Case Study – Term Deposit Marketing

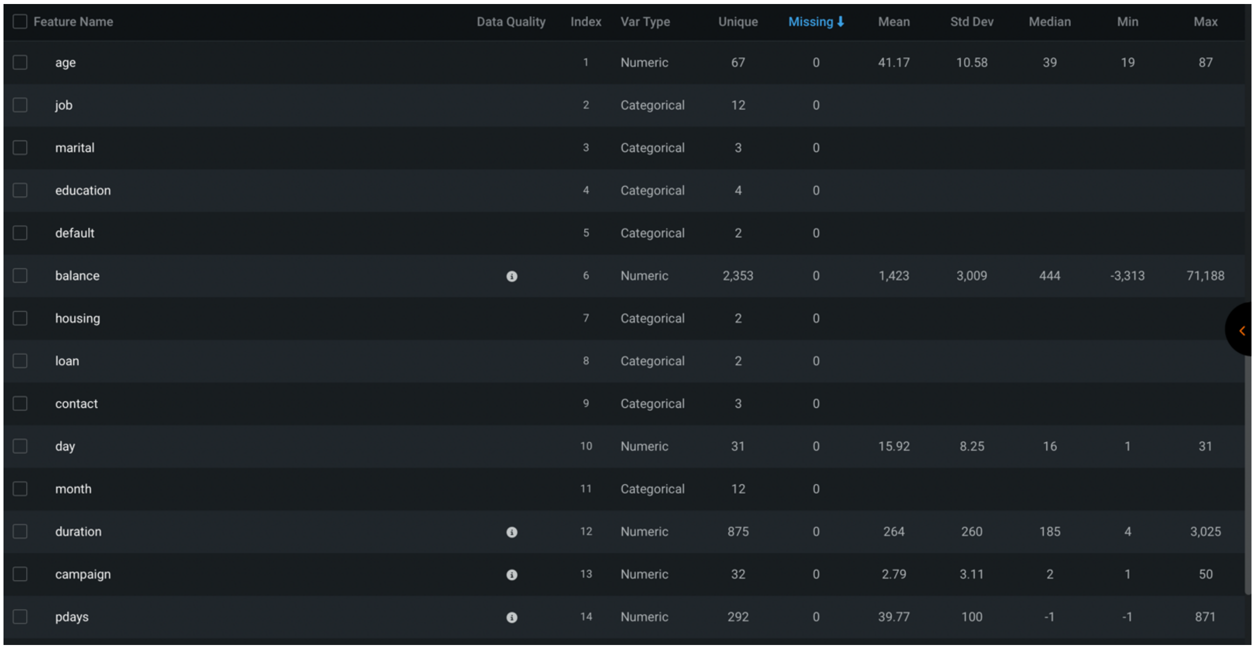
Business Case –

Within the banking industry, optimizing targeting for telemarketing is a key issue, under a growing pressure to increase profits and reduce costs. The 2008 financial crisis dramatically changed the business of European banks. Portuguese banks were pressured to increase capital requirements (e.g., by capturing more long-term deposits). The process of reaching out to customers for fixed-term deposit campaigns involves significant time and financial investment. **Without a pre-screening mechanism, there's a risk of squandering resources by approaching customers indiscriminately.**
The banks aspire to develop models with predictive capabilities to forecast whether a prospective customer is willing to deposit funds or not. The **objective is to precisely target marketing campaigns towards candidates with the highest potential to invest in fixed-term deposits.** The bank aims to improve the efficiency and effectiveness of its marketing strategies, focusing efforts on customer segments most likely to respond positively to fixed-term deposit offers. This approach empowers the bank to **enhance marketing campaign efficiency, optimize resource utilization, and precisely target candidates with the highest likelihood of investing in fixed-term deposits.** The aim is to sift through potential customers diligently, targeting those with the highest potential and optimizing both time and cost.

We have used dataset of Portuguese bank which has 4521 records.

Q1. Assess if any features are missing values.

