IDS 572 - Data Mining for Business

Assignment 3: Target Marketing Fundraising (Contd.)

Team Members:

Mrunal Ghorpade (677441117)

Neha Chimata (655196210)

Tanvi Sethi (672107527)

1. Modelling - Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345). In the last assignment, you developed decision tree, logistic regression, random forest and boosted tree models. Now, develop support vector machine models for classification. Examine different parameter values, as you see suitable. Report on what you experimented with and what worked best. How do you select the subset of variables to include in the model? What methods do you use to select variables that you feel should be included in the model(s)? Does variable selection make a difference? Provide a comparative evaluation of performance of your best models from all techniques (including those from part 1, ie. assignment 2)

In the last assignment, we performed modelling on the dataset with 60-40 split ratio and obtained best models using logistic regression with LASSO. The observations from the last assignment are as below:

MODELS	VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
		TRAINING	TEST	TRAINING	TEST
LASSO	Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood)	51.47	50.65	71.68	69.55
RIDGE	Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood)	54.83	52.59	71.15	66.96
Random Forest	Base Variables+PC of Donor Interests+ PC of Donor's response to other mails+ PC of Donor's Neighborhood	78.52	79	0	0
AdaBoost	Base Variables+PC of Donor Interests+ PC of Donor's response to other mails+ PC of Donor's Neighborhood	78.52	79	0	0
J48	Base Variables+PC of Donor Interests+ PC of Donor's response to other mails	93.21	69.01	78.71	21.73
Naïve Bayes	Base Variables+PC of Donor Interests+ PC of Donor's response to other mails	57.82	56.98	55.73	52.85

Adaboost and Random forest gave static results for all the subsets of variables selected and recall rate = 0. So, they were not picked for comparison of best models. While developing models from support vector machines, we took 7 cases defining different parameters for each case.

	CASE 1	CASE 2	CASE 3	CASE 4	CASE 5	CASE 6	CASE 7
Kernel Type	Dot	Polynomial	Polynomial	Polynomial	ANOVA	ANOVA	Radial
Kernel Degree		2	2	1	1	1	
kernel cache	0	0	0	0	0	0	200
C	5	5	7	7	7	6	3
Convergence Epsilon	0.1	0.1	0.2	0.15	0.2	0.2	0.001
max iterations	100	100	100	200	100	100	100

Scale	No	No	No	No	No	No	Yes
Kernel Gamma	-	-	-	-	1	2	0.05
Balance Cost	No	No	No	No	No		Yes

Based on the above cases, we developed models for different subsets of the variables chosen during variable selection described in the last assignment i.e. PCA and Randomforest. The SVM model with each case was applied on all the base variables and performance was analysed. Then principal components were added subsequently to the base variables to analyse their performance. The output for different cases and variable subset is as follows:

ATTRIBUTES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
Base Variables Case1	43.61	43	53.08	55
Base Variables Case2	78.52	79	0	0
Base Variables Case3	78.52	79	0	0
Base Variables Case4	39.55	41	59.1	62
Base Variables Case5	51.21	48	64.79	56
Base Variables Case6	76.38	76	9.2	6.5
Base Variables Case7	31.97	26	100	94
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) (Case1)	76.28	77	6.07	6.8
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 2	56.46	56	42.97	41
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 3	56.8	56	41.85	40
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 4	67.77	68	19.7	21
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 5	64.21	64	28.32	27
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 6	58.46	58	48.7	46
Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood) Case 7	75.39	73	19.61	11
Base Variables+PCA1+PCA2 (Case1)	23.75	24	95.57	96
Base Variables+PCA1+PCA2 (Case2)	78.52	79	0	0
Base Variables+PCA1+PCA2 (Case3)	78.52	79	0	0
Base Variables+PCA1+PCA2 (Case4)	75.82	76	6.55	7.8
Base Variables+PCA1+PCA2 (Case5)	60.48	57	45.52	38
Base Variables+PCA1+PCA2 (Case6)	76.53	76	8	5.4
Base Variables+PCA1+PCA2 (Case7)	31.97	26	100	93
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case1)	74.88	76	8.48	9.8
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case2)	56.18	56	42.97	41
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case3)	78.09	79	1.06	1.3
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case4)	76.55	77	5.64	5.9
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case5)	43.82	44	67.97	67

Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case6)	64.47	64	27.26	27
Base Variables+PCA2 + Random Forest (Donor's Neighborhood) (Case7)	38.8	27	91.52	83

Among the above outputs, the top three best cases are as follows:

1. SVM with radial kernel type (as highlighted) gave the best recall rate in both training and test data, with base variables and principle components of donor's interests and donor's neighboorhood to other mailers.
2. SVM model with kernel type dot and the same variables.
3. SVM with kernel type ANOVA on Base variables, Principal components of donor's response and neighbourhood was the third best case.

