

Assignment 4 - Term deposit marketing

Business Case: Term deposit marketing case study signifies the case for Portuguese banking institution whose marketing team wants to evaluate the number of customers who are more likely to accept a bank term deposit offer based on phone calls. We'll focus our evaluation based on Portuguese bank data set containing 4521 values. Logistic Regression and Decision trees are used to make the analysis on given features. Models are further then compared against each other to find the better model to maximize the profit.

Financial Implications:

(Assumption based on data from banco de Portugal
<https://clientebancario.bportugal.pt/en/interest-rates>)

Term deposit amount : 15,000 €

annual interest given to customer for term deposit: 4.3%

annual interest received by bank on Mortgage loan: 7.7%(financing up to 90% of the property value)

duration: 5 years

Cost for call : 12 €, assuming per hour 36 €/hr and 3 calls are made per hour

Cost of servicing and other miscellaneous amount : 500 €

True positive:

$-(15,000 \times 5 \times 4.3)$ (Interest paid to customer for 5 years) $+(15,000 \times 5 \times 0.077 \times 0.90)$ (Interest earned due to personal loan) -500 (miscellaneous/servicing cost) -12 (cost for calling)

$=(-3225+5197-500-12)$

$=1460$ €

True Negative: 0 € (no call, no revenue , no cost)

False Negative: 0 € (no call, no revenue , no cost). This can however be alternatively seen

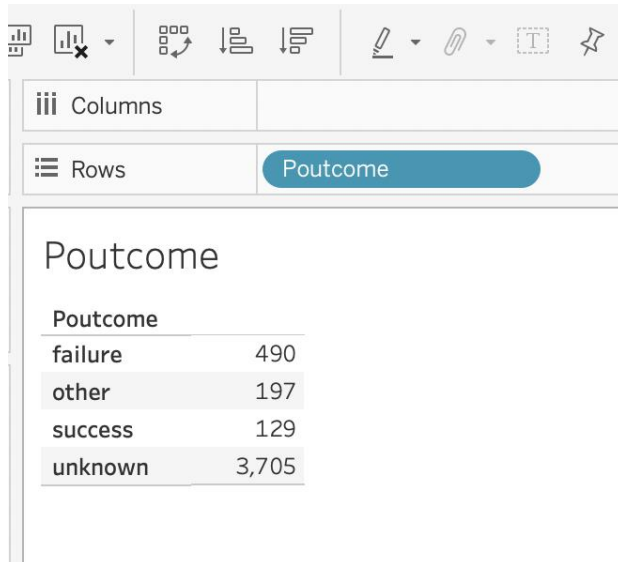
as -1460 € opportunity cost

False Positive: -12 € (cost for call but no revenue)

Question 1:

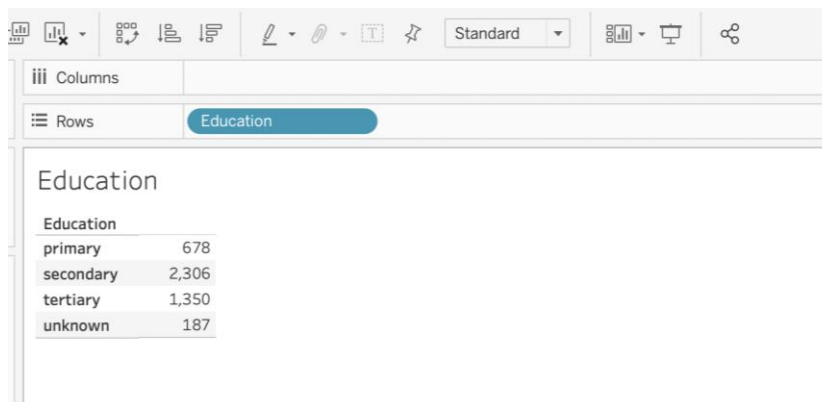
The following features have missing values:

1. Poutcome- Number of values missing are 3705 out of 4521 total values (approx. 81.95% data)



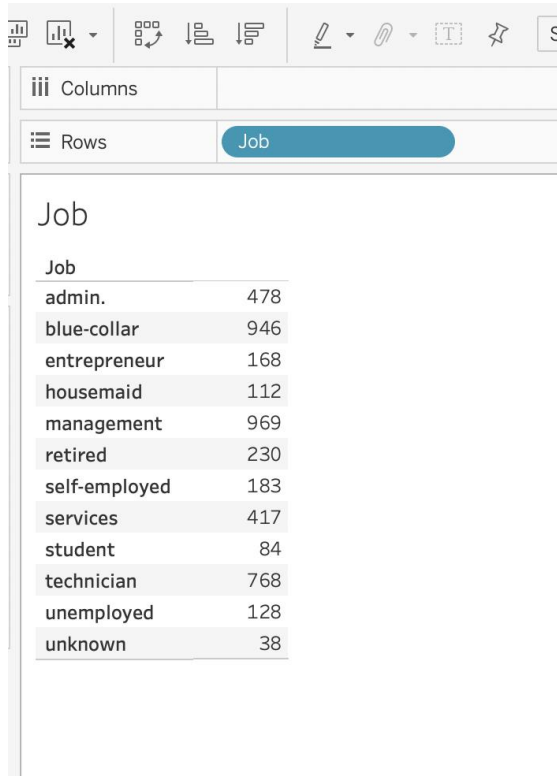
Poutcome	
failure	490
other	197
success	129
unknown	3,705

2. Education - Number of values missing are 187 out of 4521 total values (approx. 4.14% data)



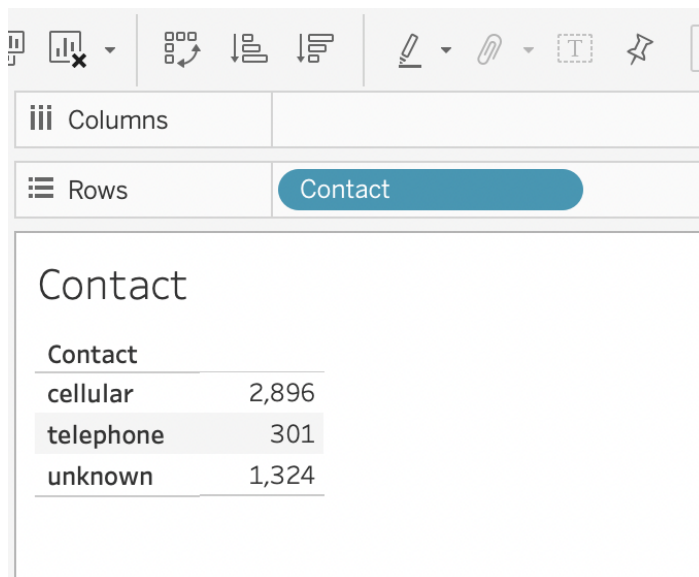
Education	
primary	678
secondary	2,306
tertiary	1,350
unknown	187

3. Job - Number of values missing are 38 out of 4521 total values (approx. 0.84% data)



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

4. Contact- Number of values missing are 1324 out of 4521 total values (approx. 29.29% data)



Contact	
cellular	2,896
telephone	301
unknown	1,324

Poutcome seems to have very large data missing which may cause issues while creating predictive models. (However, in our case datarobot itself handles the missing values, therefore we need not make any necessary steps to correct this before processing data in the model). Contact, education, job seems to have insignificant amount of data missing when compared to the dataset size.

Question 2:

In our case, no feature is available with no variance.

