Business Case.

This scenario revolves around enhancing the marketing strategy for term deposits, with a particular emphasis on targeting customers. The intended audience for the results is the marketing team, aiming to furnish them with a prioritized list of bank customers who are more likely to engage in subscribing to term deposits.

We will use decision trees and logistic regression models to analyze the data. The dataset has a total of 17 features and 45211 data points.

Financial implications:

Assumed Cost	
Term Deposit Amount	$10,000
Rate of interest	5%
Time period (years)	3
Cost per term deposit	$1,500
Cost of call	$20
Servicing	$500
Total Cost incurred	$2,020

Here the assumed cost is the term deposit interest that we will pay off to the customers. The average term deposit is considered to be $10,000 dollars for 3 years at a rate of 5% per annum.

We have also considered the cost of call and servicing. The total cost is $2,020 dollars.

Assumed Revenue (Through loans)	
Deposit Amount	$30,000
Rate of interest	10%
Time period	3
Money earned	$9,000

The revenue will be generated by the loan interest that the bank will earn by giving this money out to others as a loan. Here, the average loan amount is considered $30,000 at a rate of 10% per annum for 3 years.

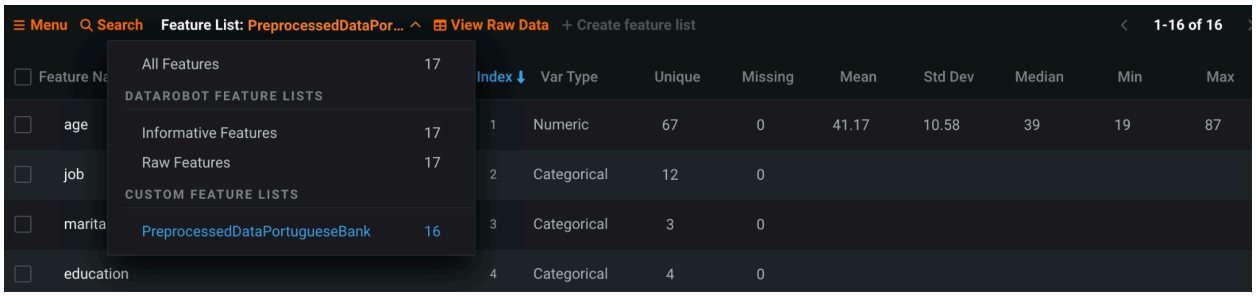
The money earned is $9000.

TP	Revenue - Cost = $9000 -$2020 = $6980
TN	No call, no revenue, no cost
FP	No call, no revenue, no cost
FN	Cost of call = $20

		Predicted	
		0	1
Actual	0	TN (0)	FP (-20)
	1	FN (0)	TP (6980)

Data Preprocessing

Our target variable is y as we want to know whether the customer will subscribe to the term deposit or not. Also, we have excluded 'duration' as it wouldn't be present as a data point before we actually make the calls. Hence, there is a new feature list created without 'duration' in it called PreprocessedDataPortugueseBank.



