

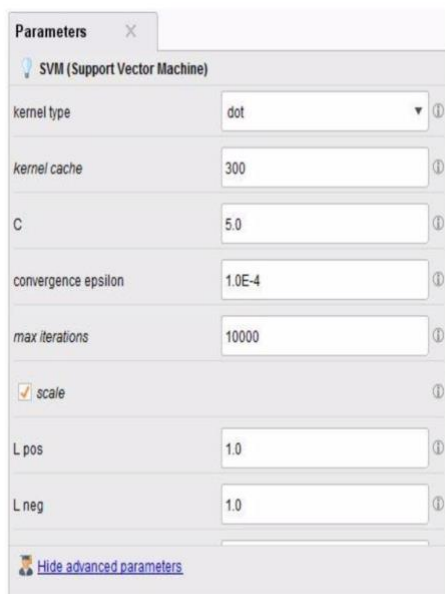
Target Marketing – Fundraising - PART 2

-Sheethal Heremagalur Sridhar

After Preprocessing the data in Part 1, we managed to reduce the dataset and the number of variables through variable transformation, replacing missing values, PCA for dimension reduction and Random Forest for selecting variables.

Now, We use the same dataset to perform **SVM(Support Vector Machine)**

We used different parameters like Dot, Polynomial and Radial Kernel to develop the classifier

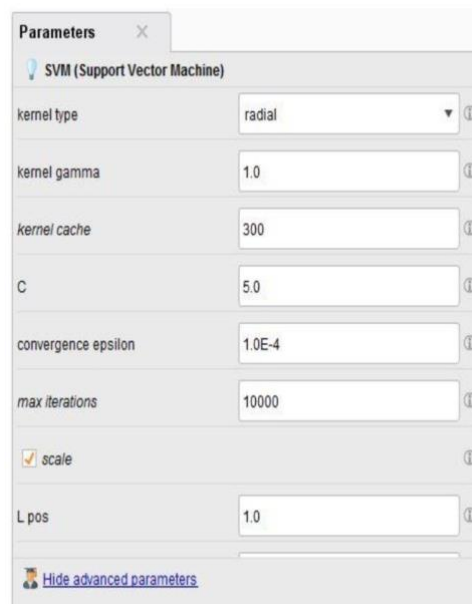


Parameters

SVM (Support Vector Machine)

kernel type	dot
kernel cache	300
C	5.0
convergence epsilon	1.0E-4
max iterations	10000
<input checked="" type="checkbox"/> scale	
L pos	1.0
L neg	1.0

[Hide advanced parameters](#)



Parameters

SVM (Support Vector Machine)

kernel type	radial
kernel gamma	1.0
kernel cache	300
C	5.0
convergence epsilon	1.0E-4
max iterations	10000
<input checked="" type="checkbox"/> scale	
L pos	1.0

[Hide advanced parameters](#)

Parameters

SVM (Support Vector Machine)

kernel type polynomial

kernel degree 3.0

kernel cache 200

C 5.0

convergence epsilon 0.001

max iterations 100000

☒ scale

L pos 1.0

Hide advanced parameters

According to the results, we found that Polynomial Kernel was our best SVM Model

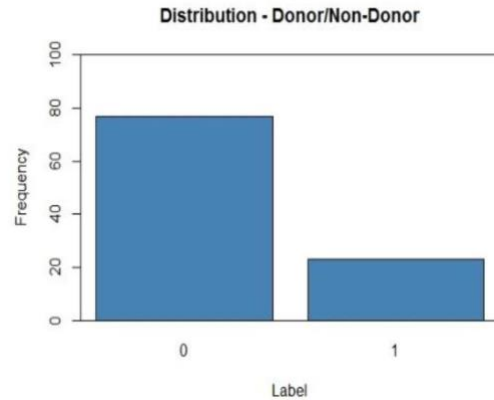
Summarizing the test results for all the models:

	TRAINING		TESTING	
Models	Accuracy	Recall(True=1)	Accuracy	Recall(True=1)
Naive Bayes	68.25%	6.45%	67.13%	5.98%
SVM	84.09%	32.40%	68.48%	26.38%
Lasso(Logistic Regression)	52.77%	1.00%	51.67%	0.64%
Random Forest	79.20%	82.3%	78.98%	0.05%
Decision Tree	77.16%	0.10%	76.70%	0.05%
Gradient Boosted Tree(GBT)	55.34%	73.38%	52.84%	63.8%

Based on the Test Data, Random Forest has the best Accuracy while Gradient Boosted Tree has the best Recall value at True=1 so we choose GBT as our best model.

