IDS 572 Data Mining Assignment -3

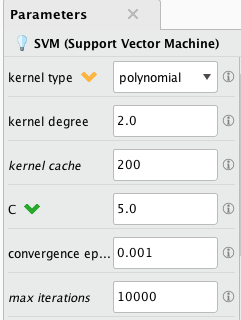
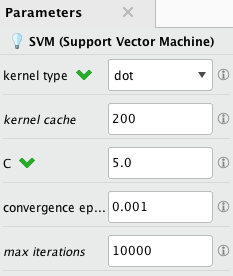
Q1) Modeling Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345). In the last assignment, you developed decision tree, logistic regression, random forest and boosted tree models. Now, develop support vector machine models for classification. Examine different 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)

Answer1)

We have tried the SVM models for classification on our preprocessed data from Assignment 2 and we used two kernels “DOT” and “Polynomial”. Our findings are listed below in the table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | SVM(All PCA) | | SVM(No PCA) | | PCAs without Interests and Hobbies | |
|  | Dot | Polynomial | Dot | Polynomial | Dot | Polynomial |
| Training | 29.09 | 69.38 | 75.77 | 38.28 | 75.85 | 78.52 |
| Testing | 27.13 | 68.45 | 73.64 | 37.77 | 71.26 | 73.21 |
| Testing-Recall (True1) | 84.11 | 33.38 | 14.51 | 74.21 | 16.43 | 11.33 |

**Finding:**

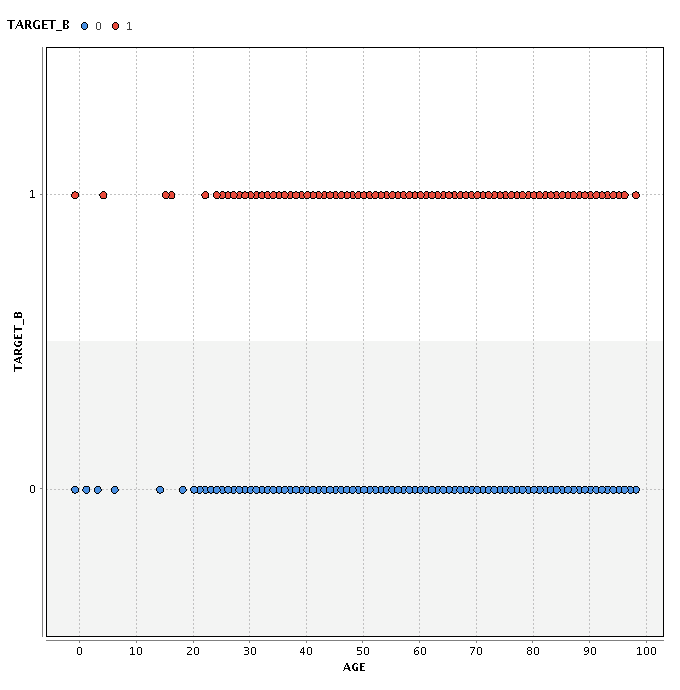
From the table above we compared the dot and polynomial (degree 2) kernels. For dot kernel(All PCA), we get the highest recall performance for true 1 but not the highest accuracy. In the polynomial kernel(All PCA) the recall performance is low as compared to the dot kernel but accuracy is much higher. However, after considering trade-off between accuracy and recall we decided there is less variance between the training data and validation data accuracies for polynomial as compared to the dot kernel. We select the model in which all PCA’s are included to be our best model with polynomial kernel parameter.

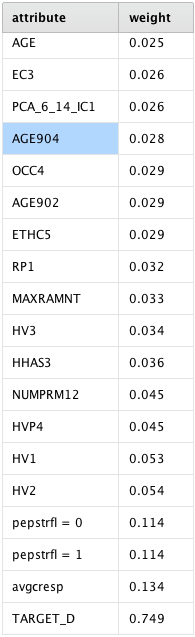
Methods used for variable selection:

1. Variable elimination
2. Variable Mapping
3. Variable Transformation
4. PCA’s

For selecting the variables we found the correlation of variables used in assignment 2 with TARGET\_B variable by using weight by correlation operator which gives weight to predictor variables by their correlation with the label variables and decide upon their selection. Below is the weights of correlation of a few variables with TARGET\_B.