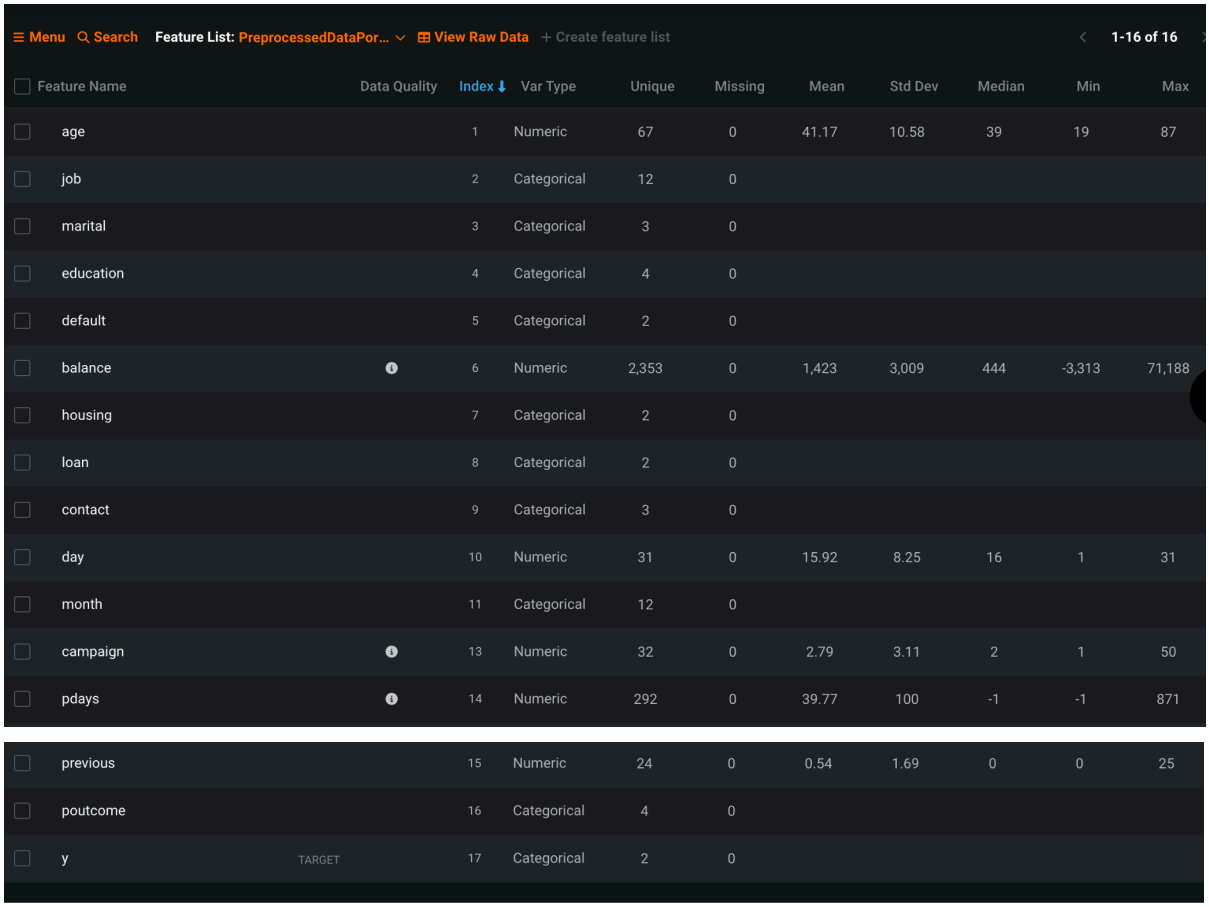
Q1. Missing Values.

There are missing values which are coded in a different way. Here,

- contact : There are values flagged as, 'unknown' which means they are missing values.
- pdays : There are values flagged as '-1' stating that the person hasn't been contacted before, which means either it's a missing value or the data is corrupted.
- poutcome : There is data which has 'unknown outcome' which means it has missing values for this case.

Q2. Features that have no variance.

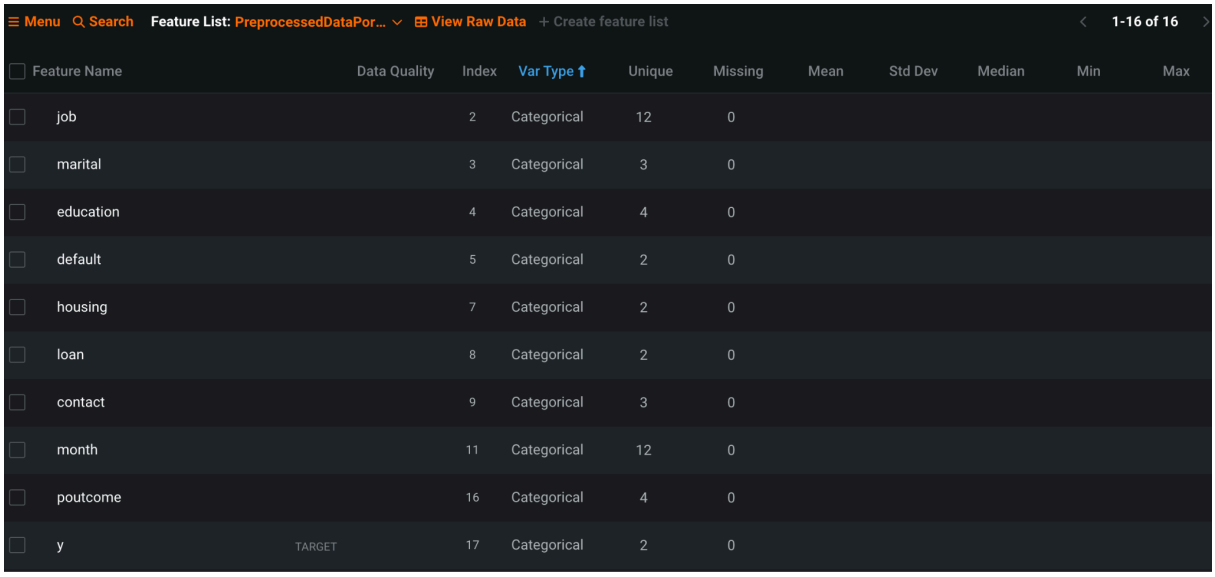
All the features have more than one unique value which means all the features have variance.