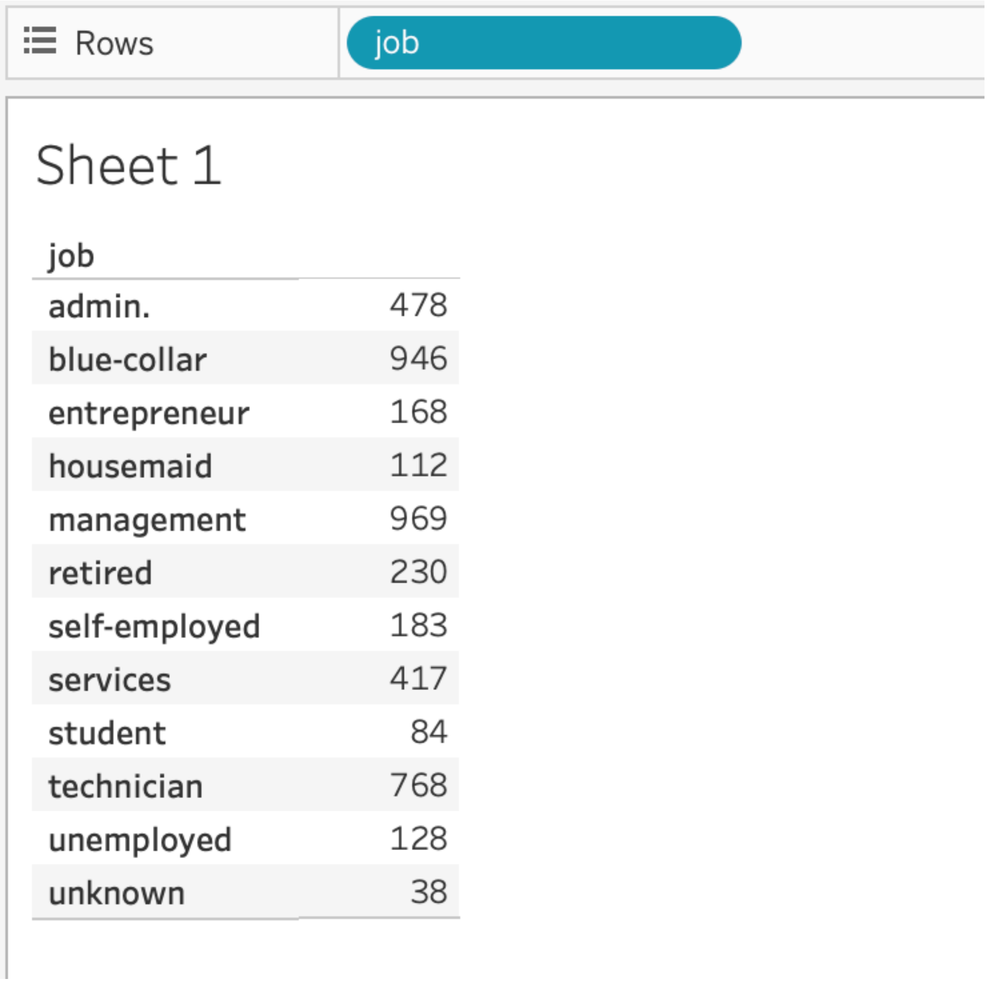
None of the features have missing values as such but there are certain values in the data which might be the possible outcome of either data corruption or representation of missing values in certain features.



□	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					
□	duration	ⓘ	12	Numeric	875	0	264	260	185	4	3,025
□	campaign	ⓘ	13	Numeric	32	0	2.79	3.11	2	1	50
□	pdays	ⓘ	14	Numeric	292	0	39.77	100	-1	-1	871

Fig:- No features have missing values

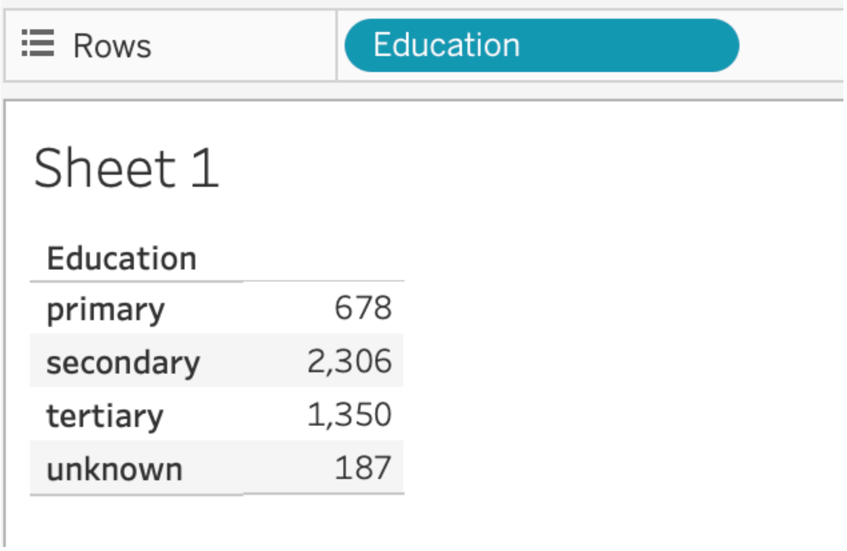
a) Job has 38 rows with values as unknown which signifies that these values might be missing, or we may not have record of job for these rows.



