Anchit Jhingan (663246540)

Dushyant Singh Khinchi (662632496)

Pallavi Srikanth (664941288)

**IDS 572 – Assignment 2**

1. **Data cleaning, exploration. The dataset has many variables – some (many?) of them may not be useful for our purpose. Your first task is to explore the data, determine missing values and how you might handle these, which variables you think need not be considered. This is a major task – and can take significant time, much more than the modeling step that comes next. (You will find below two tables with subsets of variables that were found useful in earlier analysis).**

**(a) Which attributes will you omit from the analyses and why? How do you approach this problem - for example, certain variables in the raw data may not be of interest, while for others, you should conduct some analyses (univariate analyses?) to determine which variables may be useful.**

We started reducing the number of variables by manually going through each given in the input data set. Variables that were not related to the analysis were removed based on their description and our general understanding. There were many variables that were present in the dataset along with several others which recorded the same data for a segment (e.g. AGE904 - Average age of population and AGE905 – Average age of population >=18). The main variables that were kept for the analysis while the variables related to the segmentation was removed.

The status of each variable after this first round of data exploration has been included in the excel workbook given below. With the manual data exploration, we reduced the variable count to 88 on which we preformed further analysis.



**(b) How do you clean the data, handle missing values?**

There were many variables in the data that used null for denoting negation/absence of attribute. To ensure that our model could accurately predict the response, we replaced these nulls with either character strings or numeric values. We also transformed some variables using single characters and null values to numeric values. We have listed the variables with missing values and the transformation done to them in the excel workbook below. We have also listed the different types of transformations done for your quick reference.

1. Character response (e.g. H/X/Y) to 1
2. Null to 0
3. Null to -1 (to differentiate where null has a meaning)
4. Null to ‘missing’ (for variables with character labels)

****

**(c) What new attributes/values do you derive?**

As given in the excel workbook above, we transformed the labels of some variables (e.g. HOMEOWNR) from a categorical value (e.g. H) to a numerical value (1). Further, we derived some new attribute so that we could discard the original variables. The list of derived variables has been given below:

1. Separate the two levels of the variable ‘DOMAIN’ – one for the urbanicity level of Donor and the other for the socio-economic status of the Donor into two separate variables, ‘domainU’ and ‘domainSES’ respectively
2. totweeks – the duration between ‘ADATE\_2’ (date when 97NK promotion was mailed) and ‘LAST\_DATE’ (date associated with most recent gift)
3. totmaxweeks – the duration between ‘ADATE\_2’ (date when 97NK promotion was mailed) and ‘MAXADATE’ (date of the most recent promotion received)
4. avgAllResp – average value of response to all card promotions: ‘NGIFTALL’ (no. of lifetime gifts to date)/’NUMPROM’ (lifetime number of promotions received till date)
5. LAST\_MAX\_GIFT – ratio of ‘LASTGIFT’ (dollar amount of most recent gift) and ‘MAXRAMNT’ (dollar amount of largest gift till date)
6. MIN\_MAX\_AMNT – ratio of ‘MINRAMNT’ (dollar amount of smallest gift till date) and ‘MAXRAMNT’ (dollar amount of largest gift till date)
7. **Which variables will you consider for modeling (and why)?**

**How do your findings relate with the variable subsets given in Tables 1 and 2?**

**Explain how you approach data reduction, variable selection? What methods do you try and find useful? Summarize your findings.**

**Consider the following approaches for variable selection, data reduction. Experiment with all these. Then describe which methods you use to obtain the set of variables to use in different models (in the next question). - Random forests and/or decision trees can help determine which variables to include in a predictive model for donors. - Univariate analyses, considering the value of each individual predictor of potential interest for predicting response (for example, through its AUC for predicting response). - Principal Components Analysis (PCA) can also help in data reduction.**

**Explain what you do, findings, and whether/how you think this is useful.**

Our approach to deriving the list of variables for modelling started with understanding the meaning and values of each variable – with this were able to decide which variables to include/exclude them from further processing. This resulted in the 88 variables as detailed in part (a). After dividing the data into training and test datasets (70%-30%) ratio, we then further decided to use Random Forest and Linear Regression on the list of 88 variables. Each method gave us a list of important variables on which we performed a union – 38 variables were thus kept from the 88 variables. Finally, on these 38 variables, we performed Principal Component Analysis (PCA) to obtain our dataset for modelling.

