Anchit Jhingan (663246540)

Dushyant Singh Khinchi (662632496)

Pallavi Srikanth (664941288)

1. **Modeling Consider the training/test data partitions and the balanced training data as in the last assignment. In the last assignment, you developed random forest, logistic regression and gradient 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 SVM model? Provide a comparative evaluation of performance of your best models from all techniques (including those from part 1, ie. assignment 2) (Be sure NOT to include “TARGET−D” in your analysis.)**

In the first part of the assignment, we started by reducing the variables we were going to use for modelling by going through the data dictionary and removing those variables that we felt did not seem relevant for predicting the response.

After reducing the variables through manually, we were left with 88 variables. We then used Random Forest and GBM techniques to understand the variables importance and selected the 37 most important variables on which performed Principal Component Analysis.

We finally selected the top 15 principal components on which we built models to predict TARGET\_B.

Below is a comparison of each model based on accuracy and recall for both the training and test data:

*Training Data:*

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Recall** |
| **Lasso (lambda - 1se)** | 79.92% | 0.00% |
| **Lasso (lambda - min)** | 80.22% | 5.07% |
| **Ridge (lambda - 1se)** | 80.27% | 5.83% |
| **Ridge (lambda - min)** | 80.46% | 8.07% |
| **Random Forest** | 98.37% | 6.7% |
| **GBM** |  |  |

*Test Data:*

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **Recall** |
| **Lasso (lambda - 1se)** | 79.33% | 0.00% |
| **Lasso (lambda - min)** | 79.31% | 0.57% |
| **Ridge (lambda - 1se)** | 79.32% | 0.58% |
| **Ridge (lambda - min)** | 79.18% | 0.58% |
| **Random Forest** | 95.62% | 0.68% |
| **GBM** |  |  |

1. **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. 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’?)**

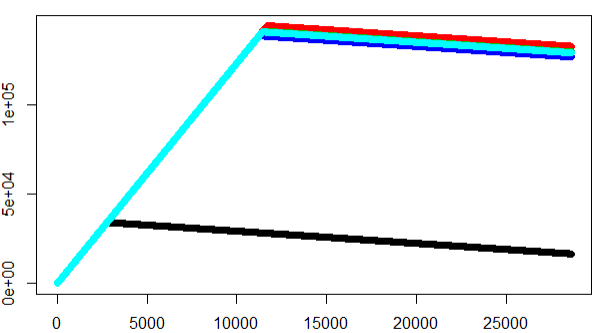
For the above techniques, to measure the net profit on the test set, we considered different thresholds of 0.2, 0.3 and 0.5 for classifying the response of 0 or 1, based on the probabilities. We observed that a threshold of 0.2 gave us the maximum profit as compared to the other thresholds. To calculate the cumulative profit, after classifying the response as 0/1, we sorted the output of the models based on probability from high to low and allocated a profit of $12.32 for all 1 responses and loss of -$0.68 for all the 0 responses as given below:

* Response 1: $12.32 - Expected donation is $13.00 however the cost of mailing is $0.68. Thus, our total profit is $13.00 - $0.68 = $12.32
* Response 0: -$0.68 - Expected donation is $0 however the cost of mailing is $0.68. Thus, our total profit is $0 - $0.68 = -$0.68

Below is the comparison of each model’s maximum profit for thresholds 0.2, 0.3 and 0.4

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Threshold | | |
| 0.2 | 0.3 | 0.4 |
| Ridge (lambda - min) | $144,883.20 | $22,594.88 | $2,020.48 |
| Ridge (lambda - 1se) | $146,275.40 | $9,671.20 | $61.60 |
| Lasso (lambda - 1se) | **$152,841.90** | $21,904.96 | $1,884.96 |
| Lasso (lambda - min) | $145,770.20 | $4,607.68 | $36.10 |
| Random Forest | $33,658.24 | $3,092.32 | $258.72 |

The cumulative profit curves for the models have been given below:

****

Our criteria for judging the best model is the max profit obtained by implementing a certain threshold. From the numbers above, we are obtaining maximum profit with Lasso (lambda 1se) at a threshold of 0.2, thus this is our best model.

