Exploratory Data Analysis

Bank Marketing Campaign

Aditi Dadariya - July 15th 2023

Agenda

Executive Summary

Problem Statement

Data Details

Exploratory Data Analysis Approach

Exploratory Data Analysis Summary

Recommendations

References

Data Glacier

Your Deep Learning Partner

Executive Summary

Company: ABC Company

Interested in: Selling bank term deposits

Strategy: Marketing Campaign

Location: Portugal

Problem Statement

ABC Bank aims to introduce a new term deposit product and seeks to develop a model that can predict whether a customer is likely to purchase the product. This prediction will be based on the customer's previous interactions with the bank or other financial institutions. The purpose of this model is to provide valuable insights to ABC Bank, enabling them to effectively target potential customers and optimize their marketing strategies for the new product.

Data Details

The dataset is downloaded from Kaggle [1]. The dataset provided, known as the Bank Marketing Campaign dataset, is specifically designed for a supervised classification problem. Its main objective is to predict whether a customer will subscribe to the bank's term deposit product.

The dataset is available in two versions: "bank-additional-full.csv," which contains the complete dataset spanning from May 2008 to November 2010, and "bank-additional.csv," which is a subset of the full dataset comprising only 10% (i.e., 4,119) randomly selected examples.

Exploratory Data Analysis Approach

1. Understanding the Data
2. Data Cleaning
3. Data Visualization

Understandi-ng Data

In total, the dataset consists of 41,188 records and 21 variables. These variables include 'age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', 'nr.employed', and 'y'. The target variable is denoted by 'y', while the remaining variables are considered as features.

The data types of the variables vary. 'age', 'duration', 'campaign', 'pdays', and 'previous' are of the int64 data type, while 'emp.var.rate', 'cons.price.idx', 'cons.conf.idx', 'euribor3m', and 'nr.employed' are of the float64 data type. The variables 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'poutcome', and 'y' are of the object data type.

Data Cleaning

- **Missing Values:** In some categorical attributes, there are instances of missing values denoted by the label "unknown". The decision was made not to drop these labels due to their substantial presence in the dataset. Removing these values could potentially impact the integrity of the training data. Hence, these labels can be approached in different ways by first replacing them into null values and later an imputation techniques (i.e. SimpleImpute) has been applied to handle these missing values.
- **Inconsistent datatypes:** Once the "unknown" values were transformed into NaN values, the categorical variables ('job', 'marital', 'education', 'default', 'housing', 'loan') were examined for potential mixed datatypes. However, it was confirmed that these variables do not exhibit multiple datatypes, as they are consistently of the same datatype throughout the dataset.
- **Encoding the categorical variables:** Prior to converting the categorical variables into numeric values, an assessment was conducted to determine whether each variable falls under binary categorical or multi-categorical classification. Binary categorical variables were encoded using Label Encoder, while multi-categorical variables were encoded using One-Hot Encoder.
- **Outliers Detection:** Outliers were identified using two methods: the Interquartile Range (IQR) method and Z-score. In order to minimize excessive changes to the dataset, a quantile imputation technique was employed to replace the outliers.
- **Imbalance Dataset:** The dataset is imbalance as the target variables has 36,548 values for the category 'no' and 4,640 values for the category 'yes'. To address this issue, the SMOTE (Synthetic Minority Over-sampling Technique) method was employed.

	AGE	JOB	MARITAL	EDUCATION	DEFAULT	HOUSING	LOAN	CONTACT	MONTH	DAY_OF_WEEK	...	CAMPAIGN	PDAYS	PREVIOUS	POUTCOME	EMP.VAR.RATE	CONS.PRICE.IDX	CONS.CONF.IDX	EURIBOR3M	NR.EMPLOYED	Y
0	56	0.0	0.0	1.0	0	0	0	1	0.0	0.0	...	1.0	999.0	0.0	0.0	1.1	93.994	-36.4	4.857	5191.0	0
1	57	0.0	0.0	0.0	0	0	0	1	0.0	0.0	...	1.0	999.0	0.0	0.0	1.1	93.994	-36.4	4.857	5191.0	0

Fig 1: Few records of Dataset

Outlier Detection

```
469 number of row are Outliers in column AGE
4612 number of row are Outliers in column MARITAL
4176 number of row are Outliers in column EDUCATION
3 number of row are Outliers in column DEFAULT
6248 number of row are Outliers in column LOAN
2632 number of row are Outliers in column MONTH
7827 number of row are Outliers in column DAY_OF_WEEK
2963 number of row are Outliers in column DURATION
2406 number of row are Outliers in column CAMPAIGN
1515 number of row are Outliers in column PDAYS
5625 number of row are Outliers in column PREVIOUS
4252 number of row are Outliers in column POUTCOME
447 number of row are Outliers in column CONS.CONF.IDX
4640 number of row are Outliers in column Y
```

Fig 1: Outliers detected by Interquantile Range method

```
2632 number of row are Outliers in column MONTH
861 number of row are Outliers in column DURATION
869 number of row are Outliers in column CAMPAIGN
1515 number of row are Outliers in column PDAYS
1064 number of row are Outliers in column PREVIOUS
```