* *Variable selection using Random Forest:*

By building a Random Forest on our data, we were able to observe the importance of each variable. We selected the top 30 variables (out of 88) by highest variable importance as given below:

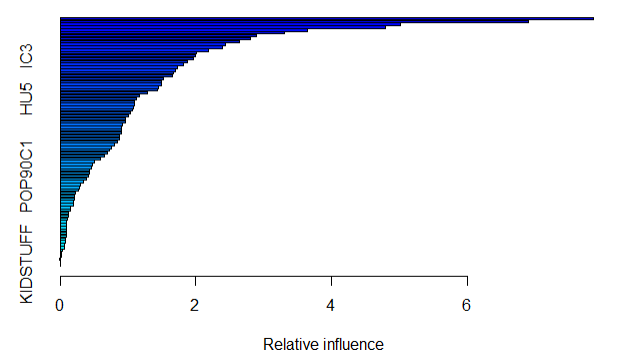
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Variable Name** |  | **Index** | **Variable Name** |
| 1 | LASTGIFT |  | 16 | POP903 |
| 2 | avgCardResp |  | 17 | AFC5 |
| 3 | CARDGIFT |  | 18 | AGE904 |
| 4 | HV2 |  | 19 | HHAS1 |
| 5 | IC5 |  | 20 | WWIIVETS |
| 6 | IC4 |  | 21 | AGE901 |
| 7 | NUMPROM |  | 22 | HHAGE1 |
| 8 | IC3 |  | 23 | EC3 |
| 9 | CARDPROM |  | 24 | NUMPRM12 |
| 10 | totWeeks |  | 25 | IC20 |
| 11 | LAST\_MAX\_GIFT |  | 26 | HU2 |
| 12 | MIN\_MAX\_AMNT |  | 27 | AGE |
| 13 | HHAS3 |  | 28 | HHAS4 |
| 14 | POP901 |  | 29 | LFC5 |
| 15 | POP902 |  | 30 | HU1 |

* *Variable selection using Linear Model*

Using Linear Models, we selected the top 30 important variables as given in the table below:

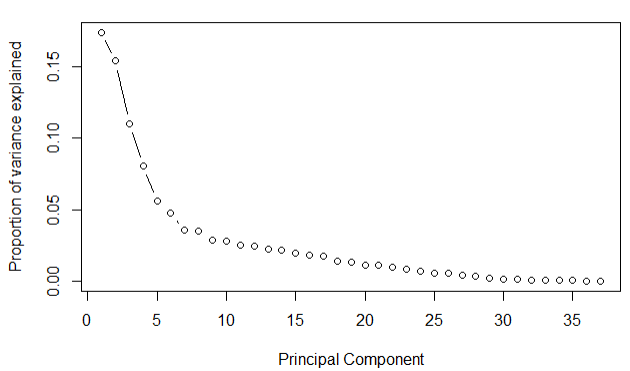
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Index** | **Variable Name** |  | **Index** | **Variable Name** |
| 1 | avgCardResp |  | 16 | LFC4 |
| 2 | LASTGIFT |  | 17 | CARDPROM |
| 3 | HV2 |  | 18 | CARDGIFT |
| 4 | totWeeks |  | 19 | IC5 |
| 5 | AGE |  | 20 | VIETVETS |
| 6 | NUMPRM12 |  | 21 | POP901 |
| 7 | CARDPM12 |  | 22 | EC3 |
| 8 | WWIIVETS |  | 23 | POP902 |
| 9 | STATEGOV |  | 24 | HHAGE1 |
| 10 | POP903 |  | 25 | IC4' |
| 11 | IC3 |  | 26 | HHAS3 |
| 12 | NUMPROM |  | 27 | MALEVET |
| 13 | LFC5 |  | 28 | INCOME |
| 14 | LAST\_MAX\_GIFT |  | 29 | PUBHLTH |
| 15 | MIN\_MAX\_AMNT |  | 30 | HU2 |

