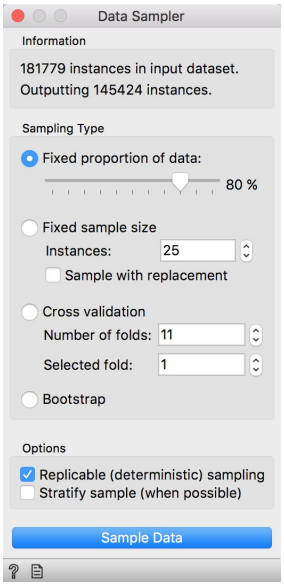
Red Cross Case

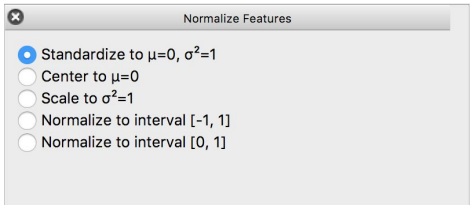
In the model, I am predicting whether or not a past financial donor will donate blood(target: “class”). Variables that I use to predict the target are: Income, First date of donation, Last date of donation, Amount, Frequency of donation, Star donator, Amount of last gift, Amount of average gift. I also decided to skip the variable “rfaa2: donation amount code” because we already have a numerical variable to indicate the donation amount and the categorical variable, labeled as code, does not provide helpful information for what the model is predicting.

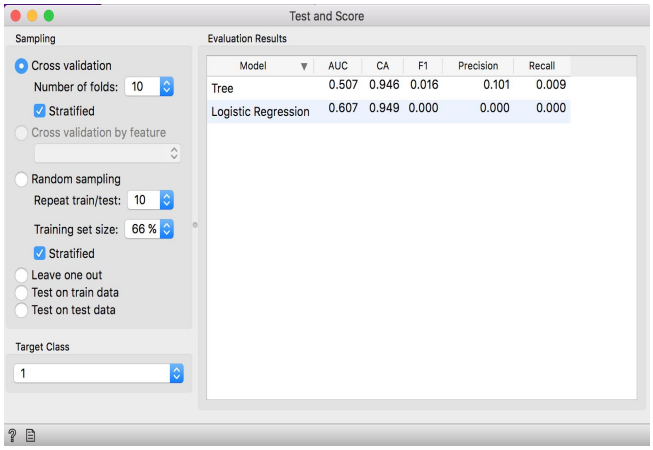
	Name	Type	Role	Values
1	**Id**	**numeric**	**meta**	
2	Income	numeric	feature	
3	Firstdate	numeric	feature	
4	Lastdate	numeric	feature	
5	Amount	numeric	feature	
6	rfaf2	numeric	feature	
7	**rfaa2**	**categori...**	**skip**	**D, E, F, G**
8	pepstrfl	categorical	feature	0, X
9	glast	numeric	feature	
10	gavr	numeric	feature	
11	**class**	**categori...**	**target**	**0, 1**

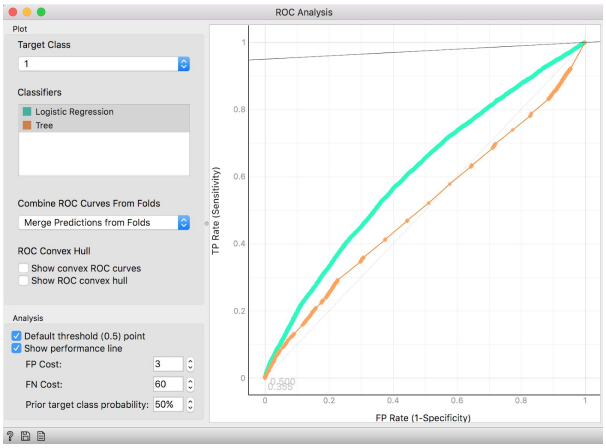
To prepare the data, I separated 80% of the data for the training set and 20% as the test set to get an independent accuracy reading for our model. This will be helpful to test the model’s generalization performance on the test data. The sampling is replicated to maintain consistency.



