Target Marketing

ASSIGNMENT 3

Honey Salve | 669167842

Objective:

We are given a dataset for Paralyzed Veterans of America (PVA) fundraising and we have to develop various models to estimate the number of donors. We build models using Decision trees, Random forests, Logistic regression (using Ridge and Lasso), Boosted trees and SVM models. We need to develop a model to predict the donation amount i.e.TARGET_D and see how to combine the response and donation amount models to identify the profitable target individuals.

Question 1 - Modelling:

The dataset from assignment 2 is partitioned into training and testing datasets with a 60:40 split (with random seed set to 12345). The random seed is set to ensure that we obtain the same random partitioning every time we run the dataset. With no specified seed, the system clock is typically used to set the seed, and a different partitioning can result in different runs. We will develop support vector machine models for classification with examine different parameter values.

Dataset cleaning & exploration: We have already cleaned the dataset and handled the missing values in the first part in assignment 2.
Missing values: Earlier, we treated missing values using average, median and mode values. We have converted the variable data types into categorical variables and created new corresponding variables depending on the situational need.

Report on what you experimented with and what worked best:

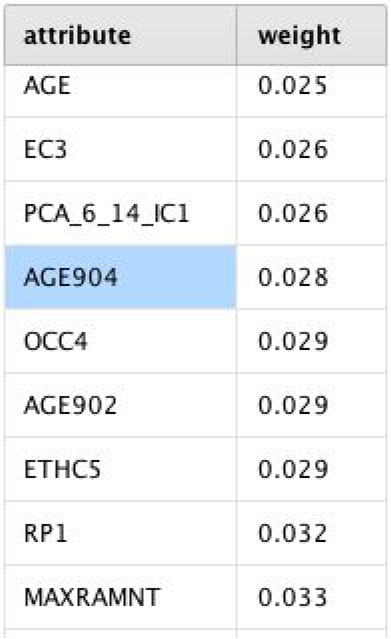
MODELS	TRAINING		TESTING	
	Accuracy	True Recall	Accuracy	True Recall
Random Forest	94.54%	5.2%	67.95%	4.18%
Logistic Ridge regression	63.5%	14.2%	63.34%	11.42%
Logistic Lasso Regression	61.37%	16.98%	61.24%	13.61%
Gradient Boosting	91.50%	51.29%	69.48%	45.67%

How do you select the subset of variables to include in the SVM model?

To start with SVM, we'll filter the variable selection to best fit the model with the following:

- Eliminate variables based on intuition
- Create corresponding new variables
- Transform variables
- PCA analysis

Correlation plot for variables:
For variable selection, we used the correlation plot from assignment 2 with TARGET_B variable which uses weight by correlation operator to predictor variables by its correlation with the label variables.
Weights of correlation for few variables with the target variable TARGET_B:



attribute	weight
AGE	0.025
EC3	0.026
PCA_6_14_IC1	0.026
AGE904	0.028
OCC4	0.029
AGE902	0.029
ETHC5	0.029
RP1	0.032
MAXRAMNT	0.033

After careful observation and considering the variables and their relevance, we decided to eliminate these variables:

ADATE_2-ADATE_24, AFC2-AFC6, AGEFLAG, DATASRC, DOB, FISTDATE, GEOCODE2, HHPAGE3, HPHONE_D, IC15-IC23, LASTDATE, LIFESRC, MAILCODE, MINRDATE NEXTDATE, NOEXCH, RAMNT_3-RAMNT_24, RDATE_3-RDATE_24, RFA_2-RFA_24, RFA_2A, RFA_2F, TCODE, WEALTH2

We need to divide the variables in somewhat similar groups. We have implemented the following four PCAs:

Interests_PCA: Here we used the 17 variables that depicted interests/hobbies of people which might help us in predicting the likeliness of them donating. Attributes used were BIBLE, BOATS, CARDS and so on...

Neighbourhood_PCA: 59 variables are used related to people's nativity, ancestry. This helped us reduce the number of relevant attributes.