Feature Name	Data Quality	Index	Importance ↑	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
<input type="checkbox"/> poutcome		16	<div><div></div></div>	Categorical	4	0					
<input type="checkbox"/> month		11	<div><div></div></div>	Categorical	12	0					
<input type="checkbox"/> contact		9	<div><div></div></div>	Categorical	3	0					
<input type="checkbox"/> previous		15	<div><div></div></div>	Numeric	23	0	0.55	1.72	0	0	24
<input type="checkbox"/> pdays	<div><div></div></div>	14	<div><div></div></div>	Numeric	264	0	39.63	99.78	-1	-1	808
<input type="checkbox"/> housing		7	<div><div></div></div>	Categorical	2	0					
<input type="checkbox"/> job		2	<div><div></div></div>	Categorical	12	0					
<input type="checkbox"/> age		1	<div><div></div></div>	Numeric	67	0	41.43	10.72	40	19	87
<input type="checkbox"/> Updated_d...tegorical		10	<div><div></div></div>	Categorical	31	0					
<input type="checkbox"/> balance	<div><div></div></div>	6	<div><div></div></div>	Numeric	2,044	0	1,424	2,959	459	-3,313	71,188
<input type="checkbox"/> loan		8	<div><div></div></div>	Categorical	2	0					
<input type="checkbox"/> campaign	<div><div></div></div>	13	<div><div></div></div>	Numeric	30	0	2.77	3.01	2	1	50
<input type="checkbox"/> marital		3	<div><div></div></div>	Categorical	3	0					
<input type="checkbox"/> education		4	<div><div></div></div>	Categorical	4	0					
<input type="checkbox"/> default		5	<div><div></div></div>	Categorical	2	0					

Question 3:

None of the categorical features have high cardinality. Days have highest cardinality of 31 among all categorical data but when compared to dataset size of 4521 it can be neglected.

Question 4:

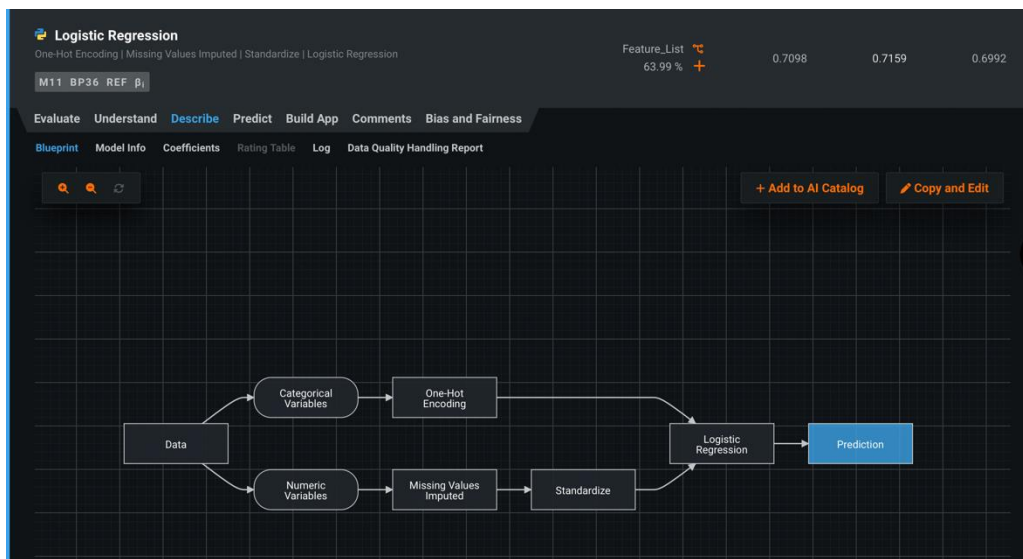
Data preprocessing steps:

1. As we are building predictive model, we have removed duration from the feature list for the model as the call duration cannot be known before making the required call. Same has been highlighted in the data dictionary.