Since this is a classification problem, I considered tree and logistic regression algorithms for the model. I compared the two algorithms using AUC and ROC with the assumption that the higher value of AUC and an uppermost ROC line equates to a better performing model. Before regularizing with the logistic regression algorithm, I standardized the dataset. Since the data set is relatively large and was time-consuming to test with different complexity parameters that would maximize AUC, I only tested a few complexity parameters. However, the results all agree that logistic regression is a better model with a higher AUC and higher ROC line.







I then proceeded to choose the threshold for the logistic regression model. I considered two ways to choose the threshold: with the ROC curve and with Excel. For the ROC curve, the FP:FN cost ratio is 3:60 as the cost of a FP is the cost of sending the campaign ($3) and the cost of a FN is the cost of how much Red Cross would be willing to pay to a blood donor ($60). The threshold found by the intersection of the iso-performance line and the ROC line is 2.6%. In other words, according to the ROC analysis, Red Cross should send out campaign requesting blood donations to financial donors who have a probability of 2.6% or higher of donating blood.

	0	1
0	TN: model predicts rejection and we rejected the applicant = 57 (=60-3)	FP: model predicts acceptance and we rejected the applicant = -3
1	FN: model predicts rejection and we accepted the applicant = 0	TP: model predicts acceptance and we accepted the applicant = 0

class	Logistic Regression (	VALUE	POP. %	PROFIT CURVE	DECISION	BENEFIT			
0	0.354992071	-3	0.00%	$ (3.00)	1	-3		SIZE	153,424
0	0.19556568	-3	0.00%	$ (6.00)	1	-3		THRESHOLD	5.02%
0	0.192006712	-3	0.00%	$ (9.00)	1	-3		TOTAL BENEFIT	$ 70,917.00
1	0.16133605	57	0.00%	$ 48.00	1	57			
0	0.147842466	-3	0.00%	$ 45.00	1	-3			
0	0.146195387	-3	0.00%	$ 42.00	1	-3		MAX BENEFIT	$ 71,181.00
1	0.145629887	57	0.00%	$ 99.00	1	57		ROW	60194
1	0.143230846	57	0.01%	$ 156.00	1	57		BEST THRESHOLD	5.02%
0	0.141720179	-3	0.01%	$ 153.00	1	-3		BENEFIT PER DECISION	$ 0.46
0	0.140867012	-3	0.01%	$ 150.00	1	-3			
0	0.140209563	-3	0.01%	$ 147.00	1	-3			

In addition to the ROC threshold, I also found a threshold with Excel given the benefit matrix as seen above. Given the analysis, the threshold that maximized total benefit is 5.02%. I proceeded with this threshold over the ROC threshold since we know the value of TP and FP and the Excel threshold allows the model to produce a maximum total benefit.

I considered two methods in evaluating the model: with AUC and Excel evaluation. The AUC on the test data is 0.616, which is an improvement on the original AUC of 0.607. The Excel evaluation can be seen below.

class	Logistic Regression	value	decision	benefit			
1	0.077849467	$ 57.00	1	$57.00		Threshold	5.02%
0	0.03239538	$ (3.00)	0	$0.00		Total benefit	$19,605.00
0	0.041458681	$ (3.00)	0	$0.00		# of people	36355
0	0.035957434	$ (3.00)	0	$0.00		# of follow ups	15145
0	0.045846142	$ (3.00)	0	$0.00		Prob. Of Follow U	41.7%
0	0.061058824	$ (3.00)	1	-$3.00		Savings per perso	$ 0.54
1	0.060033526	$ 57.00	1	$57.00		Baseline (target al	$ 0.01
0	0.05215599	$ (3.00)	1	-$3.00			
0	0.083820732	$ (3.00)	1	-$3.00			
0	0.032738541	$ (3.00)	0	$0.00		People Per Year	100
0	0.068689822	$ (3.00)	1	-$3.00		Impact Per Year	$ 53.41
1	0.039689376	$ 57.00	0	$0.00		Capital Investmen	$ 208,293.22

Given the selected threshold, the total benefit would be around $20,000. In summary, given the model, Red Cross should send out campaigns requesting blood donations to financial donors who have a probability of 5.02% or higher of donating blood.