GOV_PCA: 6 variables like FEDGOV, LOCALGOV, etc grouped as the federal or government associates.

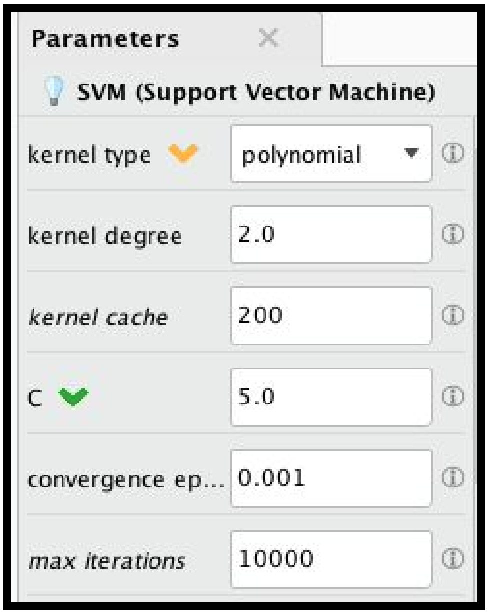
Promotion_PCA: 25 variables - RFA related values are considered to analyze based on the designation of the people in society.

This PCA process did help us a lot in compressing redundant information into few variables.

SVM model:
Why we use SVM? Well, it helps us to classify the dataset into two classes where each shows the max margin in the data. In our SVM model, we have used the below kernel methods:

- Dot kernel
- Polynomial kernel

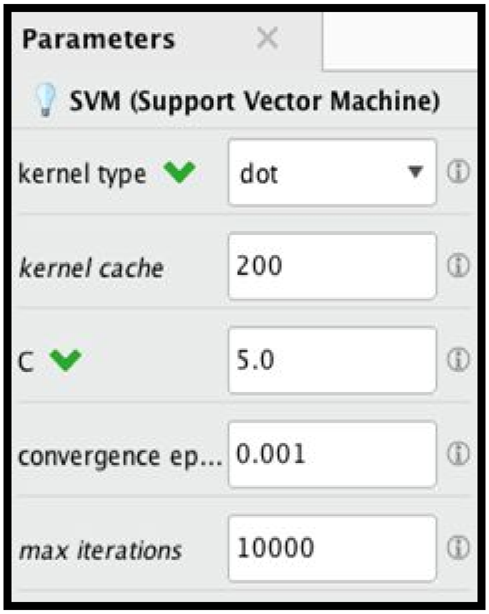
Out of these, the Polynomial kernel method is found to give the most positively optimal results. These parameters generate the optimal results for best model:



Parameters

SVM (Support Vector Machine)

kernel type	polynomial
kernel degree	2.0
kernel cache	200
C	5.0
convergence ep...	0.001
max iterations	10000



Parameters

SVM (Support Vector Machine)

kernel type	dot
kernel cache	200
C	5.0
convergence ep...	0.001
max iterations	10000

Interpretation - Comparing the Dot and polynomial kernels:

	SVM (With all PCAs)		SVM (Without PCAs)		Without interest PCA	
	Dot	Polynomial	Dot	Polynomial	Dot	Polynomial
Training	28.09	69.35	75.77	38.28	75.85	78.52
Testing	26.11	67.42	73.64	37.77	71.26	73.21
Testing- True Recall	83.15	33.34	14.51	74.21	16.43	11.33

Dot kernel: From these values, we get the highest recall performance for true 1 for dot kernel (With all PCAs), but the accuracy is not at the peak.
Polynomial kernel: For the polynomial kernel with all PCAs, the true recall performance is low in comparison to Dot kernel but here the accuracy is higher.
By keeping in mind the trade-off between accuracy and recall, we can see that training and validation data accuracy is not that different for polynomial and dot kernels.
Based on these observations, we decided to select the Polynomial kernel parameter with all PCA's as our best model.