Fig 2: Outliers Detected by ZScore

Data Visualization

Data Glacier
Your Deep Learning Partner

1. Correlation between data variables:

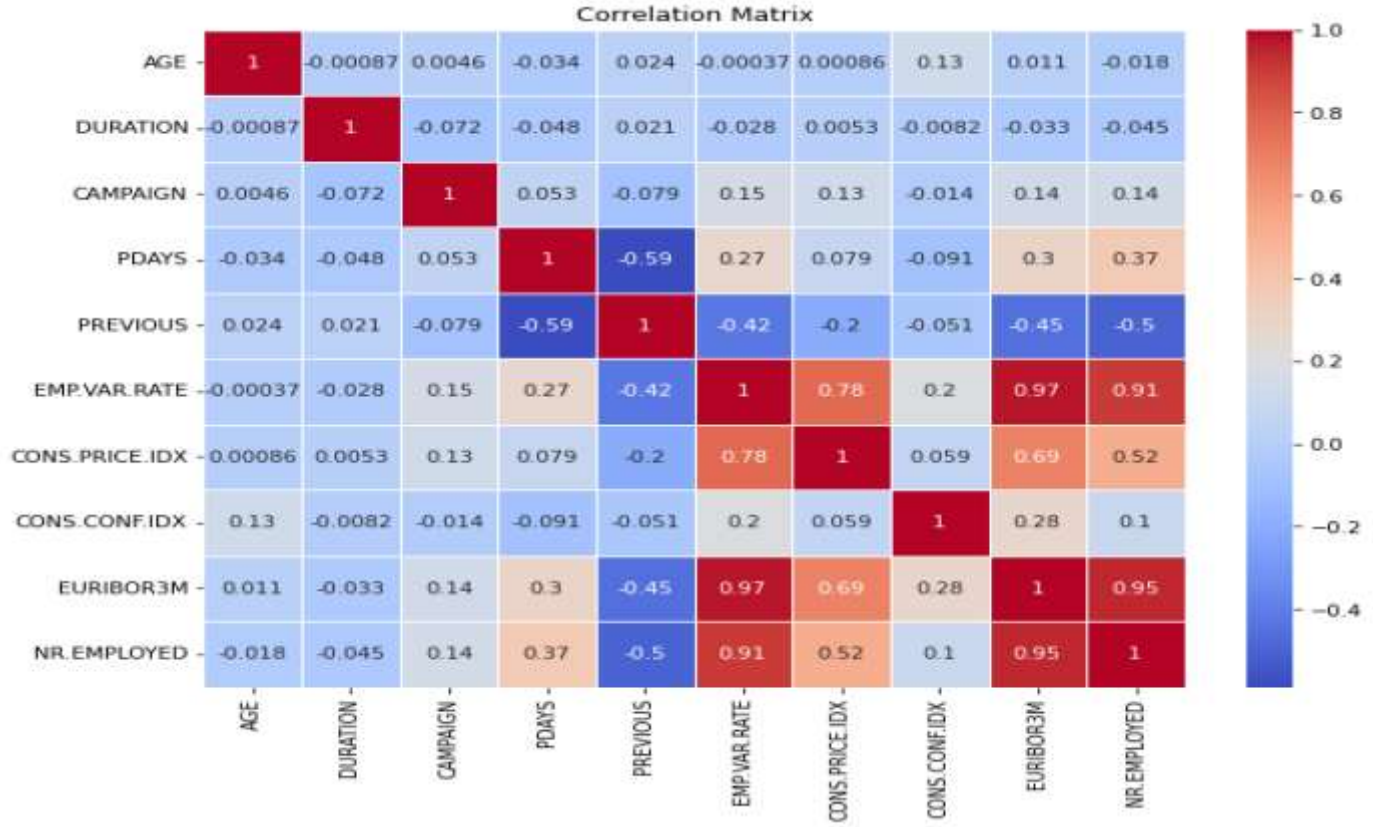

Fig 3: Heat map

As depicted in Fig 3, there seems to be a

- Correlation between EMP.VAR.RATE and CONS.PRICE.IDX: 0.7753
- Correlation between EMP.VAR.RATE and EURIBOR3M: 0.9722
- Correlation between EMP.VAR.RATE and NR.EMPLOYED: 0.9070
- Correlation between CONS.PRICE.IDX and EURIBOR3M: 0.6882
- Correlation between CONS.PRICE.IDX and NR.EMPLOYED: 0.5220
- Correlation between EURIBOR3M and NR.EMPLOYED: 0.945

Data Visualization - cont

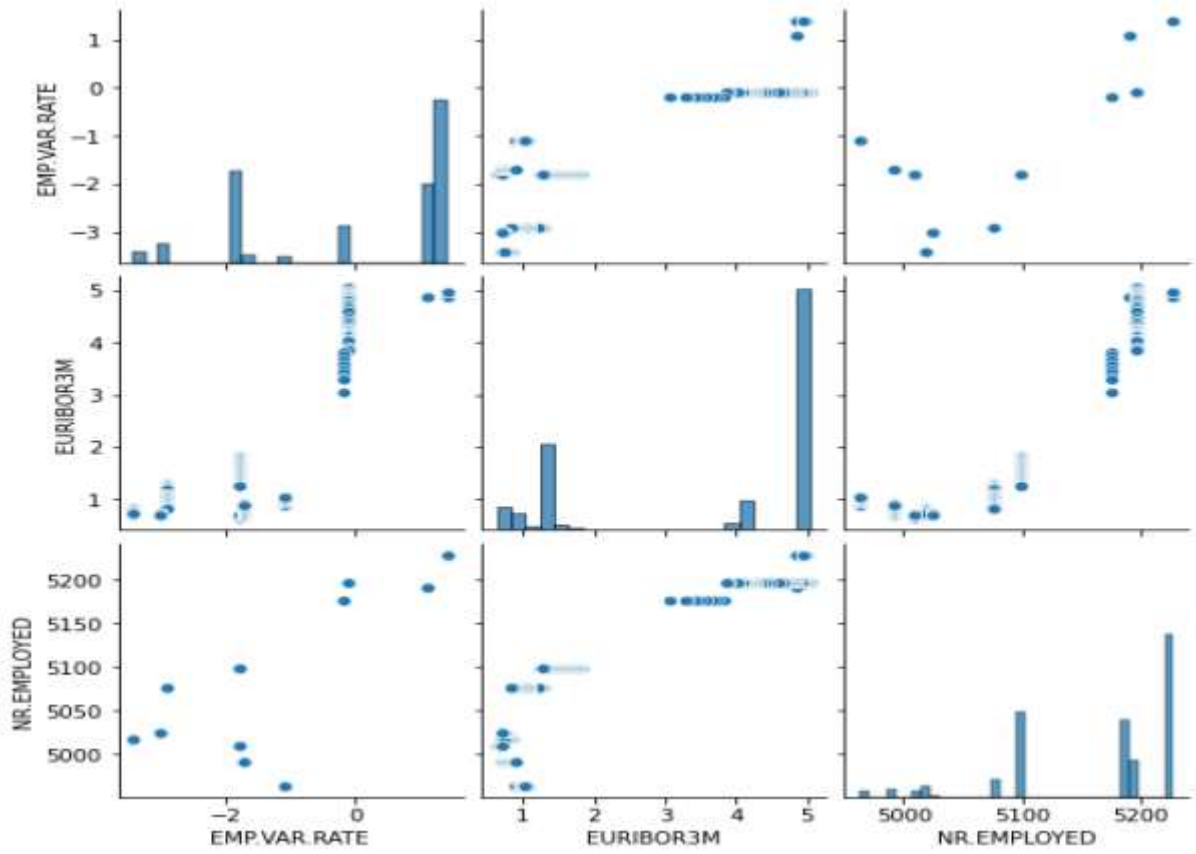

Fig 4: Pair Plot to

These findings suggest a potential relationship among EMP.VAR.RATE, EURIBOR3M and NR.EMPLOYED variables. To further illustrate this relationship, a Pair plot is showcased in Fig 4 above.

Data Visualization - cont

Data Glacier
Your Deep Learning Partner

2. **Understanding each variable individually:** The unique values of categorical variables were explored and visualized using bar plots. The numerical variables were analyzed by calculating their mean and standard deviation, providing insights into their distribution. Please refer to Figures 3 to 23 for the detailed visualizations

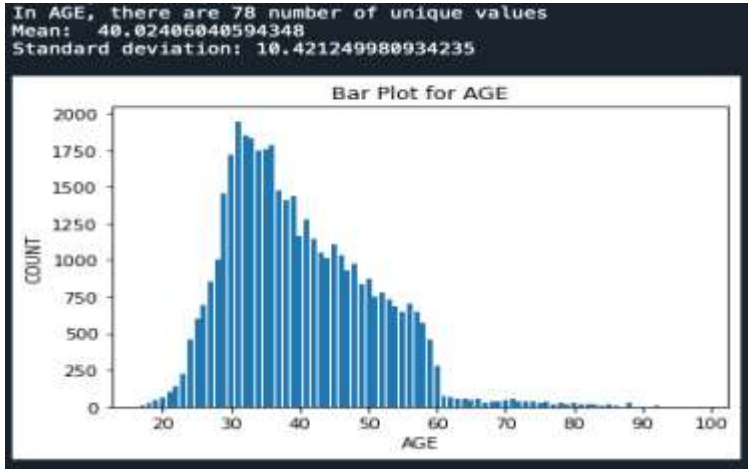

Fig 5: Age variable

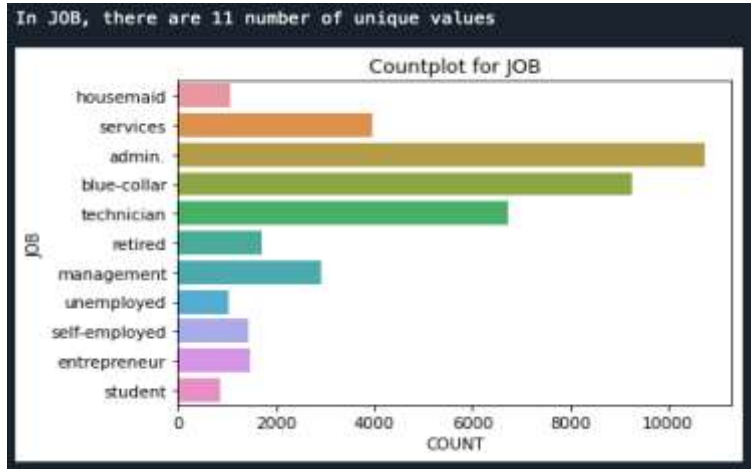

Fig 6: Job variable

Fig 5 shows that the age range of customers who are likely to be eligible for a term deposit is between 20 and 60 years. Among the various job categories, administrative jobs have the highest representation, as depicted in Fig 6.