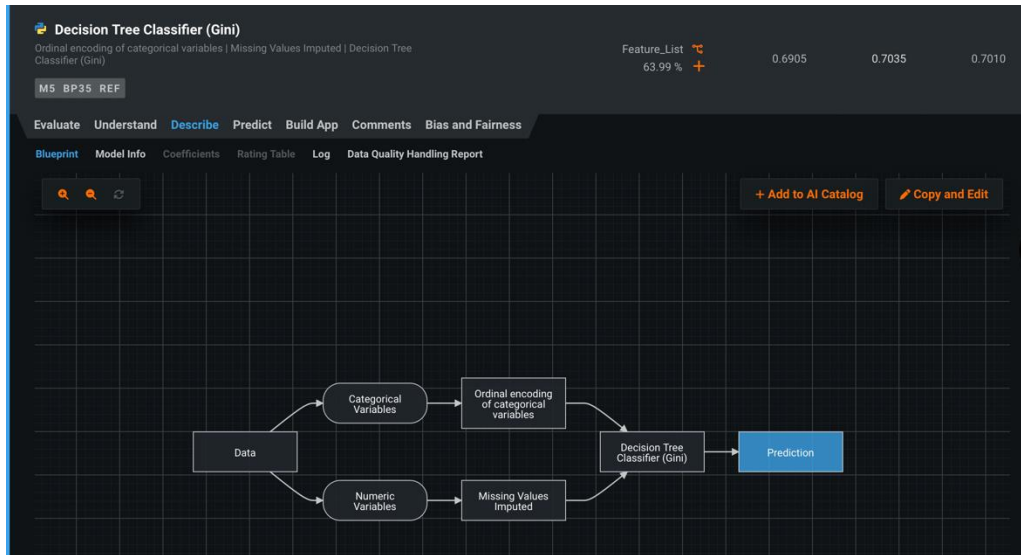
contact: contact communication type (categorical: 'cellular','telephone')
month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is usually known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
her attributes:
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

2. 'Day' is converted into categorical feature.

Logistic Regression :



Decision Tree:



Model	Validation	Cross Validation	Handout
Logistic Regression	0.7098	0.7159	0.6992
Decision Tree	0.6905	0.7035	0.7010

Decision Tree has better performance on handout data.

Question 5:

Assumptions:

Term deposit amount : 15,000 €

annual interest given to customer for term deposit: 4.3%

annual interest received by bank on Mortgage loan: 7.7%(financing up to 90% of the property value)

duration: 5 years

Cost for call : 12 €, assuming per hour 36 €/hr and 3 calls are made per hour

Cost of servicing and other miscellaneous amount : 500 €

True positive:

$-(15,000 \times 5 \times 4.3)$ (Interest paid to customer for 5 years) $+(15,000 \times 5 \times 0.077 \times 0.90)$ (Interest earned due to personal loan) -500 (miscellaneous/servicing cost) -12 (cost for calling)

$=(-3225+5197-500-12)$

$=1460$ €

True Negative: 0 € (no call, no revenue , no cost)

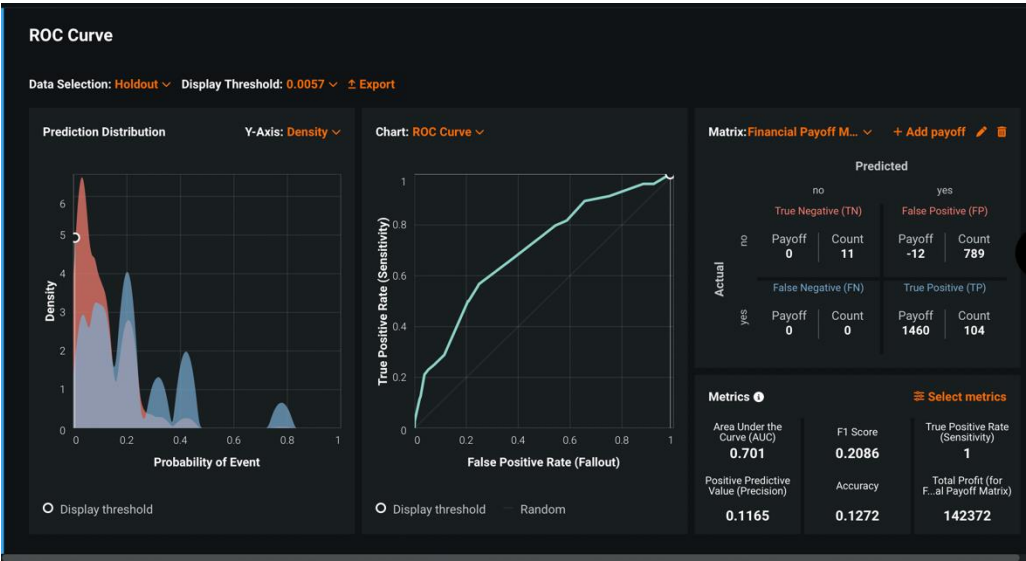
False Negative: 0 € (no call, no revenue , no cost). This can however be alternatively seen

