analysis, "bank-additional-full.csv," consists of 41,188 observations and 21 features. These features capture various aspects of customers' demographic and financial profiles, including age, job, marital status, education level, credit default status, housing loan status, and personal loan status.

Additionally, the dataset includes customer interaction details, such as the type of contact communication, the month and day of the last contact, contact duration, and the total number of contacts made. It also provides marketing campaign-related attributes, including campaign outcome, employment variation rate, consumer price index, consumer confidence index, Euribor 3-month rate, and the total number of employees.

The dataset's target variable (y) indicates whether a customer has subscribed to a term deposit (yes/no response). This variable will be used in predictive modeling to help identify potential customers who are more likely to subscribe in future marketing campaigns.

Feature Analysis:

The dataset contains a mix of categorical and numerical data.

Feature Name	Type	Data Type	# of Null or 'Unknown'	# of Outliers	Comments
age	Numerical	int64	0	4	

job	Categorical	object	330	0	*replace with mode
marital	Categorical	object	80	0	*replace with mode
education	Categorical	object	1731	0	*replace with mode
default	Categorical	object	8597	0	* Two options: leave 'unknown' as its own class or use a classification model to fill in missing values.
housing	Categorical	object	990	0	*replace with mode
loan	Categorical	object	990	0	*replace with mode
contact	Categorical	object	0	0	
month	Categorical	object	0	0	
day_of_week	Categorical	object	0	0	
duration	Numerical	int64	0	1043	* Using an upper bound defined as Q3+3*IQR to remove outliers.
campaign	Numerical	int64	0	1094	
pdays	Numerical	int64	0	0	
previous	Numerical	int64	0	5625	
poutcome	Categorical	object	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	object	0	0	

Data Issues (Missing Values, Outliers, and Skewness)

The dataset contains six categorical features with missing values: job, education, marital status, default, housing, and loan. Additionally, there is one numerical feature, "duration," that exhibits significant outliers. The mean value for "duration" is

approximately 258, while the maximum value reaches 4,918, suggesting the presence of extreme outliers.

Furthermore, the dataset is highly imbalanced, with the target variable showing a 90% skew towards the "N" (negative) class, which may impact the predictive performance of classification models.

Approaches to Overcome These Issues:

To handle missing values (NA), we will apply different techniques based on the severity of missing data in each column and its overall impact on the dataset. For features with a lower number of "unknown" values, such as "marital" and "job", we will drop the missing records. For "housing" and "loan", missing values will be imputed using the most frequently occurring category. In the case of "default" and "education", a machine learning classification model will be employed to predict and fill the missing values.

For numerical outliers, we will apply an upper outer fence method using Q3 + 3*IQR (Interquartile Range). This method will help retain approximately 97% of the original data while removing extreme values, ensuring the dataset remains robust.

To address class imbalance in the target variable, we will carefully select the appropriate evaluation metric. Specifically, AUROC (Area Under the Receiver Operating Characteristic Curve) will be used to assess model performance in distinguishing between True Positive and False Negative predictions. Additionally, given the large dataset size, we can implement under-sampling techniques to balance the class distribution.

Furthermore, during data splitting for model training, rather than completely randomizing folds, we will ensure that rare cases are consistently retained while randomly splitting only from the majority class. We can also manipulate the ratio of rare to majority cases in the training data to slightly over-represent the rare class, improving model learning for minority instances.