WEEK 9: Data Cleansing and Transformation

Name: Sainad Reddy Naini (Individual)
Email: nainisainad@gmail.com
Country: India
College: University of Bath (Graduate -UK)
Specialization: Data Science

1: PROBLEM DESCRIPTION

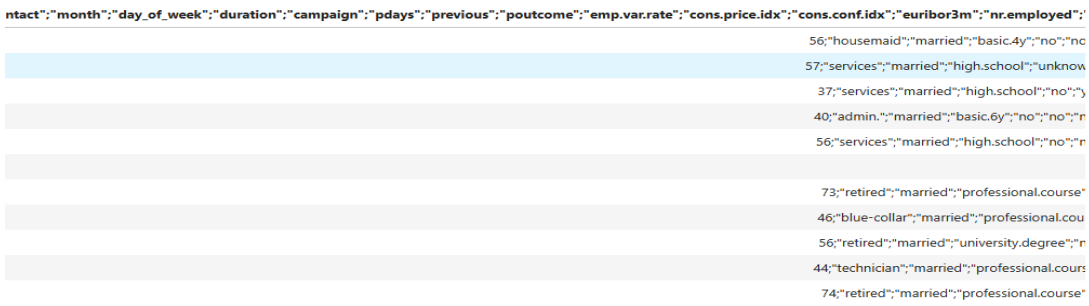
ABC Bank is preparing to launch a new term deposit product and aims to maximize the effectiveness of its marketing efforts. To achieve this, they want to develop a machine learning (ML) model that can predict whether a customer will subscribe to the term deposit based on historical data from previous marketing campaigns. This predictive model will help the bank identify potential customers who are more likely to purchase the product, allowing them to focus their marketing efforts more effectively and efficiently.

Note: Refer to the "Data Understanding" PDF from the previous week (Week 8) to get a rough idea of data problems and general approaches. In this section, we will examine the issues we have addressed and how we dealt with them, including data cleaning and transformation.

2: DATA CLEANSING AND TRANSFORMATION

2.1 Data structure: To address this problem effectively, I implemented the use of a semicolon as the delimiter when reading the file.

Before:



After:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.i
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.9
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.9
2	37	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.9
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.9
4	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.9

2.2 Duplicates: To address this problem effectively, I used the drop_duplicates method, keeping only the first instance.

Before:

```
Number of duplicates in the dataset are: 12
Total number of Rows: 41188
Total number of Columns: 21
```

After:

```
Number of duplicates in the dataset are: 0
Total number of Rows: 41176
Total number of Columns: 21
```

2.3 NA Values: Since there are no null values but only unknown values, I used the following methods to address this issue:

Before:

```
 JOB
known values : 40846
Uknown values : 330
percentage of unknown in that specific column :  0.8 %
 MARITAL
known values : 41096
Uknown values : 80
percentage of unknown in that specific column :  0.19 %
 EDUCATION
known values : 39446
Uknown values : 1730
percentage of unknown in that specific column :  4.2 %
 DEFAULT
known values : 32580
Uknown values : 8596
percentage of unknown in that specific column :  20.88 %
 HOUSING
known values : 40186
Uknown values : 990
percentage of unknown in that specific column :  2.4 %
 LOAN
known values : 40186
Uknown values : 990
percentage of unknown in that specific column :  2.4 %
```

The columns Job, Marital, Education, Default, Housing, and Loan have unknown values.

Dropping the values: The Job and Marital columns have a minimal number of unknown values compared to known values, so I dropped these columns due to their small proportion.

Mode Imputation: For the remaining unknown values, I used mode imputation. Given that the data is relatively simple and the columns in question are categorical, mode imputation is preferred over mean or median, which are more suitable for numerical data. A model-based approach was not used due to the simplicity of the dataset.

After filling all the unknown values:

```
age has no unknown values
job has no unknown values
marital has no unknown values
education has no unknown values
default has no unknown values
housing has no unknown values
loan has no unknown values
contact has no unknown values
month has no unknown values
day_of_week has no unknown values
duration has no unknown values
campaign has no unknown values
pdays has no unknown values
previous has no unknown values
poutcome has no unknown values
emp.var.rate has no unknown values
cons.price.idx has no unknown values
cons.conf.idx has no unknown values
euribor3m has no unknown values
nr.employed has no unknown values
y has no unknown values
```

2.4 Outliers: From the analysis (see the code or previous week's PDF for plots), I can conclude that there are no outliers in the categorical data. However, potential outliers are present in the numeric data, particularly in the columns age, duration, campaign, pdays, and previous.

Before addressing the outliers, it's important to determine whether they are indeed outliers. After examining the potential outliers, here are my conclusions and the methods I used to deal with them:

- **Age**: The age ranges from 17 to 98, which seems reasonable. Therefore, I didn't treat outliers for age, and no data points based on age were removed.

	age
count	40775.000000
mean	39.978541
std	10.401573
min	17.000000
25%	32.000000
50%	38.000000
75%	47.000000
max	98.000000

- **Duration**: Since we are not considering duration in our analysis, we can ignore outliers in this column. However, if you decide to include duration later, please ensure to address any outliers accordingly.
- **Pdays**: The value 999 appears in most rows, indicating that the client was not contacted previously. Removing these values doesn't make sense, so it's best to leave them as is.

pdays	
999	39285
3	431
6	404
4	116
9	64
7	60
2	59
12	58
10	52
5	46
13	35
11	27
1	25
15	24

- **Previous**: Although the values 4 to 7 appear rarely, they reflect valid client histories. These values have been included in the analysis to avoid missing important details that could impact the success of our campaign.

previous	
0	35205
1	4522
2	740
3	214
4	70
5	18
6	5
7	1

- **Campaign**: We are applying capping (Winsorization) by limiting values to a maximum of 15. This helps in reducing the impact of extreme outliers.

campaign	
1	17444
2	10486
3	5294
4	2627
5	1586
6	966
7	622
15	399
8	394
9	277
10	222
11	176
12	124
13	89
14	69

will try to adjust the values of maximum based on the model performance.

2.5 Imbalance: I have observed an imbalance in the target variable y, which could introduce bias into the model. To address this issue, we can use techniques such as sampling methods to balance the data. I will evaluate model performance and adjust the values as necessary once we start training the model. This issue will be dealt with during the training phase.

2.6 Skewness: Similar to how we handle imbalance, we will address skewness only if it impacts model performance. I will analyze the effect of skewness on the model outcome and apply transformations if necessary to improve results.

Note: Please review the source code for detailed comments and an overview of how I addressed these problems and the methods used. We will move on to Exploratory Data Analysis (EDA) next week.

3: GITHUB REPO LINK

https://github.com/sainadreddy/Data_Glacier_Intership_2024