as -1460 € opportunity cost

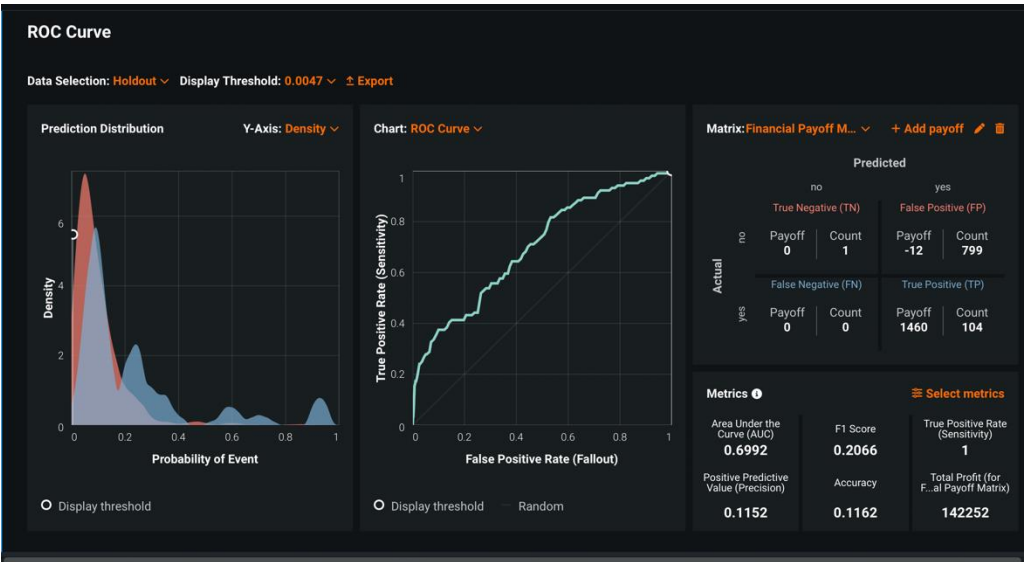
False Positive: -12 €(cost for call but no revenue)

True Negative	False Positive
0	-12
False Negative	True Positive
0	1460

Decision Tree:



Logistic Regression :



Model	Recall	Precision	F1	Accuracy	Error	ROC AUC	Maximum Payoff
Logistic Regression (threshold:0.0047)	1	0.1152	0.2066	0.1162	0.8838	0.6992	142,252€
Decision Tree (threshold: 0.0057)	1	0.1165	0.2086	0.1272	0.8728	0.701	142,372€

Decision tree model has the maximum payoff value of 142,372€.

Question 6:

Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to evaluate the profit based on the number of customers who will opt for term deposits. Maximum payoff metric assigns costs and benefits to different types of correct and incorrect predictions (true positives/true negatives and false positives/false negatives) and help evaluate the required profit/losses based on the given case. As our business case requires us to evaluate the potential customer who may opt for term deposit, we need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

ROC AUC metric (which is independent of the threshold value) is 0.701 for decision tree model when compared to 0.6992 for Logistic Regression model. Which evaluates that Decision Tree model is a better estimator. If we compare the maximum pay off value to evaluate which model maximizes the profit logistic regression has the value of 142,252€(at threshold of 0.0047)and has 799 false positive values which effects the overall profitability of the model. When compared to maximum payoff metric of the decision tree for 142,372€(at threshold of 0.0057) it has lower false positive values of 789 in our case. Hence, Decision tree models is the better model as it can evaluate more profit.