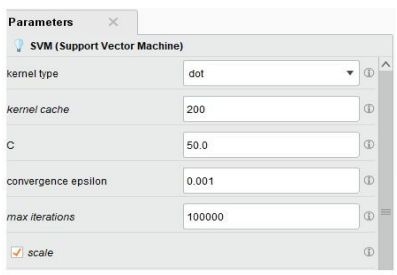
Target Marketing –Part 2

1. <u>Support Vector Machine and Gradient Boosted Trees:</u>

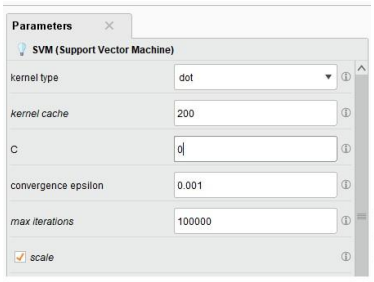
Portioned the dataset into 60% training - 40% validation and set the seed to 12345. To select the subset of variables to include in the SVM model went through multiple combinations of attributes in to find the best recall for true positives. Implemented dot, polynomial and radial kernels for SVM while designing the model. The parameters chosen are shown below. Chose the model with the best recall on positive predictions.

Used different values of C from 0 to 50. This specification assists the SVM to avoid mis-classifying the training data.



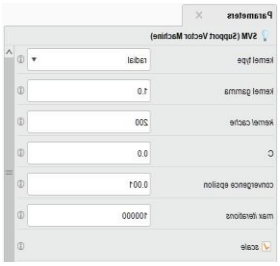
accuracy: 76.62%

	true false	true true	class precision
pred. false	6448	1954	76.74%
pred. true	14	2	12.50%
class recall	99.78%	0.10%	



accuracy: 74.72%

	true false	true true	class precision
pred. false	6181	1847	76.99%
pred. true	281	109	27.95%
class recall	95.65%	5.57%	



accuracy: 76.76%

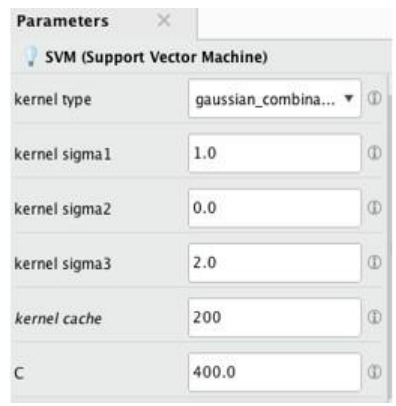
	true false	true true	class precision
pred. false	6462	1956	76.76%
pred. true	0	0	0.00%
class recall	100.00%	0.00%	

Similarly, for radial kernel- used different values of C, all values from 0 to 5 were tested.
On comparison it is seen that the recall for class positives are low.

Calculations: Given the cost of each mailing is $0.68. The profit gained per donation would be $12.32. Since the false negative, $12.32 is high we shall evaluate using recall.

The attributes and parameters of our best running model are shown below.

AVGGIFT
CARDGIFT
INCOME
WEALTH1
WEALTH2
avgGapBetwGifts
homeOwner
interestsPC1
interestsPC2
interestsPC3
interestsPC4
interestsPC5
interestsPC6
interestsPC7
interestsPC8
isMajor
isMilitary
totDays



Selected these attributes and parameters as it gave us the best SVM result with an accuracy of 63.44% and a true positive recall of 20.65%.

accuracy: 63.44%

	true false	true true	class precision
pred. false	4936	1552	76.08%
pred. true	1526	404	20.93%
class recall	76.39%	20.65%	

Comparison:

We can see the performance of our previous best fit model – Logistic Regression with **λ = 0.01** from our previous assignment on this data. We received an accuracy of only 44.70%.

accuracy: 44.70%

	true false	true true	class precision
pred. false	2246	439	83.65%
pred. true	4216	1517	26.46%
class recall	34.76%	77.56%	

Below are the parameters chosen for the Gradient Boosted tree. We see an accuracy of 75.82% with this classification function.