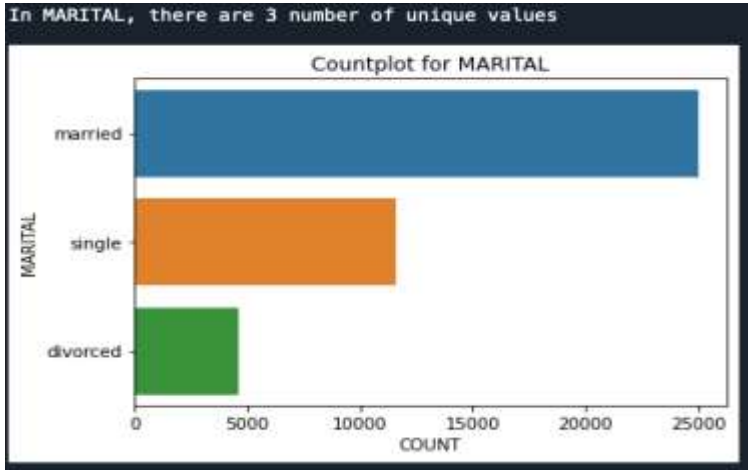

Fig 7: Marital Variable

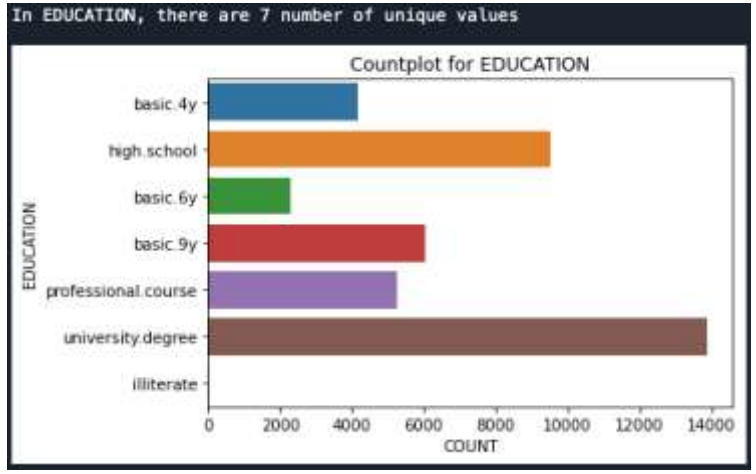

Fig 8: Education variable

Fig 7 indicates that the number of married customers surpasses that of single and divorced customers, suggesting their eligibility for term deposit offers. Fig 8 showcases the presence of university degree holders who are potentially targeted for term deposit contact.

Data Visualization - cont

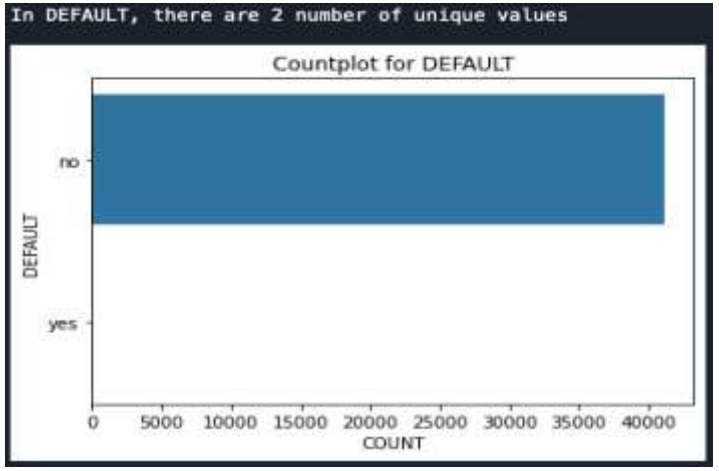

Fig 9: Default variable

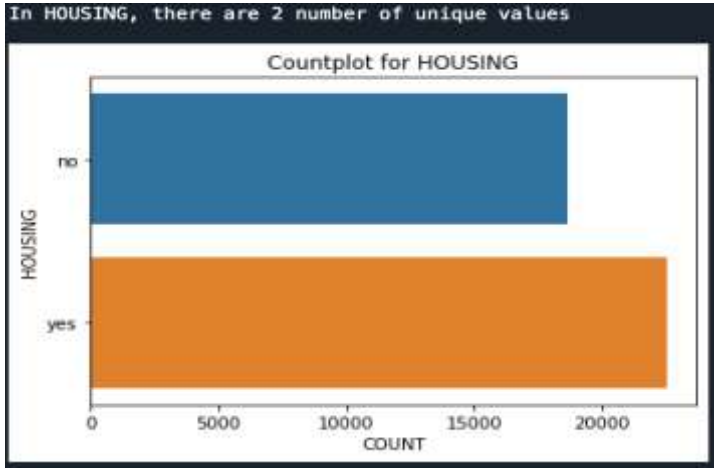

Fig 10: Housing variable

Fig 9 reveals that the majority of customers do not have any credit default. Fig 10 illustrates that the number of customers with housing loans is greater than the number of customers without housing loans.

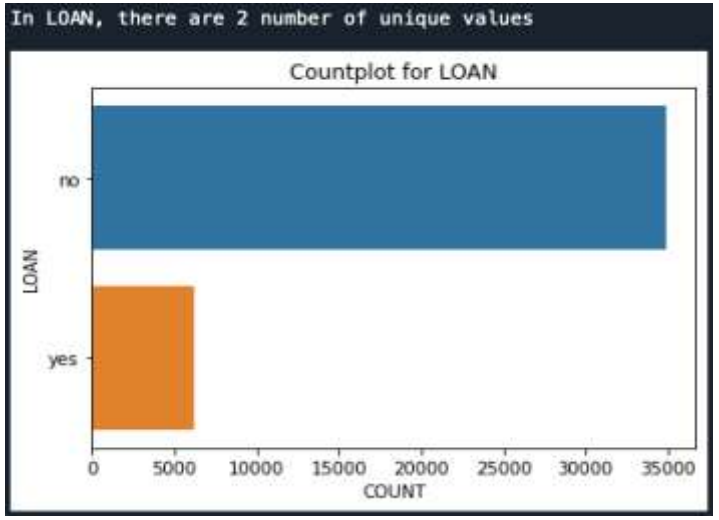

Fig 11: Loan variable

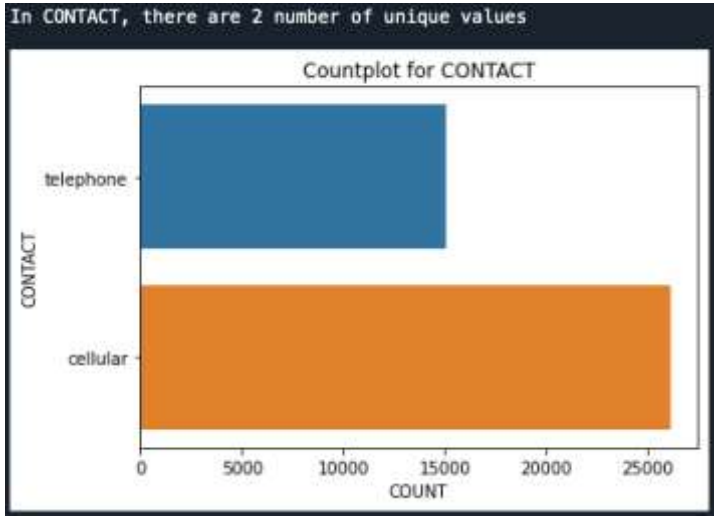

Fig 12: Contact variable

As shown in Fig 11, the count of customers with personal loans is significantly lower than the count of customers without personal loans. As depicted in Fig 12, the count of customers with cellular phones is higher compared to customers with telephone connections. December