Support Vector Machine with subset of variables (Base variables+PCA2+Random Forest) gave the best recall rates and lowest accuracy rates when compared to other models. While in terms of accuracy and recall rates balanced against each other, LASSO and RIDGE models are the better models.

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? We can calculate the net profit from given information - the expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68. Note: to calculate estimated net profit (on data with the 'natural' response rate of 5.1%), we will need to "undo" the effects of the weighted sampling, and calculate the net profit that reflects the actual response distribution of 5.1% donors and 94.9% non-donors. Draw profit curves: Draw each model's net cumulative profit curve for the validation set onto a single graph. Are there any models that dominate? Best Model: From your answers above, what do you think will be the "best" model to implement? (What criteria do you use to determine 'best'?)

Calculating the Net profit:

The sample given has 21.5% donors (TARGET–B = 1) and 78.5% of non-donors (TARGET–B = 0). Therefore, we need to adjust the cost of mailing and donation amount to match it with the actual sample distribution i.e. 5.1% for donors (TARGET–B = 1) and 94.9% for non-donors (TARGET–B = 0).

From the question we have,

Cost of Mailing = $0.68

Profit from donation = $13 - $0.68 = $12.32

We adjust the cost as below:

Cost of Mailing = $0.68 * (0.949 / 0.785) = 0.8221

Profit from donation = $12.32 * (0.051 / 0.215) = 2.9224

The above value indicates that after adjustment, the cost of mailing is $0.8221 and the donation amount is $ 2.9224, for the dataset with a distribution of 5.1 – 94.9 ratio.

We use various criteria like accuracy on training and testing data, recall on testing data, maximum profit obtained, and threshold value selected to maximize the performance of the selected model inorder to choose one best model amongst the chosen models in Question 1.

Below table summarizes the performances of the models:

PARAMETERS	LOGISTIC WITH LASSO	LOGISTIC WITH RIDGE	RANDOM FOREST	ADA BOOST	W-J48	NAÏVE BAYES	SVM (Case 7)	SVM (Case 5)	SVM (Case 1)
Accuracy (Training)	21.65%	22.47%	78.52%	78.52%	80.61%	22%	30.85%	21.66%	21.66%
Accuracy (Testing)	21.17%	21.91%	78.35%	79%	58.97%	21.38%	26.68%	21.10%	21%
Recall (Training)	100%	99.81%	3.25%	0%	91.71%	99.95%	98.84%	100.00%	100%
Recall (Testing)	99.93%	98.97%	1.77%	0%	42.94%	99.63%	92.61%	100.00%	100%
Max profit (Testing)	600.85	596.965	22	1.278	163.69	306.95	263.19	274.26	176.27
Threshold	0.075	0.214	0.215	0.215	0.08	0.469	0.445	0.241	0.195

Logistic with LASSO:

- Logistic Regression with LASSO gives highest Recall Rates for Training as well as Test data indicating that more number of predicted donors are accurate.
- Maximum profit obtained with LASSO is maximum amongst all other models, but the threshold is nearly equal to zero which means that it is no better than a no model case.

Logistic with Ridge:

- Logistic Regression with Ridge gives high Recall Rates for Training as well as Test data indicating that more number of predicted donors are accurate.
- Maximum profit obtained is also high.

Random Forest:

- Random Forest models have the least recall rate and Maximum Profit.

Adaboost:

- Adaboost models have 0% recall rate i.e. none of the donors are predicted correctly which makes this model not optimal for our analysis.

W-J48:

- Training and Testing data set performances are vastly different and lower Class 1 Recall indicating that only a few number of predicted donors are accurate.
- It has very low threshold.

Naïve Bayes:

- High Class 1 Recall along with consistency in both testing and training data sets indicates that Naïve Bayes is a good model.