Provide a comparative evaluation of performance of your best models from all techniques:

MODELS	Training	Testing	Testing - Recall (True)
Logistic Ridge Regression	63.5	63.34	14.2
Logistic Lasso Regression	61.37	61.24	16.98
SVM	69.35	67.42	33.4
Gradient Boosting	91.50	69.48	51.29
Random Forest	94.54	67.95	5.2

Here, we can see that the Gradient Boosted Trees model gives the best performance since its recall value is higher even though the accuracy is lower.

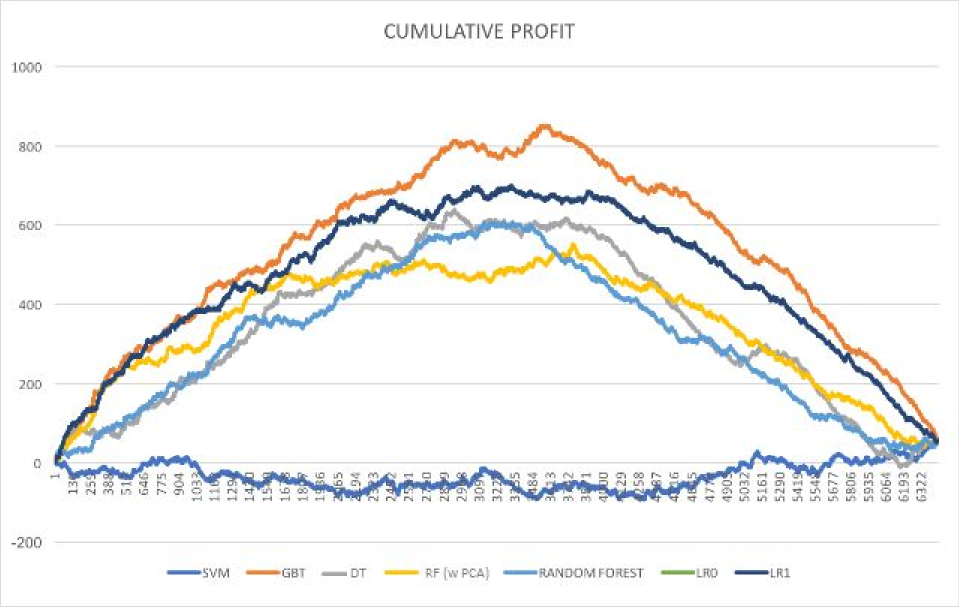
Question 2.1

What is the 'best' model for each method in Question 1 for maximizing revenue? Summarize the performance of the 'best' model from each method, in terms of net profit from predicting donors in the validation dataset; at what cutoff is the best performance obtained?

In order to calculate estimated net profit, we need to adjust the profit value and total cost of mailing based on actual distribution. Below is the calculation required to undo the effect of weighted sampling:

Average profit per successful response = $13
Cost of sending a mail = $0.68
Profit = 13 -0.68 = $12.32
Now,
The proportion of donors to non-donors in the original dataset = 5.1% to 94.9%
The proportion of donors to non-donors in the given dataset = 21% to 79%

Therefore, adjusted values of given dataset,
Donor = (12.32 * 0.05)/0.21 = 3.08
Non-donor = (12.32 * 0.95)/0.79 = 0.8075



MODELS	Training Accuracy	Testing Accuracy	Testing - Recall (True)	Testing Max Profit	Cut off Value
Logistic Ridge Regression	63.5	63.34	14.2	681.3	0.2
Logistic Lasso Regression	61.37	61.24	16.98	681.31	0.2
SVM	69.35	67.42	33.4	48.2	0.4

Gradient Boosting	91.50	69.48	51.29	**836.51**	**0.2**
Random Forest	94.54	67.95	5.2	591.26	0.2

As observed from the above figure, the maximum revenue is obtained from the **Gradient Boosted Trees**, with a value of **$836.51.** Hence, we're going to consider GBT model for further calculations. The cut-off value for the selected model is **0.2**.

For confirmation of our decision, we can refer to the Cumulative Net Curve graph below. As expected, the lift curve for the Gradient Boosted Tree, in orange, has the highest peak. This confirms that our decision is correct and that GBT is the optimal model.