Sheet 1

job	
admin.	478
blue-collar	946
entrepreneur	168
housemaid	112
management	969
retired	230
self-employed	183
services	417
student	84
technician	768
unemployed	128
unknown	38

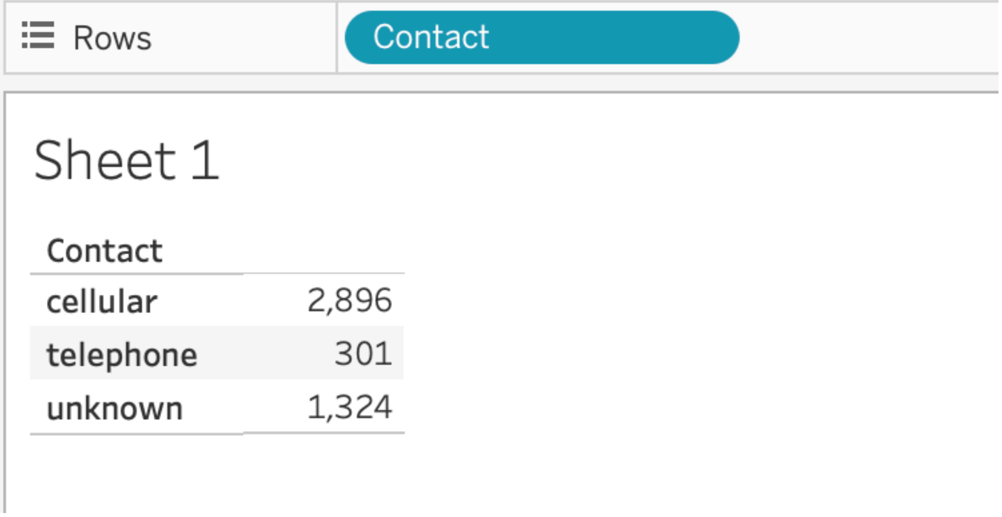
b) Education has 187 records with value as unknown.