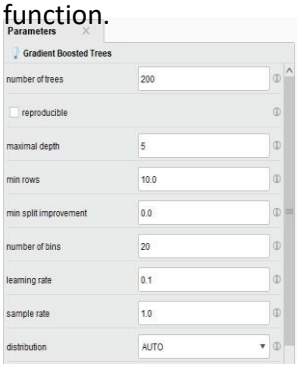


Table View Plot View

accuracy: 75.82%

	true false	true true	class precision
pred. false	6884	196	97.23%
pred. true	2857	2691	48.50%
class recall	70.67%	93.21%	

Our overall goal is to identify which individuals to target for maximum donations (profit). We have estimated the profits using different models below:

Model 1: Logistic Regression:

The parameters that we used to design the model is as follows.

The following confusion matrix was obtained for the logistic regression model that we designed

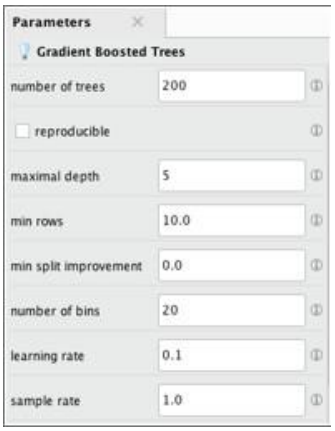
Accuracy: 23.3%

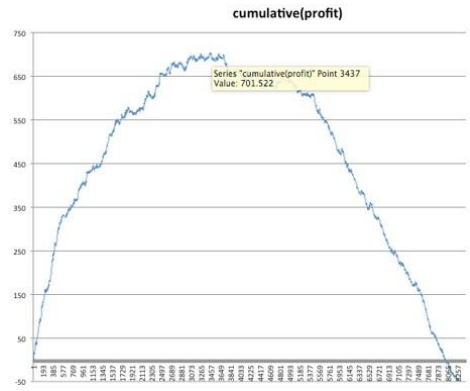
	True False	True True	Class Precision
Pred. False	16	3	84.21%
Pred. True	6446	1953	23.25%
Class Recall	0.25%	99.85%	

The observation were recorded in an ordered list (Confidence(1)). The maximum predicted profit for the model was calculated to be $779.

Model 2: Gradient Boosted Tree

The parameters that we used to design the model are as follows.





The following confusion matrix was obtained for the Gradient Boosted model that we designed

Accuracy: 53%

	True False	True True	Class Precision
Pred. False	3203	687	82.34%
Pred. True	3259	1269	28.03%
Class Recall	49.57%	64.88%	

The observation were recorded in an ordered list (Confidence(1)). The maximum predicted profit for the model was calculated to be $703.

Model 3: Random Forest

The parameters that we used to design the model are as follows.

The flowing confusion matrix was obtained for the Random Forest model that was designed

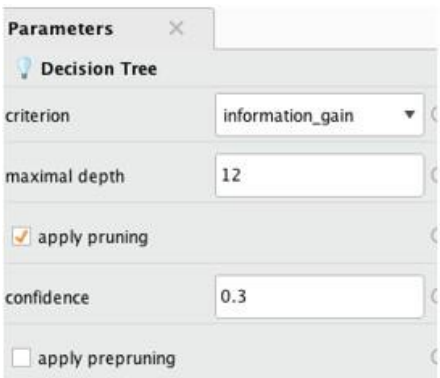
Accuracy: 54%

	True False	True True	Class Precision
Pred. False	3374	736	82.09%
Pred. True	3088	1220	28.32%
Class Recall	52.21%	62.37%	

The observation were recorded in an ordered list (Confidence(1)). The maximum predicted profit for the model was calculated to be $664.