The optimal model has the following parameters:
- Number of trees: 20
- Maximal depth: 3
- Minimum rows: 10.0
- Minimum split improvement: 0.0
- Number of bins: 25
- Learning rate: 0.2
- Sampling Rate: 1.0
- Distribution: Multinomial

Question 2.2

(a) We want to combine response as well as donation amount information to identify the individuals to solicit. Explain what approach you will take.

We perform the following steps:

1. Develop a classification model using TARGET_B as the dependent variable.
2. Calculate probability P(Donor | X).
3. Develop a regression model using TARGET_D as the dependent variable, only for data points with TARGET_B = 1.
4. Multiply P(Donor | X) and E(Donation).
5. Compare the result of Step 4 with $0.68. If the value is greater than 0.68 then person is "probable donor"; if the value is lesser than 0.68 then person is "probable non-donor".

(b) Develop a model for the donated amount (TARGET_D). What modeling method do you use (report on any one). Which variables do you use? What variable selection methods do you use? Report on performance.

For a regression model based on donation amount, that is, the TARGET_D variable, we observe that it would be optimal if we do not consider all the records. TARGET_D contains values only for individuals who are donors (i.e. TARGET_B =1). Therefore, non-donors (i.e. TARGET_B = 0) should not be included in the model to predict the donation amount.

Split Ratio of dataset used for Training and Validation = 60:40

We have used the Gradient Boosted Tree method mentioned above for building the regression model. The variable selection methods and therefore, the selected variables are the same as that of the SVM model.

(c) Based on your approach as explained in answer to 2.2 (a) above, combine the results from the response model and the donation_amount model to get an estimate of expected donation. Identify individuals to solicit and determine profit for the training and for the test set. Report your results on using the best response model from each method (as in Q 2.1 above), with the single donation_amount model. Do you notice performance differences? Do all/any of your models do better the no-model case? How does performance using this approach compare with what you saw in Q 2.1?

We multiply the results of the classification model [P(Donor|X)] and results of the regression model [E(Donation|X)] in order to obtain the target individuals. The spreadsheet attached below has the values.

We consider an individual as a donor if the predicted amount of donation is greater than $0.68, as explained earlier.

Hence, based on this model the number of potential donors are: 3017 out of 6445.

Profit obtained using TARGET_D model is given below –

For Validation Set,
Sum(Profit) = 17,227.814
Sum(tgtDonation) = 20,687.457

For Training Set,
Sum(Profit) = 42,597.633
Sum(tgtDonation) = 51, 246.952

For No-Model Case,

In this case, we'll send a solicitation mail to all the individuals on our mailing list. The profits calculated for the above test data would account for only 40% of the total profit as the dataset is split 40:60 into Training and Testing.

Calculating profit,
Number of individuals on mailing list = 6445
Maximum Profit = 20,687.457 – 6445*0.68 = 16,304.857

Question 3

Testing – chose one model, either the one from 2.1 or 2.2 above, based on performance on the test data. The file FutureFundraising.xls contains the attributes for future mailing candidates. Using your "best" model from Q 2, predict each example as donor or non-donor. Submit an xls file with two columns - the unique identifier and your prediction. (please maintain the same order of examples as in the FutureFundRaising file) The data in this file will correspond to the natural response rate of 5.1%. Will you adjust your model scores in any way – please explain what you do.

After analyzing all the above models, we decided to choose the Gradient Boosting model.
How do we predict the donors and non-donors?

- First, we handle the missing data values
- Variable selection
- Select subset of variables (2.1)
- Apply this model to the target variable TARGET_B and then to TARGET_D

This will give you the predicted donation amount.

- Get the net donation amount = confidence * expected donation amount (from above step)
- Set cut-off as0.68 for the expected donation amount.

We employ this model to get the prediction for the 'Future fundraising list'.
From this, we get the below result:

- Number of Donors = 7034

Donors_NonDonors.x
lsx