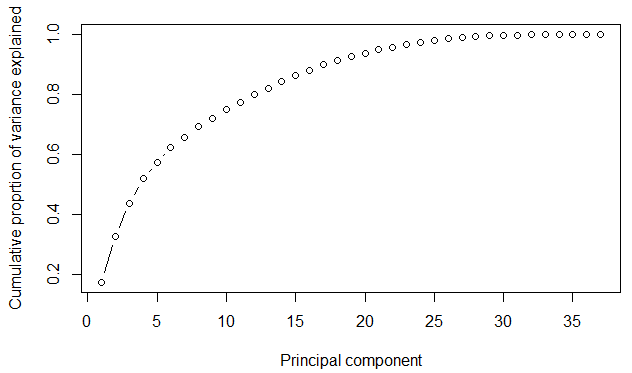
The graph given below shows the relative influence of the variable, obtained after fitting the linear model.

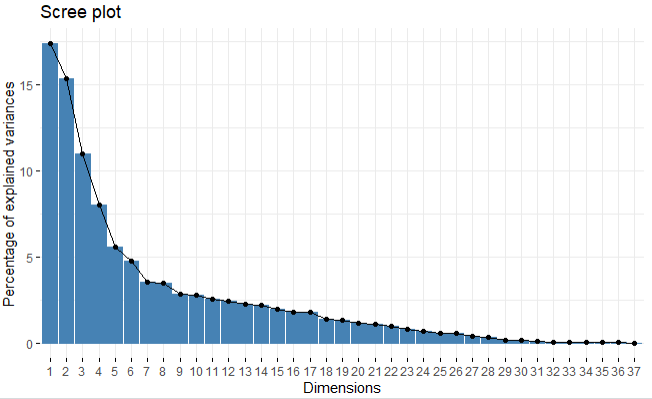


*Principal Component Analysis*

* Taking a union of the two sets of variables obtained via Random Forest and Linear models, we were left with 37 variables. We then performed Principal Component Analysis to transform the variables further. By observing the variance of the components and the skree plot, we decided to select the top 15 principal components on which to do our further analysis.



**



As shown in the ‘*Principal component vs Cumulative proportion of variance explained*’ given above, the first 15 principal components can account for approximately 80% of the total variance in the data.

The contribution of each variable to the top 15 principal components is given in the workbook below:



1. **Modeling. The data set has 5.1% responders. We will obtain a balanced training-set before developing models. For this, first split the data into training and test subsets (60:40 or 70:30), and then conduct the balancing on the training set. Note that the test data retains the original proportion of responders, which will be useful for assessing model performance on unseen data.**

**For balancing, we will use under-sampling, over-sampling, and a combination of these, using the ovun.sample function of the ‘ROSE’ package (see sample code). [https://www.rdocumentation.org/packages/ROSE/versions/0.0-3/topics/ovun.sample]**

**You should start the modeling with a training sample which has around 20-30% responders.**

**Consider the following classification techniques on the data:**

**• Logistic Regression, using Ridge and Lasso**

**• Random forest**

**• Boosted trees**

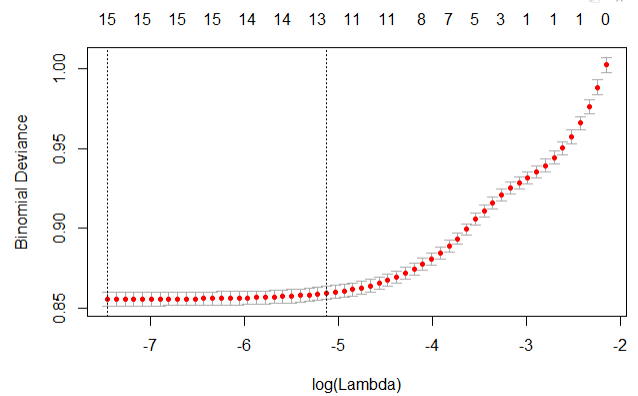
**• SVM**

**How do you determine which variables to include in the data for modeling with each of the methods above? Consider whether the different methods above incorporate mechanisms for variable selection?**

As the dataset has just 5.1% positive responses, there was a need to oversample the target variable to ensure that our models had enough positive responses for accurate prediction – for the below figures we have considered 20% over-sampling. We also removed TARGET\_D (which corresponded to the donation amount for a positive response) as this would have led to data leakage since this information would have not actually be present in a real-world scenario.