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

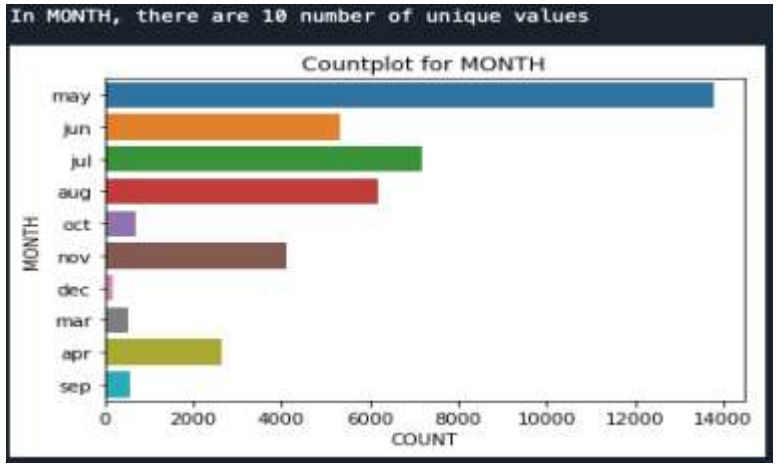

Fig 13: Month Variable

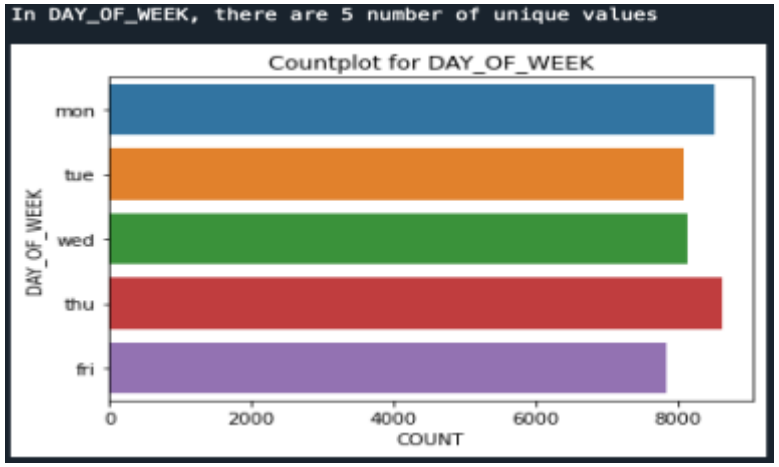

Fig 14: Day_of_Week

As shown in Fig 13, the majority of customers were contacted during the month of May in the past. As depicted in Fig 14, the number of customers contacted during the weekdays shows a slight increase on Mondays and Thursdays, although the overall impact is not substantial.

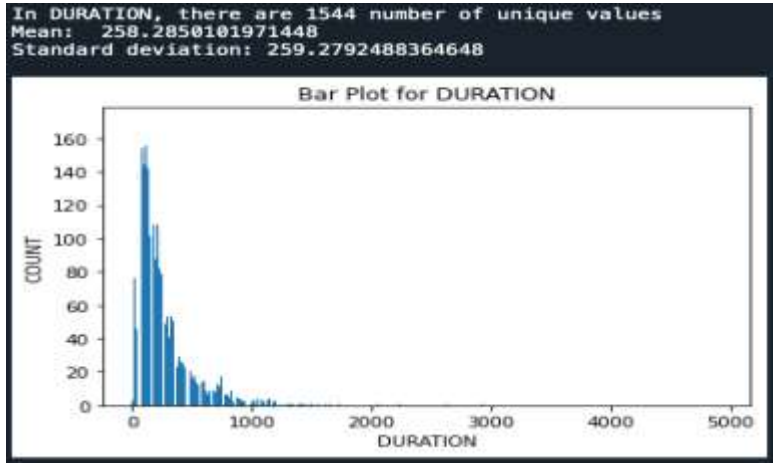

Fig 15: Duration variable

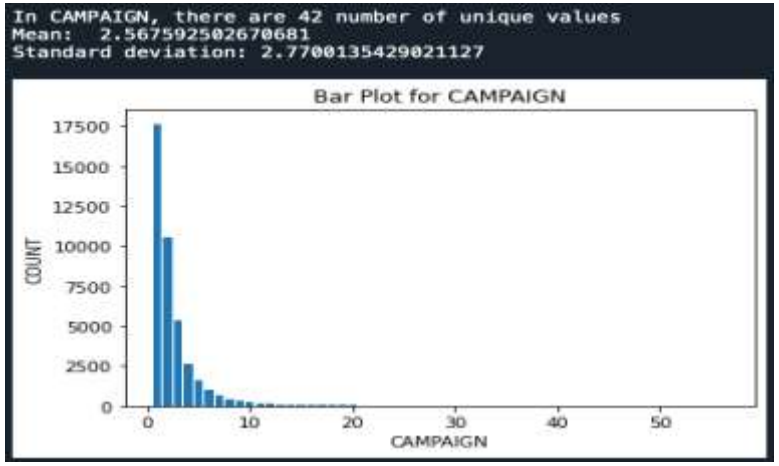

Fig 16: Campaign variable

As shown in Fig 15, the duration of calls ranged from 0 to 1000 seconds. On an average, the calls lasted for approximately 258.28 seconds with the customers. As shown in Fig 16, the number of contacts made during the campaign ranged from 1 to 10, with a significant proportion of customers being contacted only once.

Data Visualization - cont

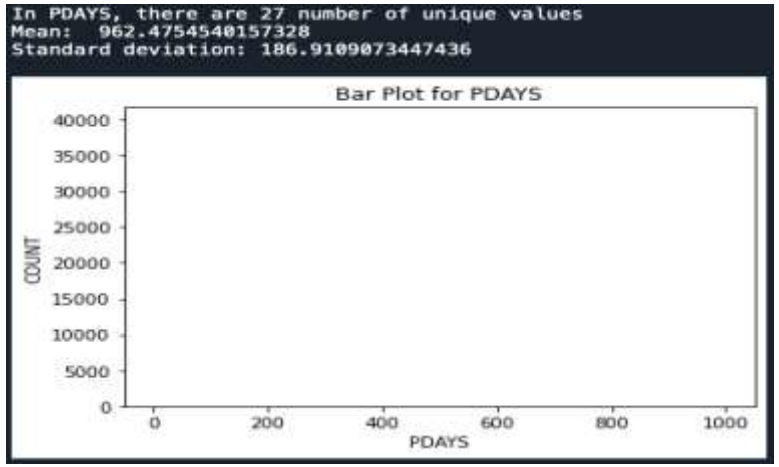

Fig 17: PDays variable

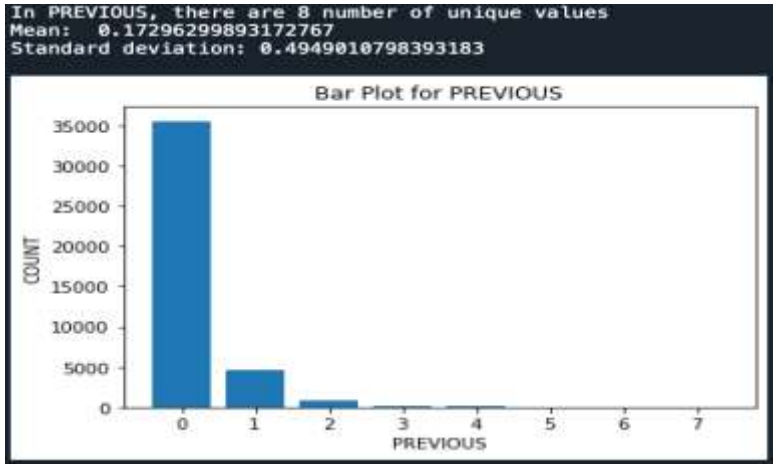

Fig 18: Previous variable

As depicted in Fig 17, the average number of times customers were contacted before the campaign was relatively low. As shown in Fig 18, the number of calls made to customers prior to the campaign is relatively low. Only a small number of customers have been contacted once or twice before the campaign