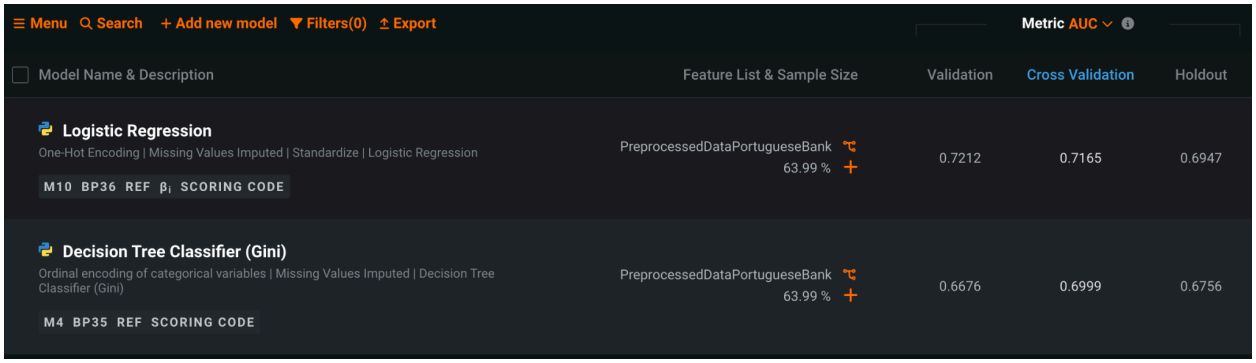
Feature Name	Data Quality	Index	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
age		1	Numeric	67	0	41.17	10.58	39	19	87
job		2	Categorical	12	0					
marital		3	Categorical	3	0					
education		4	Categorical	4	0					
default		5	Categorical	2	0					
balance		6	Numeric	2,353	0	1,423	3,009	444	-3,313	71,188
housing		7	Categorical	2	0					
loan		8	Categorical	2	0					
contact		9	Categorical	3	0					
day		10	Numeric	31	0	15.92	8.25	16	1	31
month		11	Categorical	12	0					
campaign		13	Numeric	32	0	2.79	3.11	2	1	50
pdays		14	Numeric	292	0	39.77	100	-1	-1	871
previous		15	Numeric	24	0	0.54	1.69	0	0	25
poutcome		16	Categorical	4	0					
y	TARGET	17	Categorical	2	0					

Q3. Categorical features with high cardinality.

The features with highest cardinality are job and month each with 12 unique values respectively. This isn't high cardinality given the data set we have.