Sheet 1

Education	
primary	678
secondary	2,306
tertiary	1,350
unknown	187

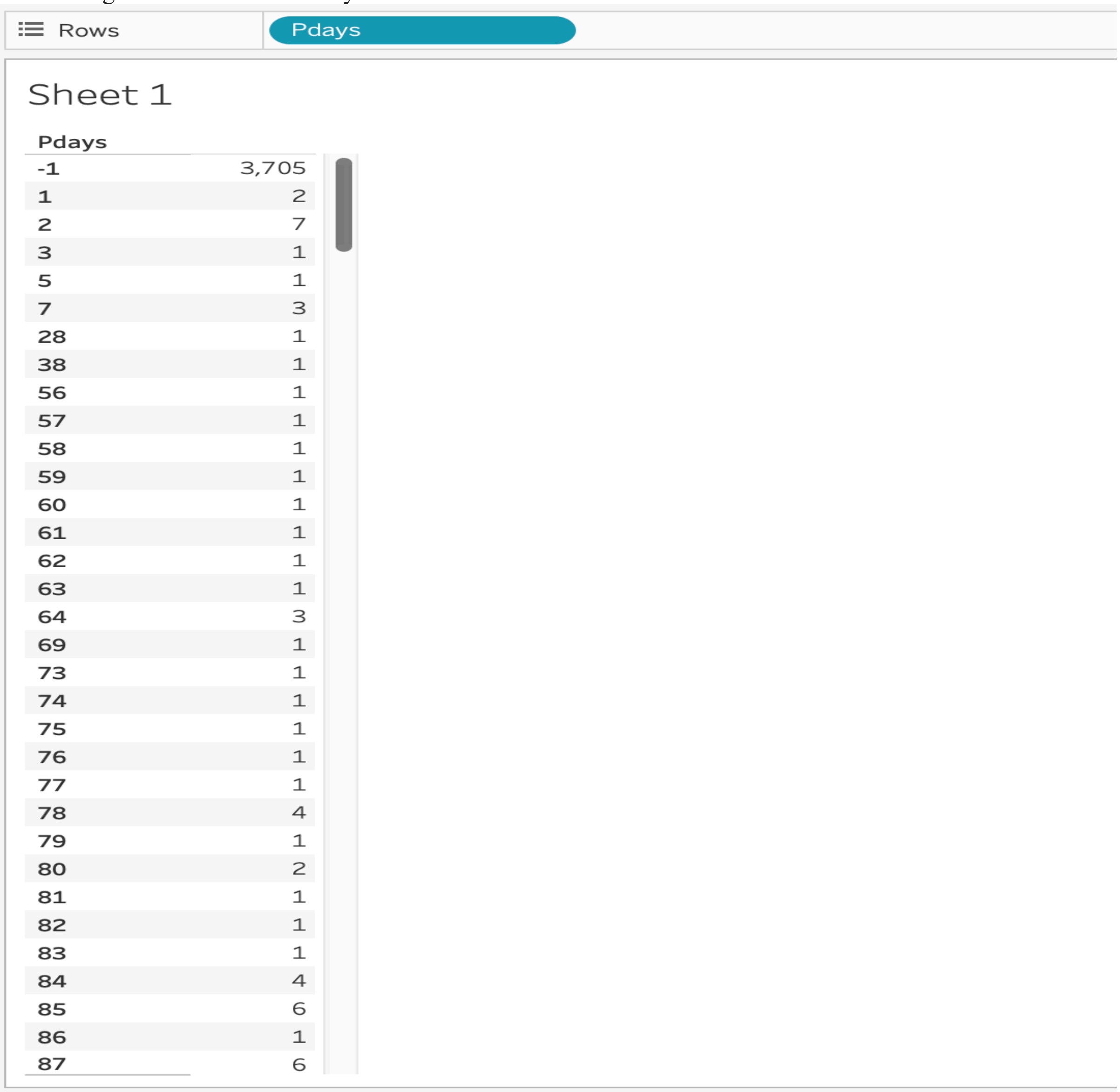
c) Contact communication type has 1324 rows with value as unknown. The data is for telemarketing and hence the value unknown for this column is indicative of probably missing values or incorrect data.



Sheet 1

Contact	
cellular	2,896
telephone	301
unknown	1,324

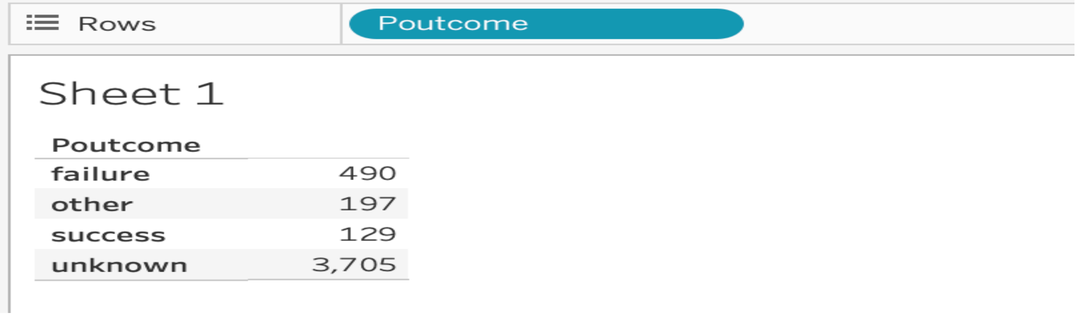
d) Number of days that passed by after the client was last contacted from a previous campaign (pdays) has 3705 with values as -1. Since number of days cannot be a negative value, it is probably just missing values disguised as a value or maybe some issue with the data.



Sheet 1

Pdays	
-1	3,705
1	2
2	7
3	1
5	1
7	3
28	1
38	1
56	1
57	1
58	1
59	1
60	1
61	1
62	1
63	1
64	3
69	1
73	1
74	1
75	1
76	1
77	1
78	4
79	1
80	2
81	1
82	1
83	1
84	4
85	6
86	1
87	6

e) outcome of the previous marketing campaign (poutcome) has 3705 rows with value as unknown which does not make much sense as it could either be that the customer agreed for the term deposit or not. So it is suggestive of either of the cases mentioned above.