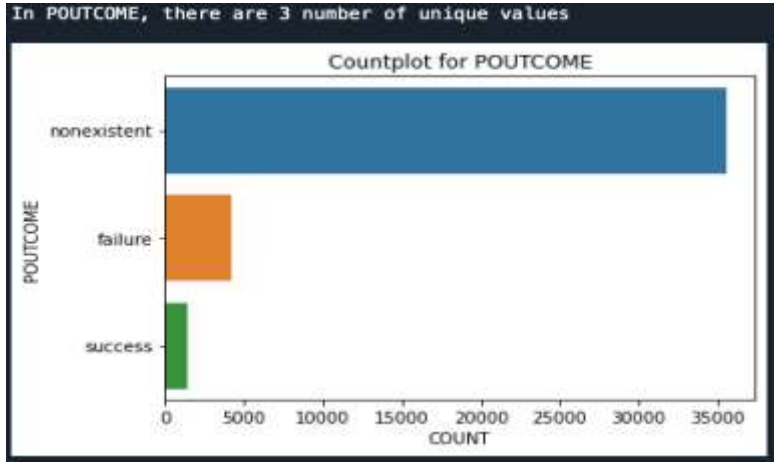

Fig 19: POutcome variable

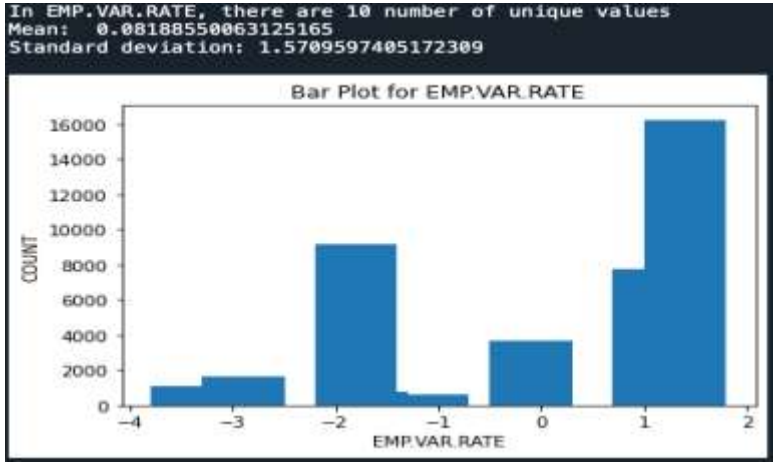

Fig 20: EMP.VAR.RATE variable

Based on Fig 19, the majority of the previous marketing campaign outcomes are non-existent, with a smaller number of successes and failures. Based on Fig 20, the employment variation rate is highest between 1 and 2, with an average of 0.0818.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

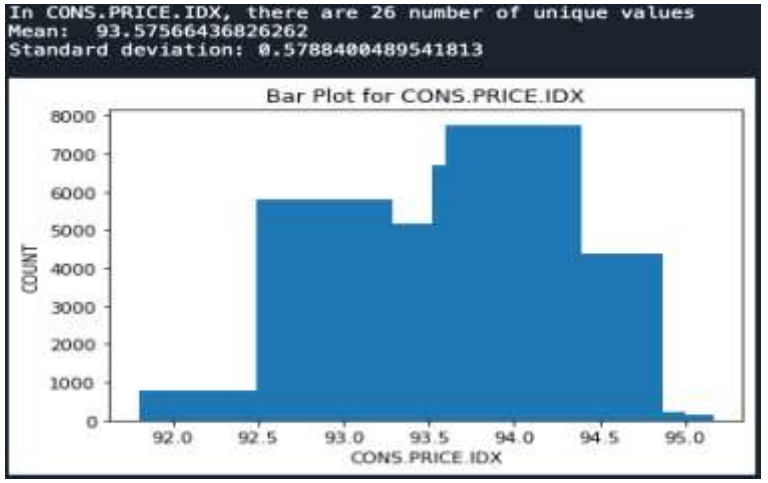

Fig 21: CONS.PRICE.IDX variable

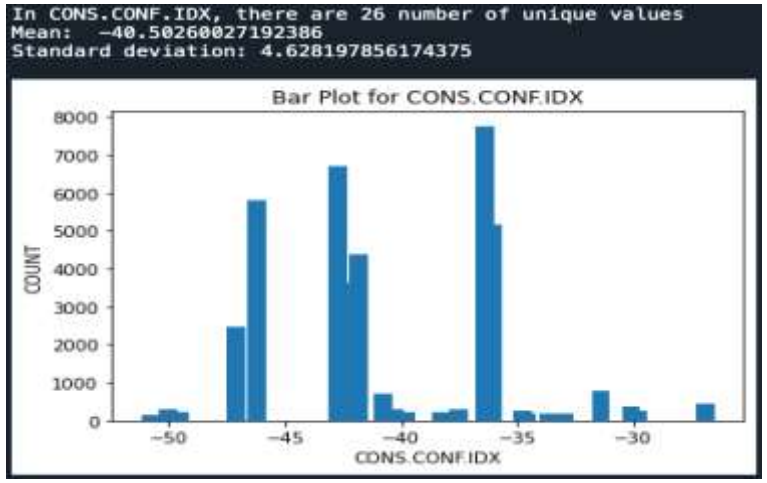

Fig 22: CONS.CONF.IDX variable

According to Fig 21, the average customer price index is 93.57, with a range between 92.5 and 94.5. Fig 22 indicates that the customer confidence index ranges from -35 to -38, with an average value of -40.50

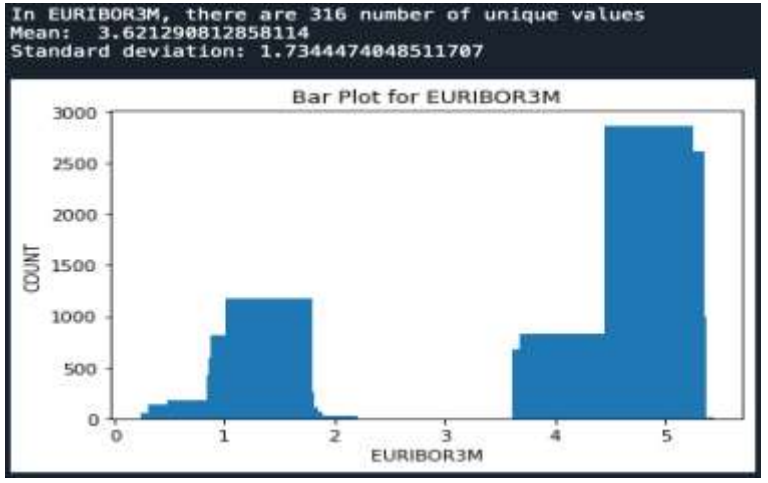

Fig 23: EURIBOR3M variable

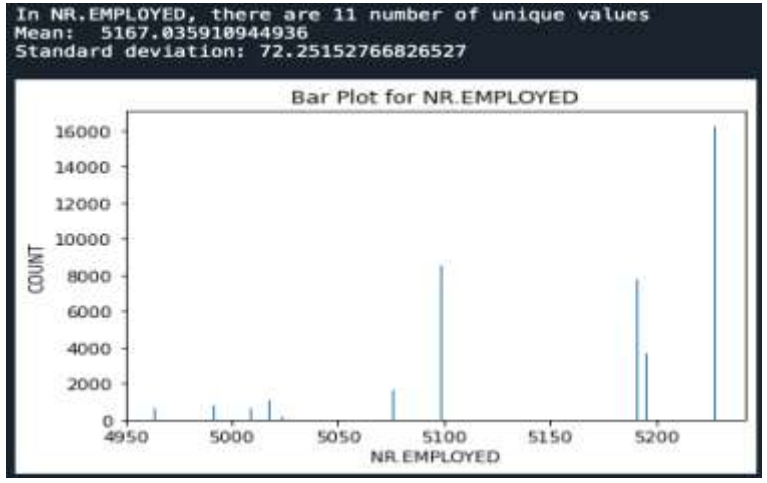

Fig 24: NR.Employed variable

Fig 23 shows that the maximum Euribor 3 month rate falls within the range of 4 to 5. Fig 24 illustrates that the average number of employees is 5167, with a maximum exceeding 5200.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

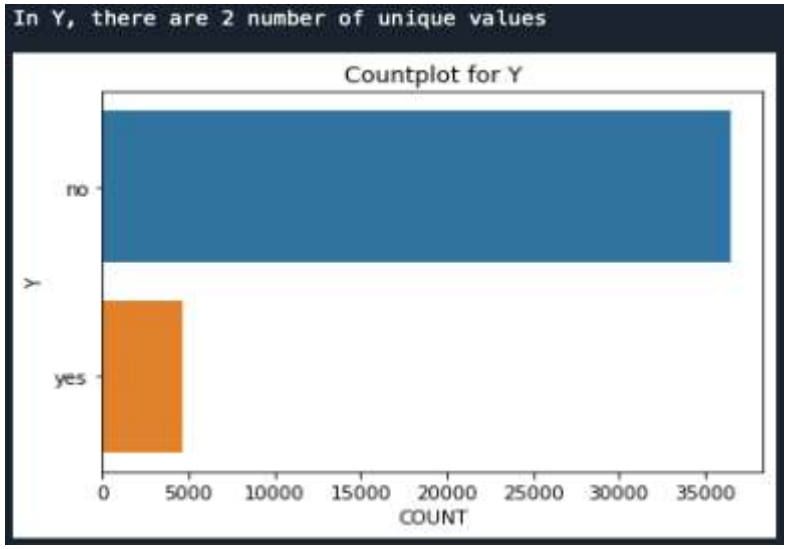

Fig 25: Y target variable

According to Fig 25, it reveals that a small number of customers already have a term deposit.

Data Visualization - cont

3. **Relation of categorical variable with target variable:** Below are the plots that illustrate the relationship between categorical variable and the target

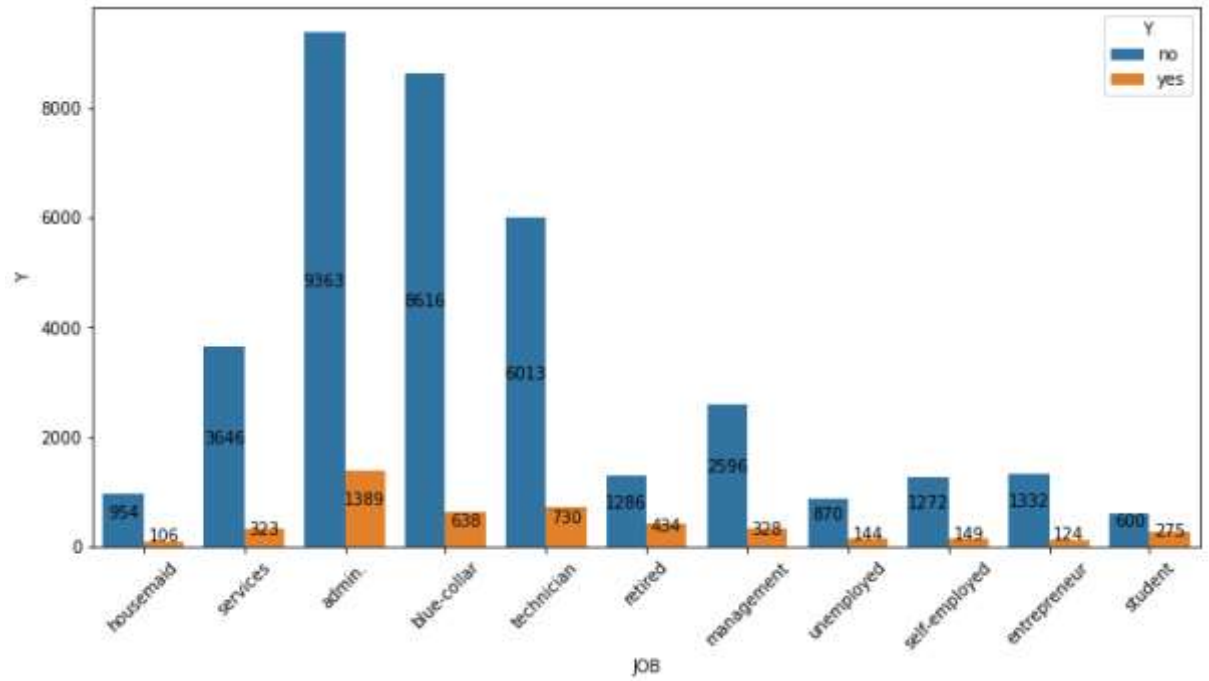

Fig 26: Job versus Y

Fig 26 demonstrates that the highest number of customers with a term deposit are employed in administrative roles. 1389 number of administrator customers has already taking the term deposit.

Data Visualization - cont

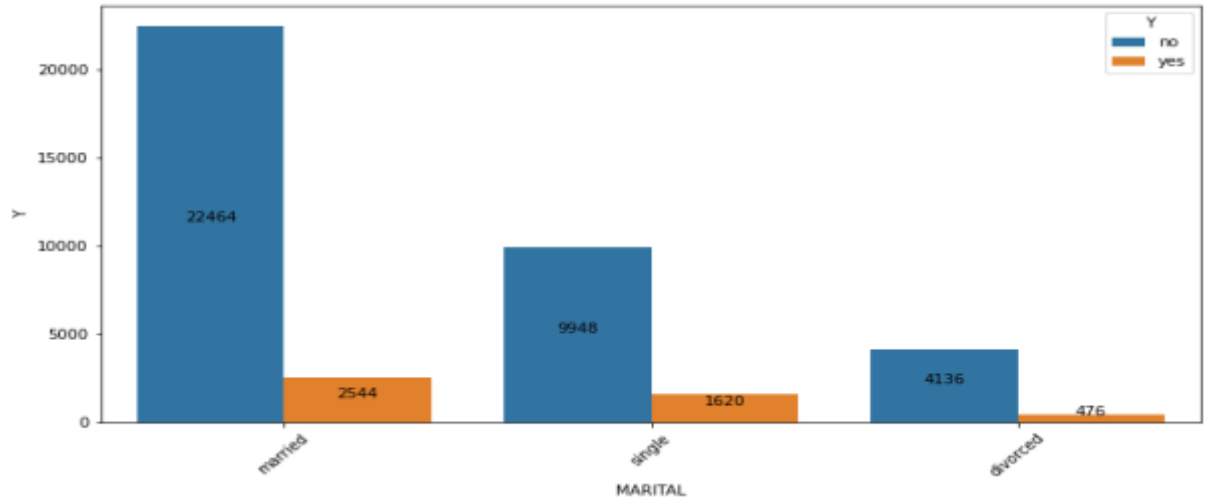

Fig 27: Marital versus Y

Fig 27 illustrates that the highest number (i.e. 2544) of customers with a term deposit are married individuals.

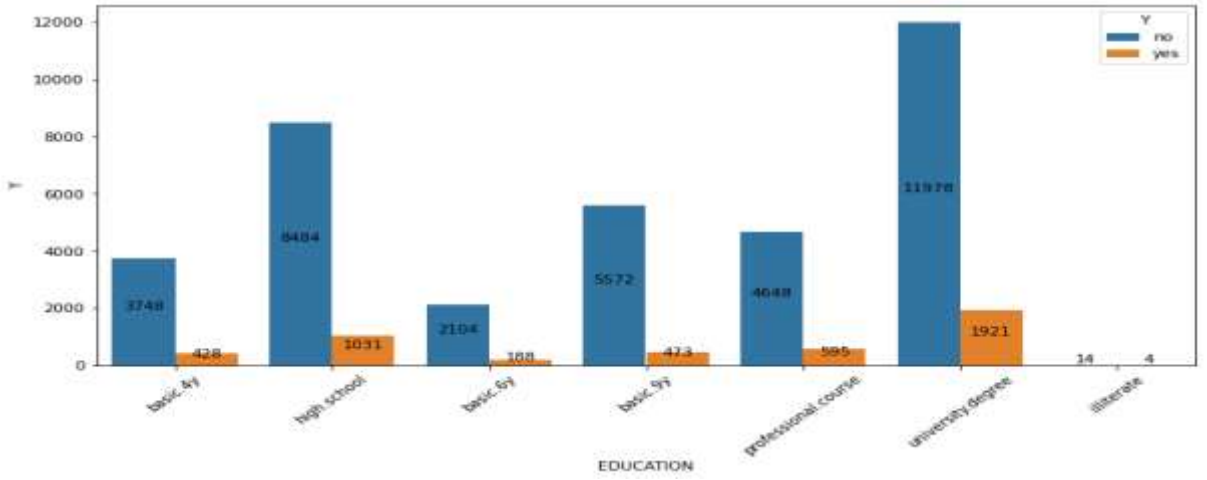

Fig 28: Education versus Y

As shown in Fig 28, the majority of customers (i.e. 1921) with a term deposit are those who hold a university degree.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