Q4. Logistic Regression and Decision Tree Models for Marketing Campaign Response



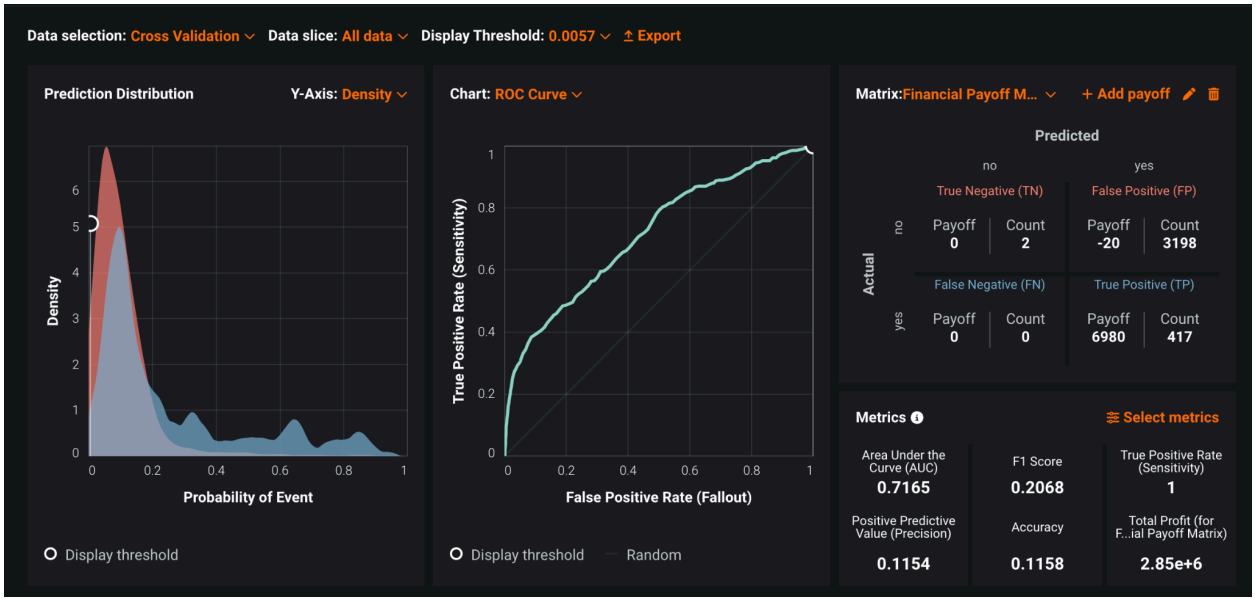
The values for AUC look good for both cross validation and holdout. It can be seen that considering AUC, Logistic Regression performs better than Decision Trees here.

Q5. Recall, precision, F1, accuracy, ROC AUC, maximum payoff for each of the models.

The payoff assumptions have been explained in the beginning of the report.

The values for the models are calculated while keeping the profit to be maximum.

Cross Validation:

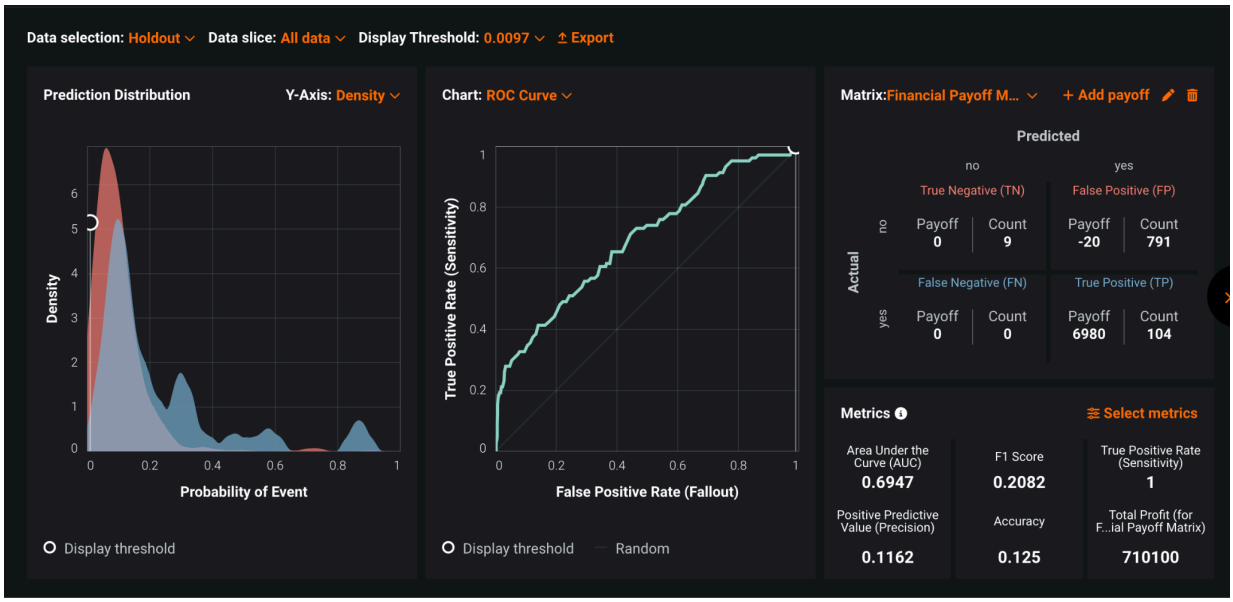


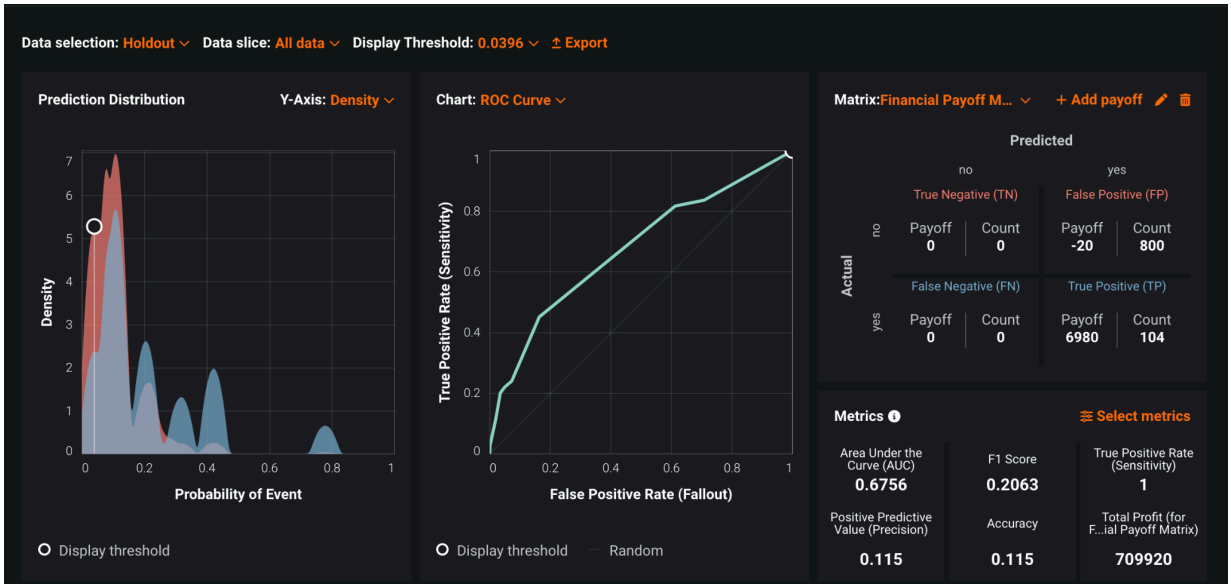
Cross validation

	Logistic Regression	Decision Tree
Recall	1	1
Precision	0.1154	0.1153
F1	0.2068	0.2067
Accuracy	0.1158	0.1153
ROC AUC	0.7165	0.6999
Maximum Payoff	2.85 million	2.85 million

For cross validation, we can see that the profit maximization is the same but the logistic regression model does a bit better on ROC AUC compared to the decision tree.

Holdout:





Holdout

	Logistic Regression	Decision Tree
Recall	1	1
Precision	0.1162	0.115
F1	0.2082	0.2063
Accuracy	0.125	0.115
ROC AUC	0.6947	0.6756
Maximum Payoff	710100	709920

For holdout we can see that the logistic regression model is doing better considering parameters like accuracy and maximum payoff.

Q6. Best metric and better model.

The best metric to evaluate the model is to choose maximum payoff, as in this case we are trying to maximize profit by making the customers subscribe to term deposit.

The better model for this case is logistic regression as it has slightly better values for some metrics than the decision tree.