Support Vector Machine:

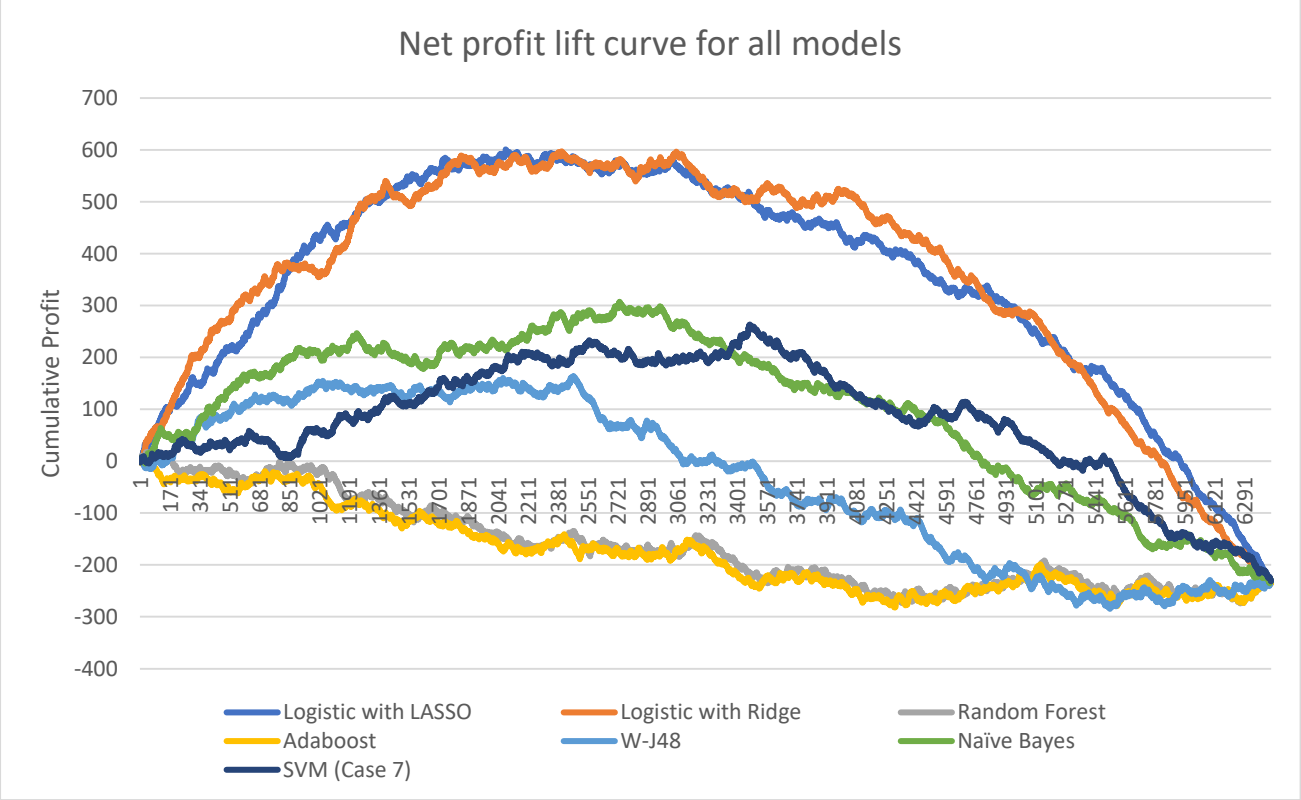
- For SVM we have tried 3 of our best models mentioned in Question 1 with different parameters. Among those, we choose SVM with case 7 as our best model. It has high recall rates, though Recall rates for other chosen SVM models are 100% they do not have Maximum Profit as SVM with case 7 and has accuracy more than others. Also, it has threshold neither high nor low making it ideal it's ideal.

To compare the best model's, we also looked at their lift charts to know "by how much" is a model better when compared with no models. If we are targeting the top decile then, lifts for our best models are calculated as given below:

Lift = (Max Profit obtained from model) / (Profit or Loss when no model is applied)

PARAMETERS	LOGISTIC WITH LASSO	LOGISTIC WITH RIDGE	RANDOM FOREST	ADA BOOST	W-J48	NAÏVE BAYES	SVM (Case 7)	SVM (Case 5)	SVM (Case 1)
Training	35.37%	35.50%	8.33%	21.48%	26.73%	31.13%	43.74%	28.75%	28.22%
Test	33.18%	41.16%	50.00%	21.00%	100.00%	28.99%	23.88%	27.29%	24.82%

Below is the graph depicting the profit curve obtained from the best model of each of the modelling techniques employed. (Note: Graphs are plotted only for validation data)



Observation:

The model which dominates the net profit lift curve and has the highest peak on the Cumulative Profit curve is the Logistic regression (with both Ridge and LASSO).

Best Model:

Logistic regression with Ridge model is the best model, with the best true 1 recall rate and profit values and Lift in first decile.

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.

Logistic regression with Ridge is the classification model with best accuracy, True 1 recall rate and the profit figures from our analysis. Therefore, we will be considering the same for further analysis.

General Linear Modelling technique can be used for the prediction of the donation amount (TARGET_D).