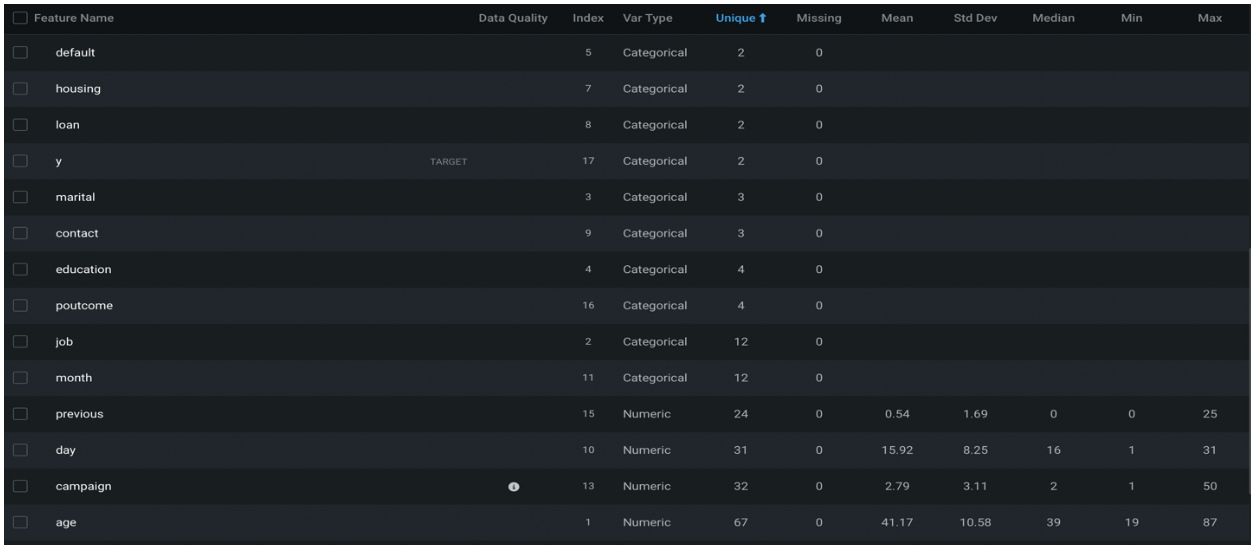


Sheet 1

Poutcome	
failure	490
other	197
success	129
unknown	3,705

Q2. Assess if any features have no variance.

Zero or no variance means that the feature values are constant across different target values. There are no features in our dataset that have constant values for all the target variables as all features have more than one unique value representing it. Therefore, none of the features in the dataset have no variance.



□ Feature Name		Data Quality	Index	Var Type	Unique ↑	Missing	Mean	Std Dev	Median	Min	Max
□ default			5	Categorical	2	0					
□ housing			7	Categorical	2	0					
□ loan			8	Categorical	2	0					
□ y	TARGET		17	Categorical	2	0					
□ marital			3	Categorical	3	0					
□ contact			9	Categorical	3	0					
□ education			4	Categorical	4	0					
□ poutcome			16	Categorical	4	0					
□ job			2	Categorical	12	0					
□ month			11	Categorical	12	0					
□ previous			15	Numeric	24	0	0.54	1.69	0	0	25
□ day			10	Numeric	31	0	15.92	8.25	16	1	31
□ campaign		ⓘ	13	Numeric	32	0	2.79	3.11	2	1	50
□ age			1	Numeric	67	0	41.17	10.58	39	19	87

Q3. Assess if any categorical features have high cardinality.

We have the below 9 categorical features in our dataset –

Categorical Features	Number of unique values
Job	12
Month	12
Education	4
poutcome	4
Marital	3
Contact	3
Default	2
Housing	2
Loan	2

None of the features above have high cardinality. The maximum number of possible unique values is for the feature job (12) and month (12) which is much lesser than the sample records in the data which is 4521 and hence is not considered as a high cardinality feature and all other features have lesser number of unique values as compared to this and hence, they also do not have high cardinality.

Q4. Develop logistic regression and decision tree models for marketing campaign response.

The model has been trained on 15 features – balance, pdays, age, education, day, campaign, previous, job, month, poutcome, marital, contact, default, housing, loan.
Day feature is kept as a numeric variable as converting it to categorical adds no additional information to the model. It is essential that the call is made a few days before or exactly on the payday in order to increase the chances of the customer accepting the term deposit offer.
The feature duration which is the last call duration has been removed from the feature set used for modeling as we cannot know the duration of the call until that call is made so we cannot use a feature which is based on action to train the model.