The Distribution of Donor/Non-Donor is denoted by the graph.

We can say that 77% are Non-Donors while 23% of our Dataset are donors



Calculation of the ADJUSTED PROFIT AND ADJUSTED COST:

Cost of each mailing = \$0.68

So, The profit per donation = \$13.00 – \$0.68 = \$12.32.

Undoing the effect of the weighted sample to reflect actual response distribution:

The data used for the training and validation (25% donors) of the model has a different proportion of donors compared to the original population proportion of (5.1%), we will calculate the weight per donor profit and cost per solicitation, which would then be used directly on the model outputs from the weighted data.

Weighted Profit = Profit per donation * (% of Actual Responders) / (% of responders in weighted sample)

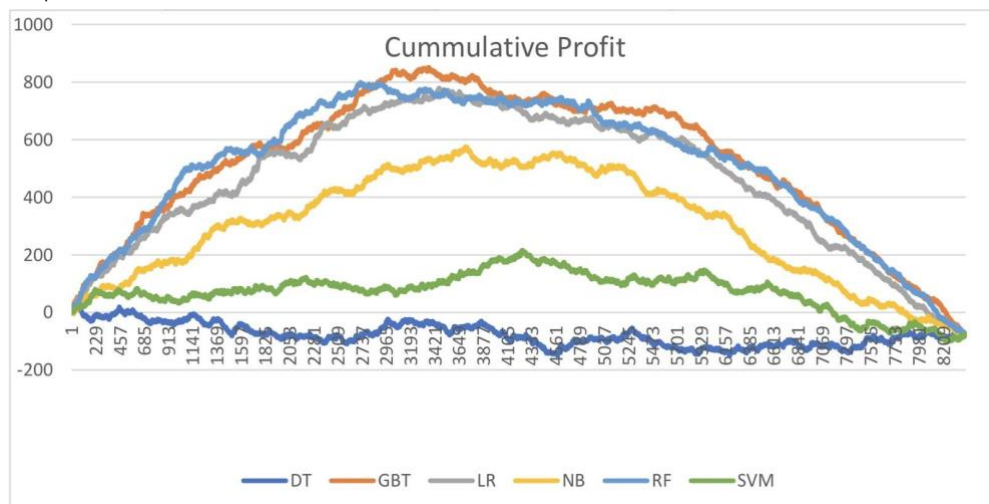
For Donors- $(12.32 \times 0.051) / 0.25 = 2.51$

For Non-Donors- $(0.68 \times 0.949) / 0.75 = 0.86$

	TRAINING		TESTING			
Models	Accuracy	Recall(True=1)	Accuracy	Recall(True=1)	Maximum Profit(\$)	Cut-Off
Naive Bayes	68.25%	6.45%	67.13%	5.98%	550.42	0.00930240
SVM	84.09%	32.40%	68.48%	26.38%	210.74	0.29484049
ssso(Logistic Regression)	52.77%	1.00%	51.67%	0.64%	735.34	0.2230594

Random Forest	79.20%	82.3%	78.98%	0.05%	703.48	0.27473375
Decision Tree	77.16%	0.10%	76.70%	0.05%	23.48	0.22830487
Gradient Boosted Tree (GBDT)	55.34%	73.38%	52.84%	63.8%	840.83	0.23475038