The Confidence (1) value obtained from classification model (after rescaling using Platt Scaling) needs to be combined with the prediction for donation amount (TARGET_D) and the same is done using the following logic:

Estimate for expected donation = Confidence (1) [Obtained from classification model] * TARGET_D [Predicted donation amount obtained from Linear regression]

Only those donation amounts that are less than $50 will be considered as regression is sensitive to outliers. Since the donation amounts exceeding $50 in our datasets are very minimal, the donations exceeding $50 are considered outliers.

2.2 (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. Note that TARGET_D has values only for those individuals who donors (that is, TARGET_D values are defined only for cases where TARGET_B is 1). What data will you use to develop a model for TARGET_D? (Non-donors, obviously, do not have any donation amount -- should you consider these as $0.0 donation, or impute missing values here? Should non-donors be included for developing the model to predict donation amount? Also, should cases with rare very large donation amounts be excluded?

After merging the data from Target B model post rescaling the confidence of prediction B, we set the role of Target D as label and select variables for TargetD modelling. Here variable selection was done using correlation by weights operator which sorted the variables by weights in descending order and we selected weights above or equal to the threshold of 0.03.

To get rid of the outliers, target D with values more than 0 and less than equal to 50 were taken. By doing this we also filter the data for Target_B =1 since only the subset which is a donor would give the donor amount. Linear regression, Gradient boosting and General linear models were applied on the training data but, General linear model gave the best model with root mean squared loss of 17.726 on the test data. The parameters and the outputs are as below:

PARAMETERS	VALUES
Family	Gaussian
Link	Family_default
Solver	Auto
Use Regularization	Yes
Lambda	0.0001
Standardize	Yes
Add intercept	Yes
Remove collinear columns	Yes

Output:

GENERALIZED LINEAR MODEL	TRAINING	TEST
Root Mean Squared Error	5.921	17.726

2.2 (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?

We combined the response and the donation model for all the response models used in 2.1. And the outputs against each combination is displayed on the following page. The combination of LASSO model with the donation model gave the maximum total profit on the test data i.e. $ 13575.11. On the training set the total profit was $42943.47 and the individuals to solicit came out to be 4832. Also, there is an observation of profit if only Target_D was considered.

Expected Donation = [confResponse]*[prediction(TARGET_D)]

Profit = if (expDonation > 0.68, TARGET_D -0.68, 0)

Profit with Target_D only = if (TARGET_D > 0, TARGET_D - 0.68, 0)

The no model case does not consider the cost of mailing while calculating profits. However, on modelling we take into consideration many factors that give us a good estimation of profits. The no model case gave $16151.63 which is far more than the value expected if we consider all the costs. So, none of the models are better than the no model case.

	METRICS	TRAIN	TEST
RIDGE	Profit	47930.83	13574.43
	Profit with Target_D only	47943.47	16151.63
	Individuals to solicit		4833
LASSO	Profit	47943.47	13575.11
	Profit with Target_D only	47943.47	16151.63
	Individuals to solicit		4832
W-J48	Profit	47747.19	13960.59
	Profit with Target_D only	47943.47	16151.63

	Individuals to solicit		4171
RANDOM FOREST	Profit	33267.32	13575.79
	Profit with Target_D only	33269.96	16151.63
	Individuals to solicit		4831
ADA BOOST	Profit	33267.32	13575.79
	Profit with Target_D only	33269.96	16151.63
	Individuals to solicit		4831
NAÏVE BAYES	Profit	23960.42	9736.09
	Profit with Target_D only	33269.96	16151.63
	Individuals to solicit		3221
SVM CASE7	Profit	47806.51	13657.95
	Profit with Target_D only	47943.47	16151.63
	Individuals to solicit		4619

There was no significant difference among the performance of the all the combination model except the one with Naïve Bayes model. All observations on the test data are nearly the same. However, there was a difference in the training data profit observations for the models.

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 3 scores in any way – please explain what you do.

Logistic regression with LASSO from question 2.2 is chosen for prediction as it had a better performance for the validation dataset compared to the model employed in Q2.1.

We built a model on data which had 20% as donors, however we are applying the model to a dataset with a response rate of 5.1%. Hence, we rescaled the confidence levels of the predictions obtained through the TARGET_B model.

Prediction of donors and non-donors is done by using the sample file (pva_futureData_forScoring) given in the question on the chosen best model.

Observation:

Number of donors = 4203

Number of Non-donors = 15797

Cut-off used to predict donors and non-donors is 0.68.

For Expected donations above 0.68 we get prediction for TARGET_B as 1 and the rest as 0 (please refer excel sheet). All data points below the cut off value are considered non-donors.

Attached is the file that contains the value of ControlN and Prediction for Target B.

Output for future fundraising v1.0.xlsx