WEEK 8 DELIVERABLES

Name: VIVIAN KERUBO MOSOMI

Email: kerubomosomi7@gmail.com

Country: Kenya

College: Jomo Kenyatta University of Agriculture and Technology (JKUAT)

Specialization: Data Science

Problem Description

ABC Bank wants to predict whether a particular customer will buy their term deposit product based on their past interactions with the bank or other financial institutions. This will allow the bank to focus its marketing efforts on customers who are more likely to purchase the product.

Data Understanding

- The data has 41188 rows and 21 columns

- These are the columns of the dataset:

- Age, Job, Marital, Education, Default, Housing, Loan, Contact, Month, Day of Week, Duration, Campaign, Pdays, Previous, Poutcome, Emloyee Variation Rate, Consumer Price Index, Consumer Confidence Index, Euribor Three Month Rate, Number of Employees and y which is the target variable.

This is the info of the dataset:

The dataset is a dataframe.

Column	Non-Null Count	Dtype
age	41188	int64
job	41188	object
marital	41188	object
education	41188	object
default	41188	object
housing	41188	object
loan	41188	object
contact	41188	object
month	41188	object
day_of_week	41188	object
duration	41188	int64
campaign	41188	int64
pdays	41188	int64

```
  Column      Non-Null Count  Dtype
 previous     41188           int64
 poutcome     41188           object
 emp.var.rate 41188           float64

dtypes: float64(5), int64(5), object (11)
memory usage: 6.6+ MB
```

Data types of the columns in the dataset:

Column Name	DataType
Age	Int64
Job	Object
Marital	Object
Education	Object
Default	Object
Housing	Object
Loan	Object
Contact	Object
Month	Object
Day of Week	Int64
Duration	Int64
Campaign	Int64
Pdays	Int64
Previous	Object
Poutcome	Float64
Employee Variation Rate	Float64
Consumer Price Index	Float64
Consumer Confidence Index	Float64
Euribor Three Month Rate	Float64
Number of Employees	Float64
y	Object

These are the categorical columns and what they represent:

- Job - Represents the job type of the customer which can be blue collar, technician, services, management, entrepreneur, self employed, housemaid, student etc
- Marital - Status of the customer which can be married, single, divorced or unknown
- Education - Level of education of the customer. That is: University degree, HighSchool, Professional Course etc
- Default - Whether the customer has a credit default or not.Can also be unknown
- Housing - If the customer has a housing loan or not.
- Loan - Whether the customer has a personal loan.
- Contact - Type of communication used to contact the customer i.e - Cellular,Telephone

- Month - Last contact month of the year
- day_of_week - Day of the week when the customer when last contacted
- poutcome - Outcome of the previous marketing campaign.It can be failure,success or nonexistent
- y - Has yes or no values which represent whether the customer will purchase the term deposit or not.

These are the numerical columns and what they represent:

- Age - Age of the customer
- Duration - Duration of the last contact with the customer in seconds
- Campaign - Number of contacts performed during the campaign for the customer
- pdays - Number of days since the customer was last contacted in the previous campaign
- previous - Number of previous marketing campaigns in which the customer was contacted before the current campaign.Eg - 4561 customers were contacted during the first previous campaign and so on.
- emp.var.rate - Represents the employee variation rate. Whether the employment rate decreased or increased by the indicated percentage. A positive value indicates an increase in employment levels while negative values indicate a decrease in employment levels.
- cons.price.idx - Consumer Price Index. It's a measure of inflation where higher values indicate inflaction and lower values indicate deflation.
- cons.conf.idx - Consumer Confidence Index. Measures consumer's confidence about the state of the economy. Values closer to zero indicate optimism in their financial situation and the economy's future. Values further from zero indicate pessimism in their financial institution
- euribor3m - Euribor 3-month rate. It represents the average interest rate at which European banks lend unsecured funds to one another for a three-month period. Higher values indicate higher short-term borrowing costs in the Eurozone hence higher market demand for funds and vice versa.
- nr.employed - Number of employees

Problems in the dataset

i)In terms of outliers:

The only column that has outliers is the duration column. And I'll leave them as they are because time taken by each customer varies as some may engage in prolonged discussions while others conclude quickly. Thus, I'll retain the outliers for analysis as they carry valuable information about customer behavior which could influence the outcomes of the predictive models or insights derived from the dataset.

ii) In terms of missing values:

The dataset has no missing values

iii) In terms of skewness:

The overall skewness is 1.76 hence positive skewness. Skewness can cause the following problems in our analysis:

- Overfitting of our model
- Underfitting
- Bias in predictions leading to poor generalization
- Class Imbalance
- Can affect model training

I’ll have to solve the skewness to avoid encountering the above. This can be done using log transformation, feature engineering, scaling etc.

GitHub Repository Link

https://github.com/Vee2002/DataGlacier_Internship/tree/vc/Data%20Glacier