**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. Note that you need to have well-calibrated probability scores before ‘combining’ the response and donation amount models – explain what you do for this. (b) Develop a model for the donated amount (TARGET\_D). What modeling method give ‘best’ performance (consider OLS, random forests, gbm). 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 have donated (that is, TARGET\_D values are defined only for cases where TARGET\_B = 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 we not omit non-donors for developing the model to predict donation amount? Also, should cases with rare very large donation amounts be excluded? [ϑleading questionsϑ] (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?**

For creating a model to predict the donation amount, we decided to include the probability values from our best model, Lasso (with lambda 1se) as in input in the model. While this one input came from the TARGET\_B model, we decided to also use the 37 variables obtained by performing variable importance analysis using Random Forest and GBM in part 1 of the assignment.

While looking at the summary of the data, we observed the following statistics for the donation amounts:

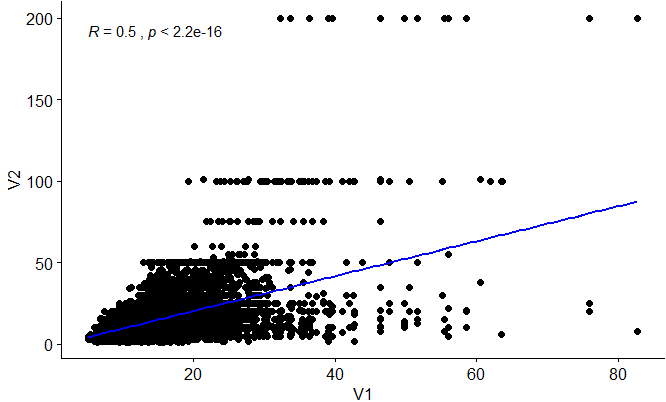
|  |  |
| --- | --- |
| **Statistics** | **Donation Amount ($)** |
| Minimum | 0 |
| 1st Quartile | 10 |
| Median | 13 |
| Mean | 15.52 |
| 3rd Quartile | 20 |
| Maximum | 200 |

We also decided to omit the non-donors (i.e. those individuals for whom TARGET\_B=0) from the data used to build the donation amount prediction model. However, we did not omit any large donation amounts from the model training data as we wanted to allow the model to be trained on all the available data as the percentage of TARGET\_D=1 response individuals was already low.

We created models using GBM, RF and OLS. To understand the performance of each model, we decided to compute the R.M.S.E for each model as given in the table below:

|  |  |  |
| --- | --- | --- |
| **Model** | **Train** | **Test** |
| **GBM** | 7.88 | 7.98 |
| **RF** | 7.44 | 7.61 |
| **OLS** | 7.87 | 7.91 |

With the Random Forest model having the lowest R.M.S.E for both the training and test data, we also looked at the correlation between the prediction from the RF model and the actual donation values. As we obtained a correlation of 50%, we considered the Random Forest as our ‘best’ model.



V1 – Model prediction, V2 – Actual donation amount

To obtain the expected donation amount, we multiplied the prediction from the RF models with the probability values from our best model in Part 1. As the cost of mailing is $0.68, we will target only those individuals who’s expected donation amount as greater than $0.68. We have attached the result from the RF model, along with the probability values and expected donation amount in the workbook below:



1. **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.**

For predicting the response on the FutureFundRaising data, we decided to go with the approach from 2.2. We first cleaned the data by imputing missing values and changing categorical variables to factors. We selected the same 37 variables that we assessed were the most important from the analysis done in the first part of the assignment and performed principal component analysis. Using the top 15 principal components and ‘best’ model from part 1 of the assignment – Lasso – we computed the probability of each candidate being a donor and classified them as 0: Non-donor or 1: Donor, using the same threshold of 0.2. Using this computed probability as another input variable, we then used the ‘best’ model from 2.2, which was Random Forest and computed the donation amounts. The donation amount corresponding to each unique identifier is given in the workbook below:

