REPORT

Banking Dataset Classification:

About Dataset:

There has been a revenue decline in the Portuguese Bank and they would like to know what actions to take. After investigation, they found that the root cause was that their customers are not investing enough for long term deposits. So the bank would like to identify existing customers that have a higher chance to subscribe for a long term deposit and focus marketing efforts on such customers.

Data Set Information:

The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be subscribed ('yes') or not ('no') subscribed.

datasets: test.csv which is the test data that consists of 8238 observations and 20 features without the target feature

Goal:- The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

The dataset contains test data. Features of test data are listed below.

Bar Graph Representation:

housing

Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.07	
Std. Deviation	0.99	
Quantiles	0	Min
	0	25%
	2	50%
	2	75%
	2	Max

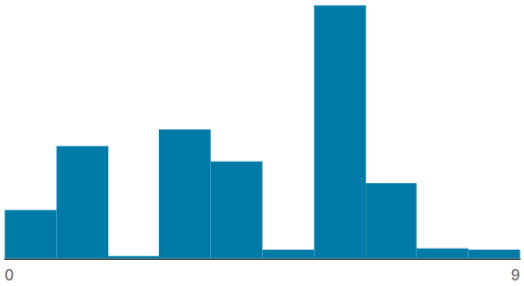
loan

Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.32	
Std. Deviation	0.72	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	2	Max

contact

Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.36	
Std. Deviation	0.48	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

month

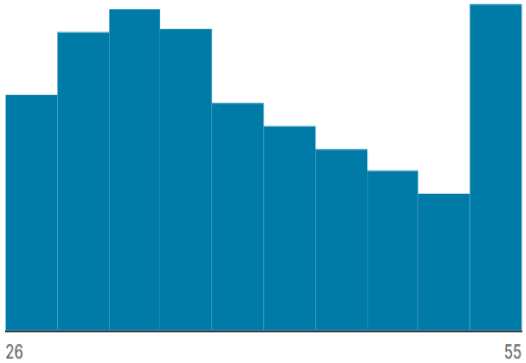


Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	4.22	
Std. Deviation	2.32	
Quantiles	0	Min
	3	25%
	4	50%
	6	75%
	9	Max

day_of_week

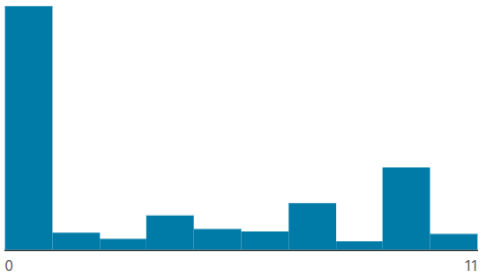
Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.02	
Std. Deviation	1.39	
Quantiles	0	Min
	1	25%
	2	50%
	3	75%
	4	Max

age



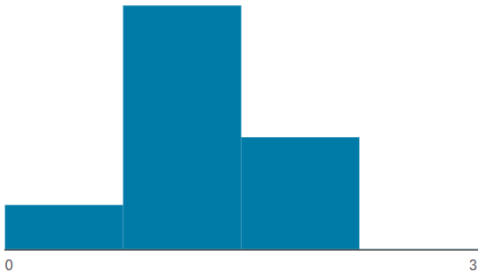
Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	39.6	
Std. Deviation	9.02	
Quantiles	26	Min
	32	25%
	38	50%
	47	75%
	55	Max

job



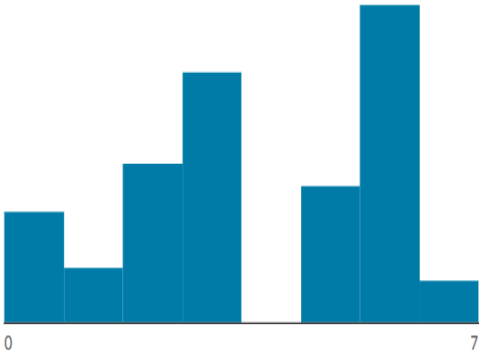
Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	3.73	
Std. Deviation	3.6	
Quantiles	0	Min
	0	25%
	2	50%
	7	75%
	11	Max

marital



Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.17	
Std. Deviation	0.61	
Quantiles	0	Min
	1	25%
	1	50%
	2	75%
	3	Max

education



Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	3.74	
Std. Deviation	2.13	
Quantiles	0	Min
	2	25%
	3	50%
	6	75%
	7	Max

default

Valid	8238	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.2	
Std. Deviation	0.4	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

Target Variable for DataSet:

Target variable (desired output):

Feature	Feature_Type	Description
y	binary	has the client subscribed a term deposit? ('yes','no')

Prediction:

Predicting if the client will subscribe to a term deposit.

In conclusion:

has the client subscribed a term deposit? ('yes','no')

Additional Work:

- Checking Correlation of feature variables
- Univariate analysis of Numerical columns
- plotting histogram for each numerical variable

Observation in plotting histogram for each numerical variable:

As we can see from the histogram, the features age, duration and campaign are heavily skewed and this is due to the presence of outliers as seen in the boxplot for these features.

Looking at the plot for pdays, we can infer that majority of the customers were being contacted for the first time because as per the feature description for pdays the value 999 indicates that the customer had not been contacted previously.

Checking Correlation of feature variables:

There are no features that are highly correlated and inversely correlated. If we had, we could have written the condition that if the correlation is higher than 0.8 (or can be any threshold value depending on the domain knowledge) and less than -0.8, we could have drop those features. Because those correlated features would have been doing the same job.