Target variable – y – has the client subscribed to the term deposit or not.

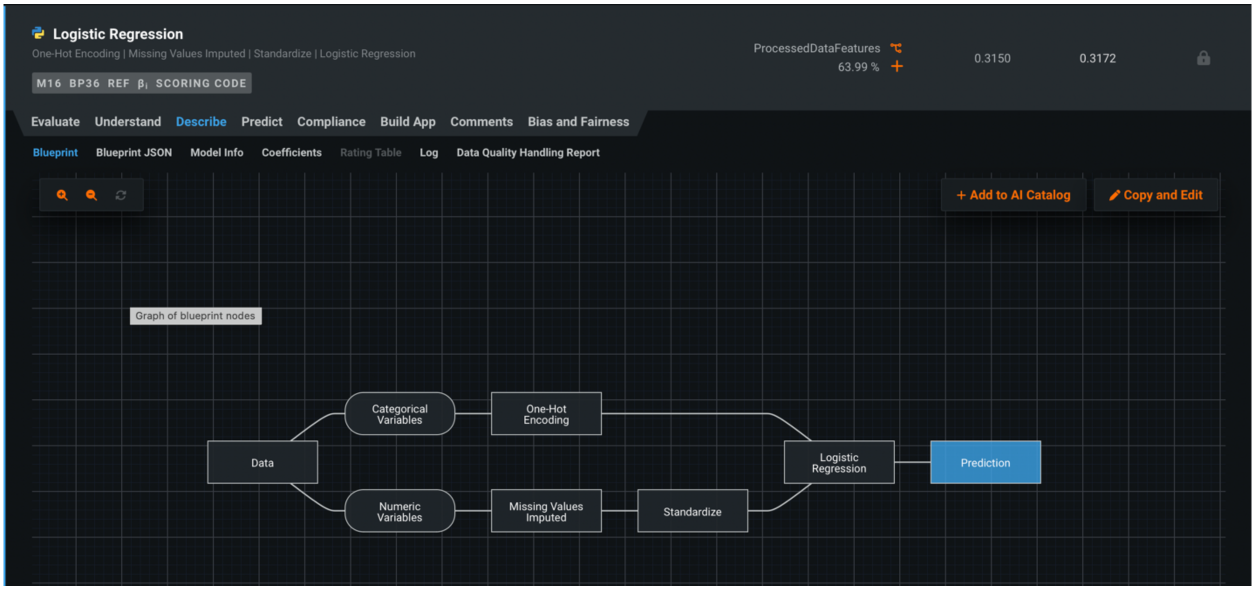


Fig: - Logistic Regression model

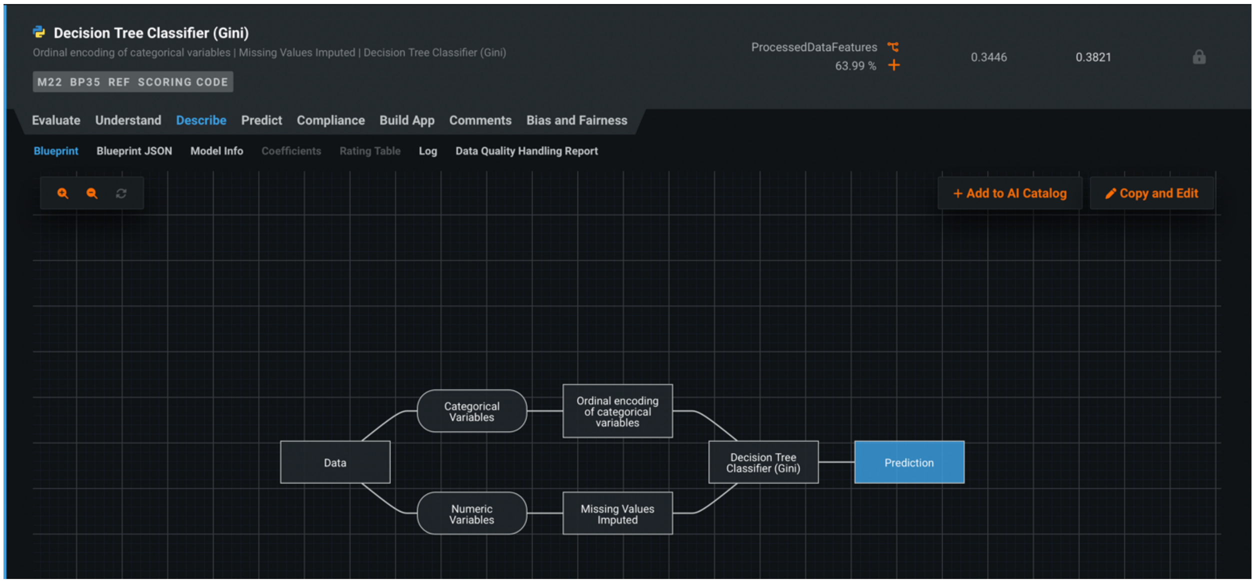


Fig: - Decision Tree Model

Q6. Report recall, precision, F1, accuracy, ROC AUC, maximum payoff for each of the models. Explicate your payoff matrix and the underlying assumptions.

For Payoff matrix we have made the following assumptions –

1. Term Deposit value – approx. €15000
2. Annual Interest received by the customer – 4.5%
3. Annual Interest received by bank – 7.5%
4. Duration for term deposit – 5 years
5. Cost of call – assuming standard rate per minute to be €0.5 we get a total cost of €30/hr and let's assume that 2 calls can be made per hour therefore cost of call would be €15