Model 4: Decision Tree

The parameters that used to design the model are as follows

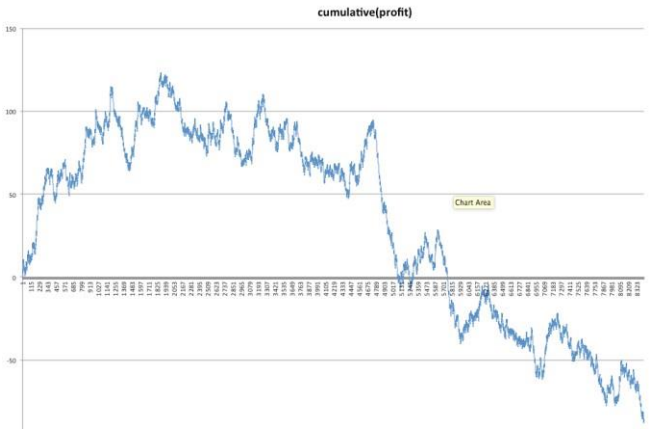


The flowing confusion matrix was obtained for the Decision Tree model that we designed

Accuracy: 42%

	True False	True True	Class Precision
Pred. False	2062	494	80.67%
Pred. True	4400	1462	24.94%
Class Recall	31.91%	74.74%	

Model 5: SVM



The flowing confusion matrix was obtained for the SVM model that was designed

Accuracy: 34.18%

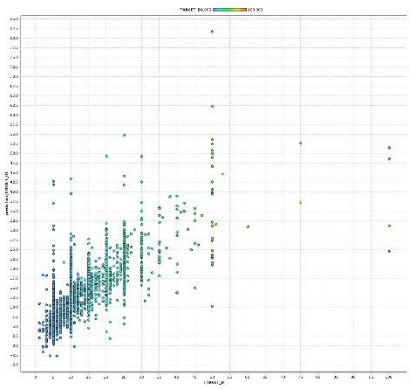
	True False	True True	Class Precision
Pred. False	1300	379	77.43%
Pred. True	5162	1577	23.40%
Class Recall	20.12%	80.62%	

Chose Logistic Regression as the best model and it has a maximum profit of $778 and an accuracy of 23.3%.

If we have to combine response as well as donation amount information to identify the individuals to solicit, We have to use a classification model and a prediction model and multiply the probability. Here we take logistic regression and gradient boosted model and multiply the probability of predicting a donor with the predicted value of Target_D which we obtain from the prediction model. ***The output of this multiplication gives us the cumulative profit curve.***

In OLM, Linear Model:

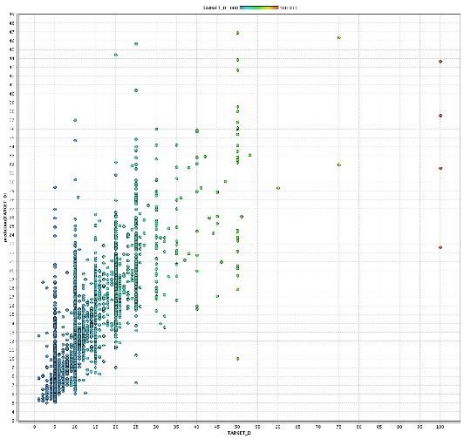
In pre-processing we would eemove cases where MAXARAMT > 99 AND TARGET_D = 0 .

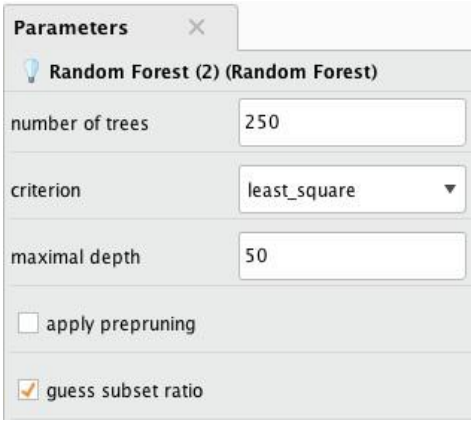


In Random Forest:

In pre-processing we would remove cases where MAXARAMT > 99 AND TARGET_D = 0

root_mean_squared_error: 6.768 +/- 0.000

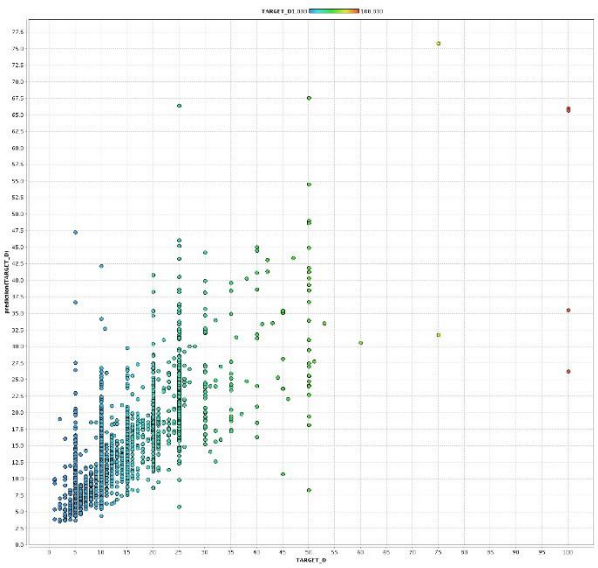


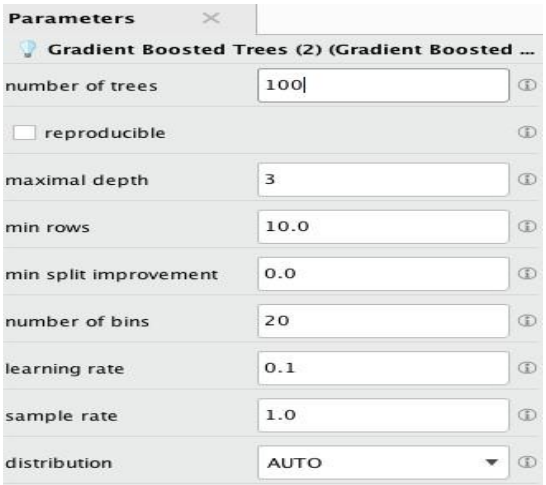


In Gradient Boosted Trees:

In Pre-processing we would remove cases where MAXARAMT > 99 AND TARGET_D = 0

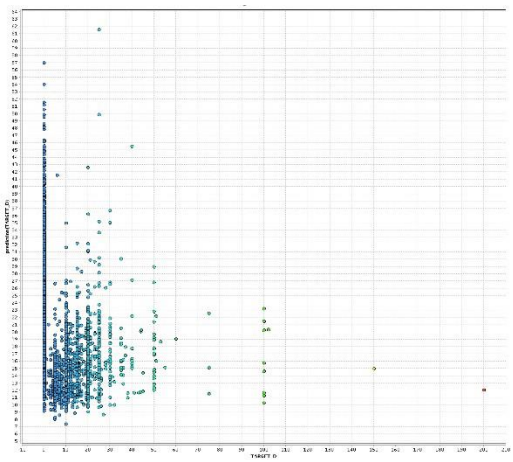
mean_squared_error: 6.690 +/- 0.000





We have chosen the donation amount from the gradient boosted trees model. The confidence values of our class predictions were multiplied with our classification model and the following process was obtained:

Profit Obtained was calculated to be **$18735** and the Target_D only profit was calculated to be **$2240**. The root mean square error was found to be **19.731+/-0.000**. From the gradient boosted tree model with an accuracy of **69%.**



Th best fit model has been identified as the Logistic Regression model. Now the same model has been implemented on the pva_futureData_forScoring.csv which contains the attributes for the future mailing candidates.

The cut-of value used to predict donor/non-donor for the validation threshold is 0.2 obtained form the cumulative gain for the trained model at confidence level1 at 0.2

Out of 20,000 records, the model has predicted 10,925 as donors. Using the above cut-off we obtain the maximum profit with 10,925 predicted donors and the output excel file has been attached with this document.