WEEK 8 DELIVERABLES

Group Name: JSSN

Name	Email	Country	University /Company	Specialization
Mha Luqman Salameh	mahasa179@gmail.com	Jordan	AABU	Data Science
Almudena Zhou Ramírez	almu180@gmail.com	Spain	UNED	Data Science
Shoug Alotaibi	Shouga.1417@gmail.com	Saudi Arabia	KFUPM	Data Science
Peter Okwukogu	peter.okwukogu@gmail.com	Nigeria	Colab Kaduna	Data Science

PROBLEM DESCRIPTION

ABC bank wants to sell term deposits to their customers, but they want to identify particular customers with a higher propensity to buy.

There's a need to identify these customers so they can focus their marketing campaigns efficiently and effectively.

DATA UNDERSTANDING

Here the details of the dataset collected:

1. It has 41,188 observations
2. It has 21 columns
3. 5 of 21 the attributes have numerical data types, 11 of the 21 attributes are objects(strings) data types, and the remaining 5 are floats.

TYPES OF DATA I HAVE FOR ANALYSIS

Attributes Information:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has a housing loan? (categorical: 'no','yes','unknown')

7 - loan: has a personal loan? (categorical: 'no','yes','unknown')

related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark

purposes and should be discarded if the intention is to have a realistic predictive model.

other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

21 - y - has the client subscribed a term deposit? (binary: 'yes','no')

WHAT ARE THE PROBLEMS IN THE DATA

1. Of the 41,188 observations, 12 rows have duplicate data. I am yet to confirm if this is an actual problem. However, I find it unusual.
2. A lot of the columns have 'unknown' values. I have to confirm if these unknown values may skew analysis.
3. The target variable is imbalanced. 89% of the consumers do not buy term deposits. This will affect the model that will learn from it

PROPOSED APPROACHES TO OVERCOME NA VALUES, OUTLIERS, ETC.

1. For duplicate values, if they are redundant, they'll be deleted.
2. A lot of the únknown values' are NA values. If the percentage of NA values are significant, we'll use feature engineering to create columns for them or encode them to fit the existing context
3. Some outliers exist because of variability while others exist because of mistakes during data imputation. Outliers that exist because of variability will be left alone, those that exist because of anomalies will be corrected or removed based on the business understanding

GITHUB REPO LINK

https://github.com/Astharen/DataScientistsGroupProject