The values for payoff matrix would be computed as below -

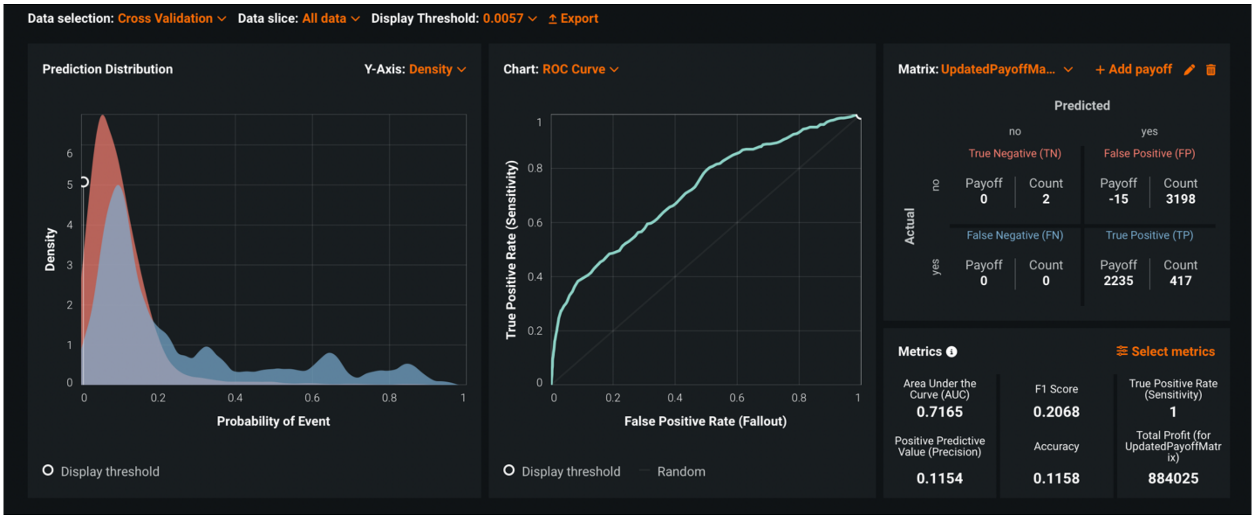
True Positive (the call was made and the proposal for CD was accepted) = 15000 * (0.075 – 0.045) * 5 – 15 = 2235
False Positive (the call was made but the offer was not accepted) = - 15
False Negative (the call was not made and there was no offer) = 0 argument can be made that this can be -2235 alternatively and can be identified as missed opportunity cost.
True Negative (the call is not made and the offer is not made) = 0

The final Payoff matrix is -

True Positive	**False Positive**
2235	-15
True Negative	**False Negative**
0	0

The metrics reported below are according to the threshold value set to maximize the profit.