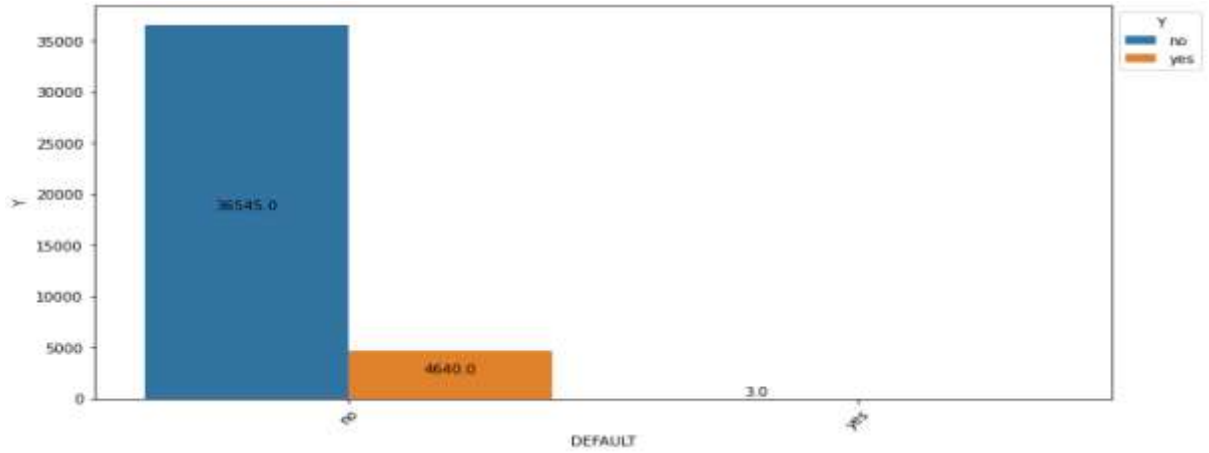

Fig 29: Default versus Y

According to Fig 29, there is a total of 4640 customers who have had a term deposit in the past.

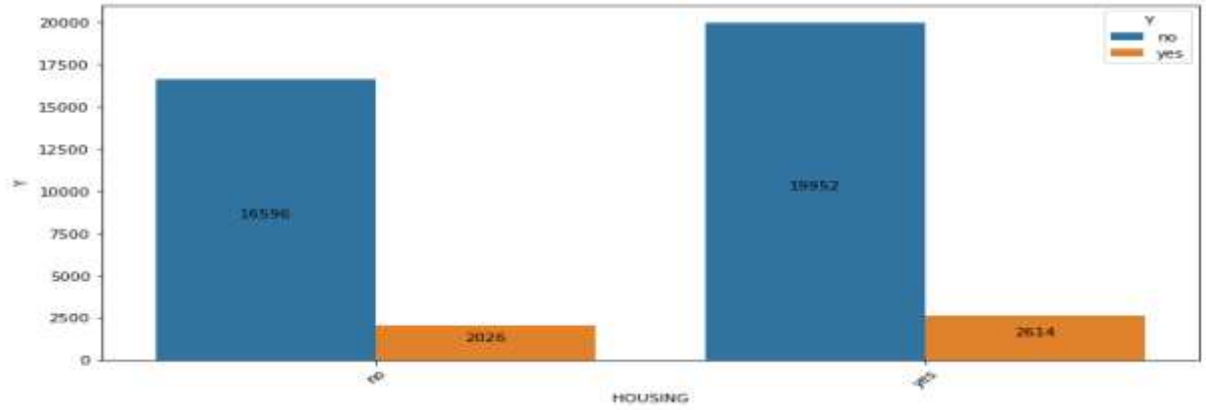

Fig 30: Housing versus Y

Based on the information shown in Fig 30, there is a relatively similar proportion between housing loan holders who have a term deposit and those who do not have a term deposit.

Data Visualization - cont

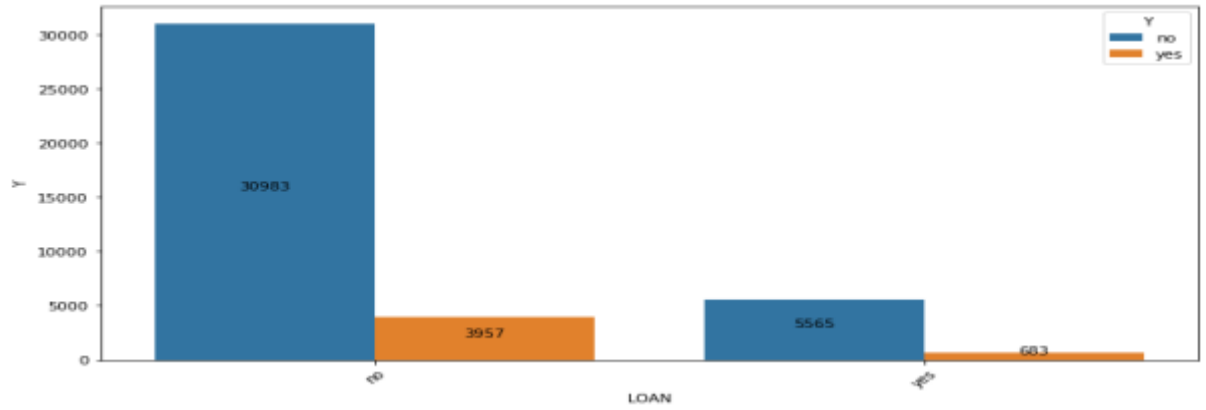

Fig 31: Loan versus Y

Based on the data presented in Fig 31, it can be observed that the number of term deposits for customers without a personal loan is significantly higher compared to those with a personal loan. deposit in the past.

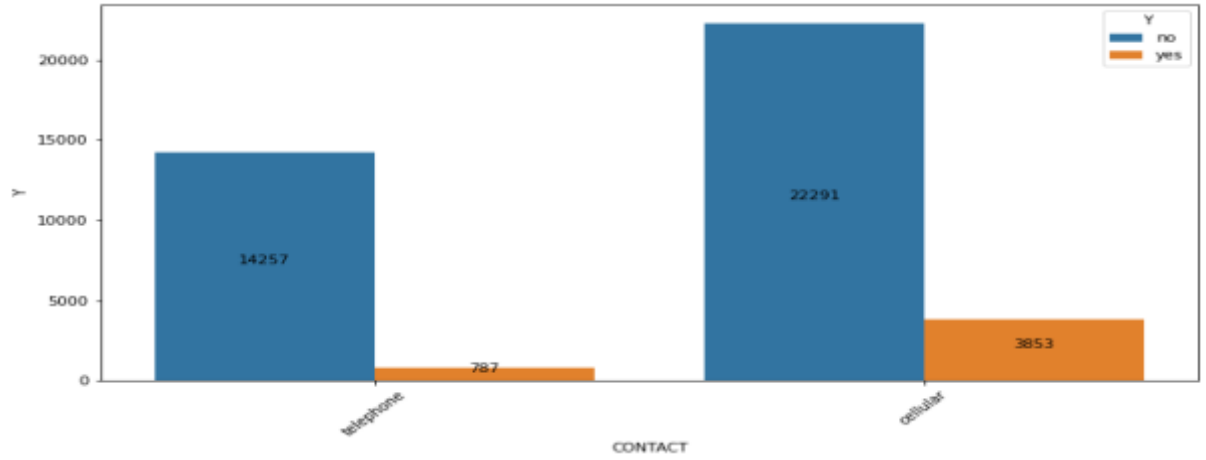

Fig 32: Contact versus Y

Based on the information depicted in Fig 32, it can be observed that the majority of customers who have opted for a term deposit have a cellular phone.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