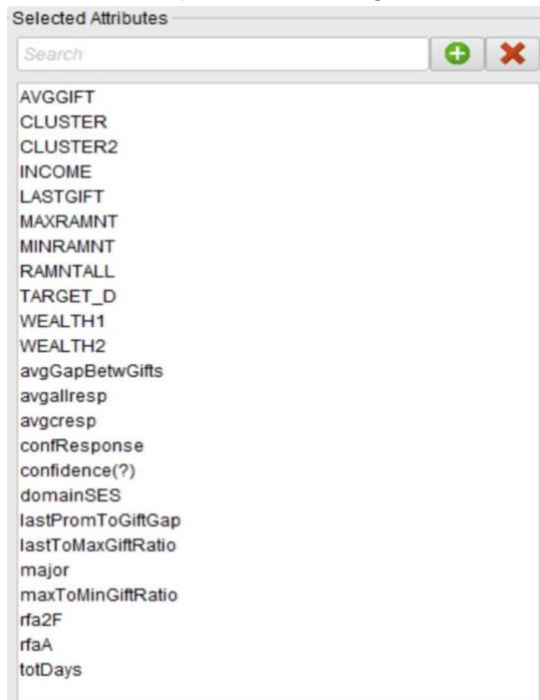
Based on the Maximum profit, we choose GBT as our best model with a cumulative profit of \$840.83 and a cut-off value of 0.229



Parameters used for our best model (Gradient Boosted Trees) -

- Number of trees: 100
- Maximal depth: 20
- Minimum rows: 10.0
- Minimum split improvement: 0.0
- Number of bins: 20
- Learning rate: 0.1
- Sampling Rate: 1.0

We select the following attributes for developing the model as Target_D has values only for donors and to develop a model using the donated amount, we do not need to take all the observations.



Split Ratio for Training and Validation is 60:40

To identify the targets : Result from the response (classification) model $P(\text{donor}|X)$ * result from the donation amount model $E(\text{donation}|x)$

Also, if the predicted amount of donation is more than \$0.6 , we would prefer that individual as a donor, as it takes \$0.68 to send a mail.

Profit obtained using TargetD model is given below -

On Training Data:

ExampleSet (1 example, 0 special attributes, 2 regular attributes)

Row No.	sum(profit)	sum(tgtDOnlyProfit)
1	35880.450	42504.330

On Test Data:

Row No.	sum(profit)	sum(tgtDOnl...
1	25476.970	29871.130

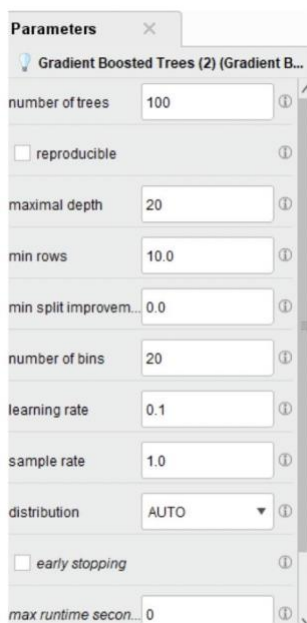
For the No Model case, the solicitation was mailed to all the individuals in our data set. Also, the profit that is obtained above for the test data set is only 40%(60:40 split), we calculate for the same.

Number of individuals = 7583

Maximum Profit = $42504 - 7583 \times 0.68 = 37347.56$

We used Gradient Boosting as our final model for the future scoring file in 2.1

The parameters chosen for Gradient Boosting are:



On the Future Scoring file, we performed the same preprocessing as done on the PVA file followed by applying the model.

Following steps were performed:

1. Performing data cleaning by removing the missing attributes and generating new ones.
2. Selection of the variables using PCA
3. Using weights from a random forest to generate census variables.
4. Applying different models to find the maximum profit, which was found by the gradient boosted trees.
5. Applying model for Target B to predict the donors and non-donors by using futureFundingraising.csv. 1 represents donors and 0 represents non-donors Amongst 20000 donor's data, the model classified 7583(37.91%) predictions as donors by the model used in 2.1.

Therefore, we proceed with TARGET_B modelling as it maximizes the expected donation and the total number of donors.