Metric	Logistic Regression	Decision Tree
Recall	1	1
Precision	0.1154	0.1153
F1 score	0.2068	0.2067
Accuracy	0.1158	0.1153
ROC AUC	0.7165	0.699
Maximum payoff	€884025	€883995



Data selection: Cross Validation Data slice: All data Display Threshold: 0.0057 Export

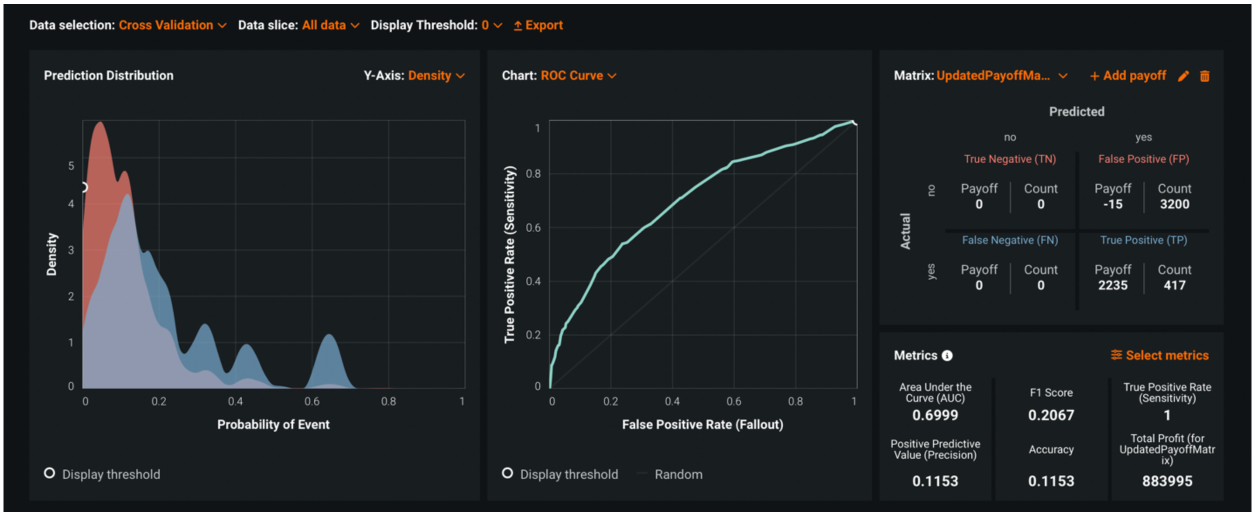
Matrix: UpdatedPayoffMa...

		Predicted: no	Predicted: yes
Actual	no	True Negative (TN) Payoff 0 | Count 2	False Positive (FP) Payoff -15 | Count 3198
	yes	False Negative (FN) Payoff 0 | Count 0	True Positive (TP) Payoff 2235 | Count 417

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.7165	0.2068	1
Positive Predictive Value (Precision)	**Accuracy**	**Total Profit (for UpdatedPayoffMatrix)**
0.1154	0.1158	884025

Fig: - Metrics reported for Logistic Regression



Data selection: Cross Validation Data slice: All data Display Threshold: 0 Export

Matrix: UpdatedPayoffMa...

		Predicted: no	Predicted: yes
Actual	no	True Negative (TN) Payoff 0 | Count 0	False Positive (FP) Payoff -15 | Count 3200
	yes	False Negative (FN) Payoff 0 | Count 0	True Positive (TP) Payoff 2235 | Count 417

Metrics

Area Under the Curve (AUC)	F1 Score	True Positive Rate (Sensitivity)
0.6999	0.2067	1
Positive Predictive Value (Precision)	**Accuracy**	**Total Profit (for UpdatedPayoffMatrix)**
0.1153	0.1153	883995

Fig: - Metrics reported for Decision Tree

Q7. What is the best metric to evaluate model performance and why? Which is the better model?

The best metric to predict the model performance is the maximum payoff metric. Our business problem requires us to maximize the profits for the bank by getting more and more customers to accept the proposals for term deposit. The payoff metric uses the costs that would be incurred by the bank for the various scenarios marked under true positive, true negative, false positive and false negative which can help us understand the distributions and maximize our profits. Since our end goal is to increase the profit by making calls to better targeted customers who would opt in for term deposits, the payoff metric is the best metric for model performance evaluation.

Logistic regression model has maximum payoff of €884025 as compared to decision tree which has a maximum payoff of €883995. Therefore, logistic regression helps us maximize the profit as it reduces the number of false positives thereby decreasing the overall cost of making the call. In terms of ROC AUC also, logistic regression model performs better than the decision tree. Hence, logistic regression is a better model for this scenario.