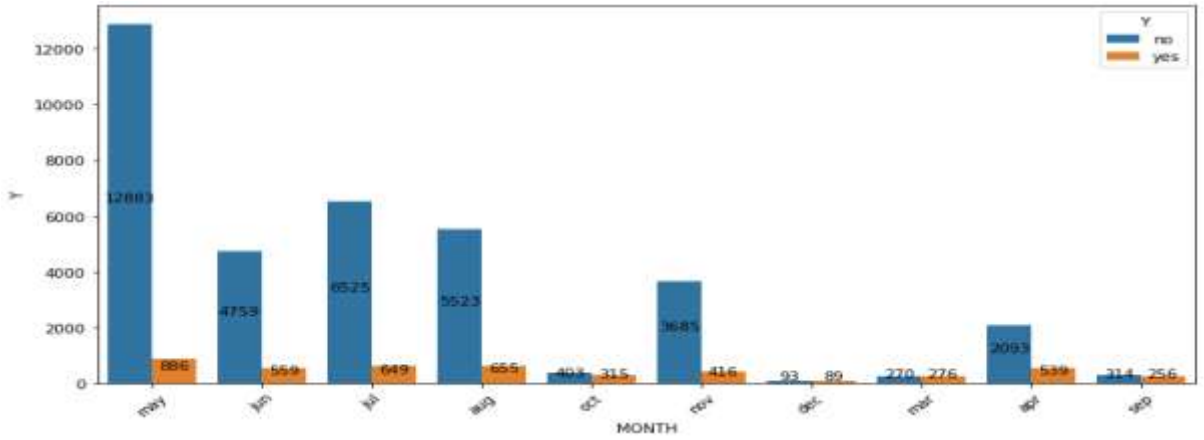

Fig 33: Month versus Y

Based on the data visualized in Fig 33, it is evident that the majority of customers were contacted during the month of May. Furthermore, this period also witnessed the highest number of conversions to a term deposit.

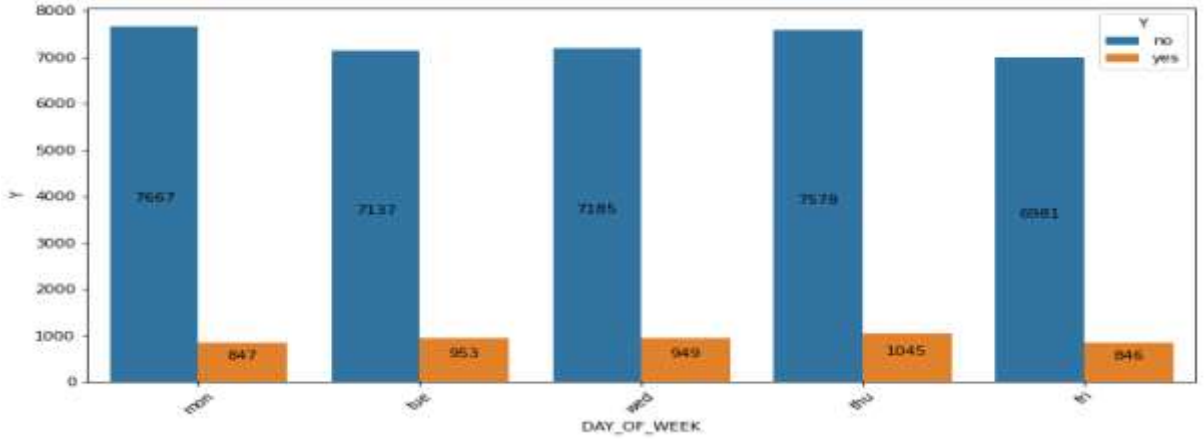

Fig 34: Day_of Week versus Y

Based on the information presented in Fig 34, it can be observed that the majority of customers were contacted on Monday and Thursday. Additionally, there are only 1,045 customers who have a term deposit.

Data Glacier
Your Deep Learning Partner

Data Visualization - cont

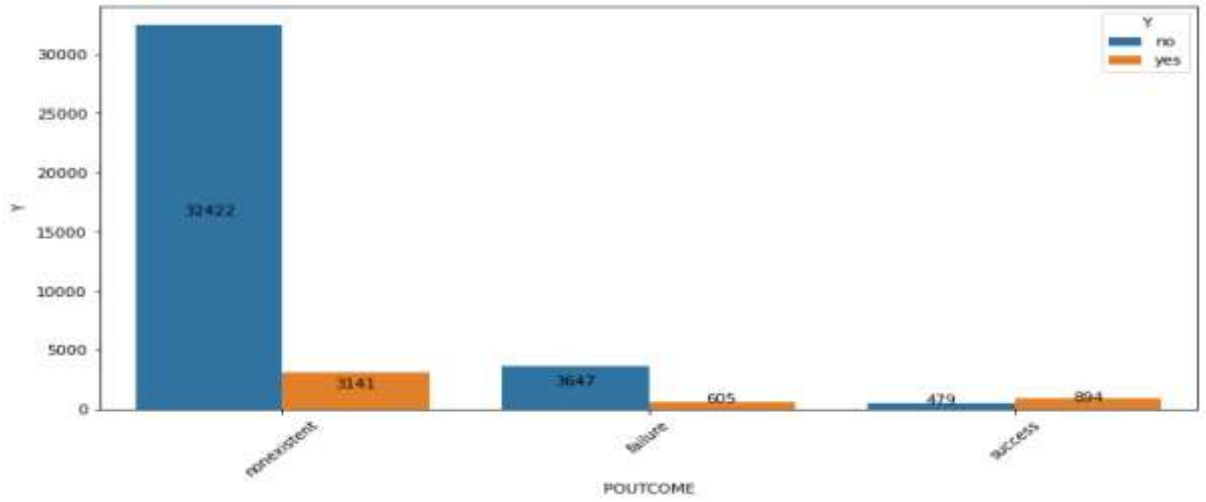

Fig 35: POutcome versus Y

According to the data presented in Fig 35, there are 3,141 non-existent customers but have a term deposit.

EDA Summary

Although there were no missing values, the “unknown” labels were present in the dataset. The correlation between EMP.VAR.RATE, EURIBOR3M, and NR.EMPLOYED, variables is more than 90 %. Understanding the fact that the categorical variables creates a relation with target variable at some level, it becomes crucial to also understand the time duration spent with the customer during the call might also impact the customers decision to subscribe for term deposit.

Recommendations

As analysed with the dataset, the occupation, education, contacted day and month are impacting the customers to opt for term deposit. Hence, the recommendation is to contact the customer keeping below observations into account.

1. Customers employed in administrative roles are more inclined to choose a term deposit option.

2. Customers with a university degree are more likely to opt for a term deposit. This indicates that students and retirees are more receptive to being contacted and converting the marketing campaign into a term deposit.

3. The month of May has had a greater impact on converting the marketing campaign into term deposits.

4. Monday and Thursday show slightly higher optimism when it comes to contacting customers during the marketing campaign.

Model Recommendations

The binary classification problem of the Bank Marketing Campaign involves predicting whether customers will subscribe to the term deposit. The following machine learning models have been selected for this task:

1. Linear regression
2. Linear Discriminant Analysis
3. Random Forest Classifier
4. AdaBoost Classifier
5. XGBoost Classifier
6. Hist Gradient Boosting Classifier

The initial evaluation of all the aforementioned machine learning models resulted in the following accuracy scores for each model. The models will be further fine-tuned with their respective hyperparameters using techniques like cross-validation and grid search.

Model Name	Accuracy
LR	0 90.402201
LDA	90.191794
RFC	91.073885
ADAB	90.887756
XGB	91.349033
HGBC	90.677349

References

[1] https://archive.ics.uci.edu/dataset/222/bank+marketing

Thank You