We also used scatter plots for finding out important variables -





After carefully observing and considering the above two approaches we decided to eliminate the following variables:

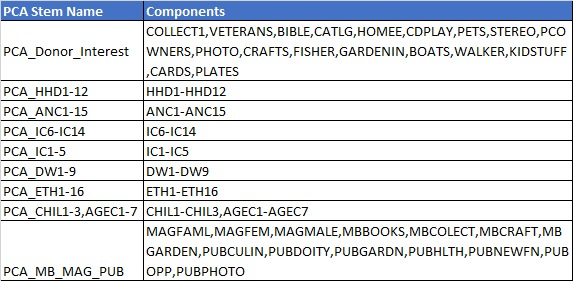
ADATE\_2-ADATE\_24, AFC2-AFC6, AGEFLAG, CHILC1-CHILC5, DATASRC, DOB, FISTDATE, GEOCODE2,HHPAGE3, HPHONE\_D, IC15-IC23, LASTDATE, LIFESRC, MAILCODE, MINRDATE

NEXTDATE, NOEXCH, ODATEDW, OSOURCE, PVASTATE, RAMNT\_3-RAMNT\_24, RDATE\_3-RDATE\_24, RFA\_2-RFA\_24, RFA\_2A, RFA\_2F, SOLP3, STATE, TCODE, WEALTH2, RECP3,RECSWEEP

The following variables are mapped and transformed:

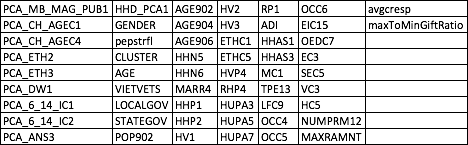
RECINHSE, DOMAIN, PEPSTRFL, CHILD03, CHILD18, HOMEOWNWER, CARDPROM, CARDGIFT, NGIFTALL, NUMPROM, LASTGIFT, MAXRAMNT, MINRAMNT, CHILD07, CHILD12, MAJOR, RECPGVG

We performed the following PCA’s and divided the attributes in similar groups as follows:



Variable selection did make a lot of difference in handling very large number of attributes that contain the similar information. Principal Component Analysis helped us compressing the overlapping information of many similar variables into 2-4 variables at most.

Final list of variables for SVM Model -



**Comparison of best models from all the techniques is given below:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MODELS USED** | **Logistic** | **Naïve Bayes** | **WJ48** | **SVM** | **RF** | **Gradient Boosted Trees** |
| Training | 78.56 | 71.8 | 81.11 | 69.4 | 78.5 | 66.73 |
| Testing | 77.97 | 65.9 | 73.82 | 68.5 | 79 | 62.66 |
| Testing-Recall (True1) | 16.98 | 20.62 | 12.12 | 33.4 | 5.2 | 51.29 |

Looking at the table above we concluded the best performance is given by **Gradient Boosted Trees**. Although the accuracy is lower than the Logistic Regression but the recall is much higher.

Also, in Naive Bayes, WJ48, SVM, and Random Forest the accuracy is higher than the gradient boosted trees but the recall is compromised.

**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?**

**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’?)**

Answer 2.1 ) The maximum revenue obtained from below table is $852.51 . So , we take Gradient Boosted Trees as our best model and proceed with it further.

TABLE - Summarization of max revenue from each model

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **MODELS USED** | **Logistic** | | **Naïve Bayes** | **WJ48** | **SVM(All PCA)** | **RF** | **GBT** |
| MEASURES | Alpha=0 | Alpha=1 |  | C=0.5;M=20 |  |  | T=20;D=3;bins=25;LR=0.2 |
| Training | 78.59 | 78.56 | 71.8 | 81.11 | 69.4 | 78.52 | 66.73 |
| Testing | 76.9 | 77.97 | 65.9 | 73.82 | 68.5 | 79 | 62.66 |
| Testing-Recall (True1) | 14.2 | 16.98 | 20.62 | 12.12 | 33.4 | 5.2 | 51.29 |
| Testing Max Profit | 699.3 | 699.31 | 552.02 | 637.897 | 56.2 | 608.132 | 852.51 |

Calculation involved to undo the effect of weighted sampling -

Average profit per successful response → $ 13.

Cost of sending per mail → $ 0.68

Total profit: 13 – 0.68 → $ 12.32

Proportion of Donor to Non-Donors in the original dataset - 5.1% to 94.9%

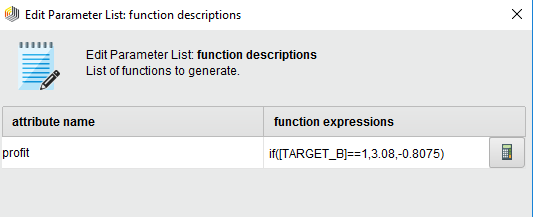
Proportion of Donor to Non-Donors in the weighted dataset - 20% to 80%

Adjusted values for weighted dataset

Donor → ($ 12.32 \* 0.05) / 0.20 = 3.08

Non Donor → ($12.32 \* 0.95) / 0.80 = 0.8075

Snapshot of calculation done in RapidMiner -



Cut-off value obtained for best model using Gradient Boosted Trees is 0.2 . Please refer below table for reference -

TABLE - Cutoff values for each model

|  |  |  |
| --- | --- | --- |
| **Model** | **Cumulative Max Profit** | **Cut-off Value** |
| Random Forest | 608.1 | 0.2 |
| SVM | 56.2 | 0.4 |
| J-48 | 637.9 | 0.2 |
| NaiveBayes | 552.3 | 0.1 |
| GBT | 852.5 | 0.2 |
| LR(Alpha=0) | 699.3 | 0.2 |
| LR(Alpha=1) | 699.3 | 0.2 |

**Criteria used to select best model -**

By observing cumulative profit curve(below), we inferred that model which dominates the net profit lift curve and has the highest peak is the Gradient Boosted Trees. Also , this model has the highest recall (true1) value equal to 51.29 as compared to other models that will help us to predict more number of donors accurately.