**Logistic Regression:**

Using logistic regression, we tried to obtain the best value of lambda that we could use for the ridge and lasso models. The two values we obtained were the (1) minimum value: 0.0005799787 and (2) 1se value: 0.005936262.



1. Lasso

Using an alpha value of 0, we obtained the following accuracy and recall on the training and testing data with the different values of lambda:

*Training Data:*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lambda min** | | | |  | **Lambda 1se** | | | |
| **Prediction** | **Reference** | | |  | **Prediction** | **Reference** | | |
|  | **Negative** | **Positive** |  |  | **Negative** | **Positive** |
| **Negative** | 52,987 | 12,728 |  | **Negative** | 53,381 | 13,408 |
| **Positive** | 484 | 680 |  | **Positive** | 0 | 0 |

|  |  |  |
| --- | --- | --- |
| **Lambda** | **Accuracy** | **Recall** |
| **Min** | 80.22% | 5.07% |
| **1se** | 79.92% | 0.00%\* |

\*For Lambda 1se, since the number of true positives is 0, the recall is also 0%.

*Test Data:*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lambda min** | | | |  | **Lambda 1se** | | | |
| **Prediction** | Reference | | |  | **Prediction** | Reference | | |
|  | **Negative** | **Positive** |  |  | **Negative** | **Positive** |
| **Negative** | 22,669 | 5,694 |  | **Negative** | 22,896 | 5,727 |
| **Positive** | 227 | 33 |  | **Positive** | 0 | 0 |

|  |  |  |
| --- | --- | --- |
| **Lambda** | **Accuracy** | **Recall** |
| **Min** | 79.31% | 0.57% |
| **1se** | 79.99% | 0.00%\* |

\*For Lambda 1se, since the number of true positives is 0, the recall is also 0%.

1. Ridge

Using an alpha value of 1, we obtained the following accuracy and recall on the training and testing data with the different values of lambda:

*Training Data:*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lambda min** | | | |  | **Lambda 1se** | | | |
| **Prediction** | Reference | | |  | **Prediction** | Reference | | |
|  | **Negative** | **Positive** |  |  | **Negative** | **Positive** |
| **Negative** | 52,659 | 12,326 |  | **Negative** | 52,831 | 12,626 |
| **Positive** | 722 | 1,082 |  | **Positive** | 550 | 782 |

|  |  |  |
| --- | --- | --- |
| **Lambda** | **Accuracy** | **Recall** |
| **Min** | 80.46% | 8.07% |
| **1se** | 80.27% | 5.83% |

*Test Data:*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lambda min** | | | |  | **Lambda 1se** | | | |
| **Prediction** | Reference | | |  | **Prediction** | Reference | | |
|  | **Negative** | **Positive** |  |  | **Negative** | **Positive** |
| **Negative** | 22,632 | 5,694 |  | **Negative** | 22,670 | 5,694 |
| **Positive** | 264 | 33 |  | **Positive** | 226 | 33 |

|  |  |  |
| --- | --- | --- |
| **Lambda** | **Accuracy** | **Recall** |
| **Min** | 79.18% | 0.58% |
| **1se** | 79.32% | 0.58% |

Between Ridge and Lasso, we are observing a better performance on the test data with Ridge regression using lambda min.

1. **Does extent of resampling make a difference in model performance? Develop and evaluate models with a higher proportion of response (40%-50%). Compare performance of models with those developed using the first balancing proportion (20%-30%). What do you conclude?**

When applying a resampling of 50%, we were able to see some changes in the performance of the models. While there was no noticeable difference to the Principal Component Analysis output (the top 15 principal components still attributed to ~80% of the variance), we did see a difference in the Ridge and Lasso models. We have compared the performance of resampling with 20% and resampling with 50% in terms of accuracy and recall on the test data as given below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Lambda** | **20%** | | **50%** | |
| **Accuracy** | **Recall** | **Accuracy** | **Recall** |
| **Lasso** | **Lambda min** | 79.31% | 0.57% | 50.09% | 5.27% |
| **Lambda 1se** | 79.99% | 0.00% | 50.14% | 1.26% |
| **Ridge** | **Lambda min** | 79.18% | 0.58% | 49.99% | 8.43% |
| **Lambda 1se** | 79.32% | 0.58% | 49.98% | 7.30% |

Overall, we can observe that on the test data, the accuracy has decreased while the recall values have increased across all models.