**Parameters used for our best model are :**

• Number of trees: 20

• Maximal depth: 3

• Minimum rows: 10.0

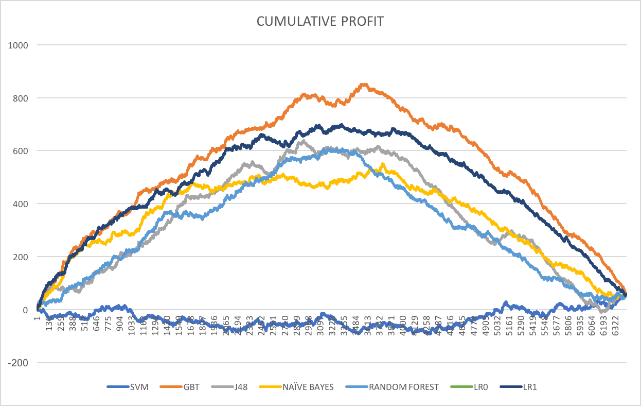
• Minimum split improvement: 0.0

• Number of bins: 25

• Learning rate: 0.2

• Sampling Rate: 1.0

• Distribution: multinomial

Figure - Cumulative Profit Curve for all models 

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.**

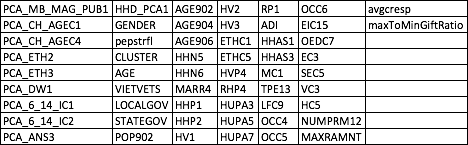
**Answer** - We develop a classification model using Target B and calculate probability P(donor|X) . Next , we develop a regression model using Target D only for data points which have Target B=1. Next, we multiply the probability and E(donation) and then compare the result with $0.68. If value is greater than 0.68 , then the person is probable donor and if value is less than 0.68 , it is a 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.**

**Answer** - For a model based on donation amount that is the TARGET\_D variable we observe that we do not have to take all the records. TARGET\_D has values only for individuals who are donors therefore non donors should not be included in the model to predict the donation amount.

Split Ratio used for Training and Validation = 60:40

We have used the same Gradient Boosted Tree method as obtained earlier to build this regression model. Variables used are the same(figure below) as included in SVM model. Variable selection methods also remains same as reported for 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?**

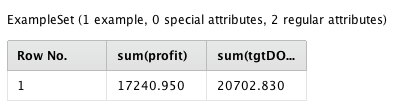
**How does performance using this approach compare with what you saw in Q 2.1?**

The results from the response(classification) model [P(donor|X)] and the result from the donation amount model [E(donation|x)] can be multiplied to identify the targets. Attached Spreadsheet has the values. Also we would prefer an individual as a donor if the predicted amount of donation is greater than $0.68 as it takes $0.68 to send a mail.

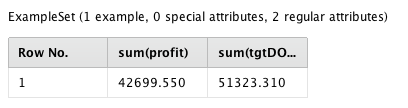
Hence based on this model the number of potential donors are: 3029 out of 6445

Profit obtained using TargetD model is given below -

Validation set



Training set



For the No Model case , we will mail solicitation to all the individuals in our data set. Since , profits obtained above for test data set are only 40%( 60:40 split we took for training and testing) , we calculate for the same.

Number of individuals = 6445

Maximum Profit = 20702 – 6445\*0.68 = 16319.4

Q3. **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.**

Answer -

With the finalized trained models in part 2.1 and 2.2, we went ahead with Gradient Boosting using following parameters:

**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

For the new FutureFundRaising.csv, we followed below steps for predicting the donors and non-donors:

1. Handling Missing Values for the new data
2. Variable selection using PCA’s
3. Selecting subset of attributes based on the attributes as selecting after Random Forest in 2.1
4. Applying model for Target B to predict the donors and non-donors
5. Applying model for Target D to get the expected donation amount
6. Multiply the confidence and expected amount to get the net expected donation amount.
7. Apply a filter to calculate the cost only if the Expected Donation amount is more than 0.68 which is the threshold value.

By using above steps we reached to a conclusion that out of 20000 donor population our model is able to predict a total of 7037(35.18%) donors which can maximize the expected donation.