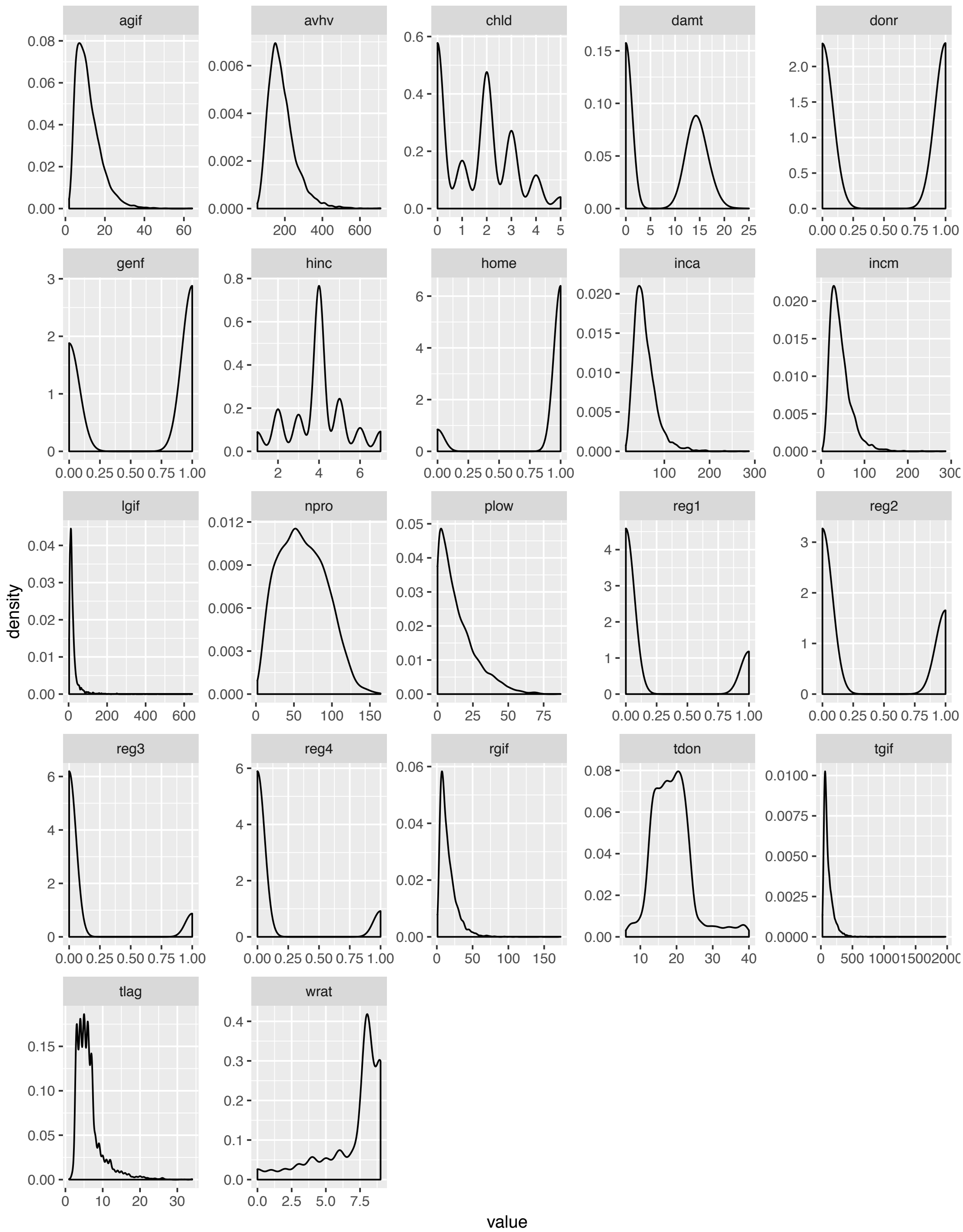
Group Project

Introduction

This research is based on the data generated from a charitable organization who wishes to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to previous donors. according to their recent mailing records, the typical overall response rate is 10%. Out of those who donate via the mailing, the average donation is $14.50. Each mailing costs $2.00 to produce and include gifts. The expected profit from each mailing is 14.50 x 0.10 – 2 = -$0.55. Our task consists of two parts, first part is to develop a classification model that can effectively captures likely donors so that the expected net profit is maximized; the second part seeks for the best model to predict the gift amount. The entire dataset consists of 3984 training observations, 2018 validation observations, and 2007 test observations. Weighted sampling has been used, over- representing the responders so that the training and validation samples have approximately equal numbers of donors and non-donors.

Analysis

There are 22 variables and 8,009 observations in the initial dataset with no missing data. The data were separated into three parts: training, validation and test. After drawing the density plot of each numeric variables (shown as follow), we found that most of variables need to be transformed. Since avhv is highly skewed, we transformed it to logistic mode. Additionally, we standardize them to provide convenience for later regression and prediction.

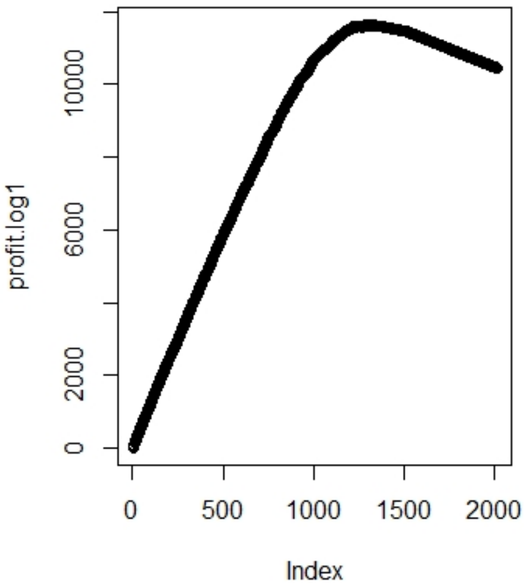


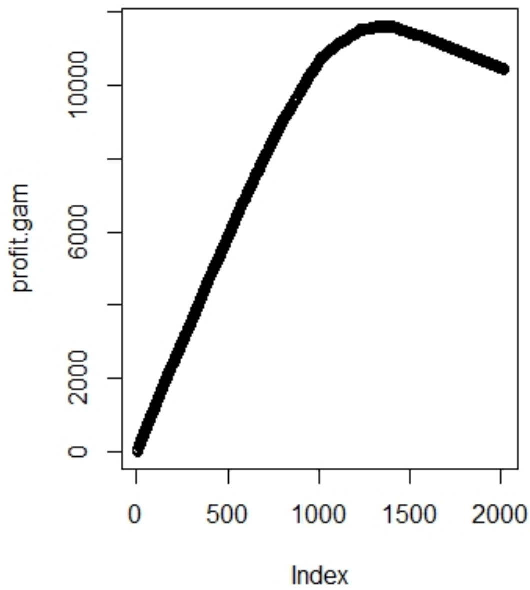
Part 1: Prediction modelling for DNOR

To select the best classification model for DONR variable, we fitted training dataset into seven different models including logistic regression, logistic regression GAM, LDA, QDA, k-nearest neighbors, random forest model and decision tree model. After running classification regressions, we used valid data to forecast DONR variable. And then we used predicted values to calculate profit under different conditions.

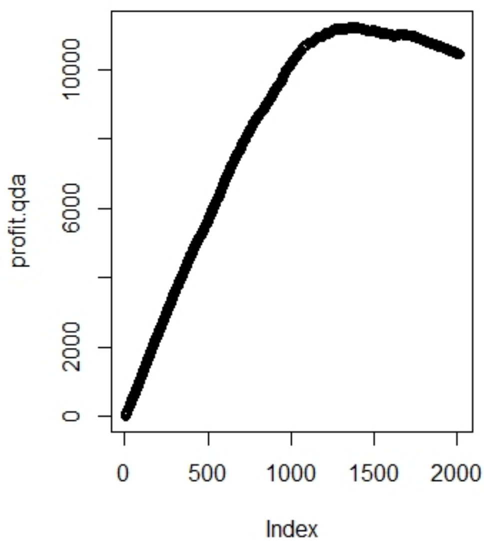
Specific results of each prediction are shown as below in the same order as above.

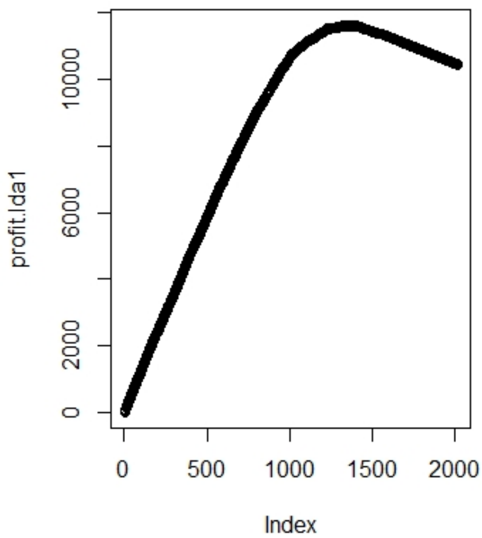
Logistic Model: Maximum profit earned with Logistic Model is 11640.5.





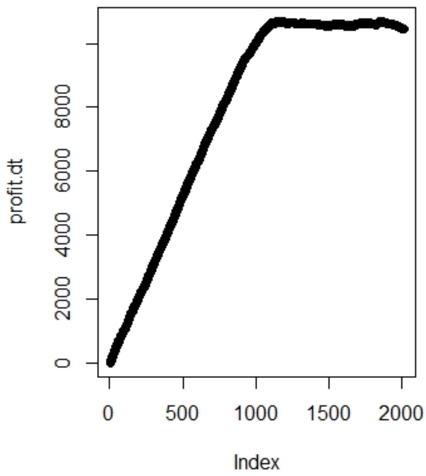
Logistic GAM Model: Maximum profit earned with Logistic GAM Model is 11624.5.





LDA Model: Maximum profit earned with LDA Model is 11624.5.

QDA Model: Maximum profit earned with Logistic Model is 11224.



Decision Tree Model: Maximum profit earned with Decision Tree Model is 10687.

```
> table(post.valid.rdmf, y.valid.donr)
               y.valid.donr
post.valid.rdmf   0   1
              0 885  95
              1 134 904
>
> table(post.valid.knn,y.valid.donr)
                    y.valid.donr
post.valid.knn         0   1
  -0.507455520244056 830 801
  1.97012143153575   189 198
```

We can see from graphs above that under each model prediction, predicted profit first increases gradually, reaches a certain level and then starts to decrease or remains the same. In general, when number of mail is about 1,000, the profit is at the top point.

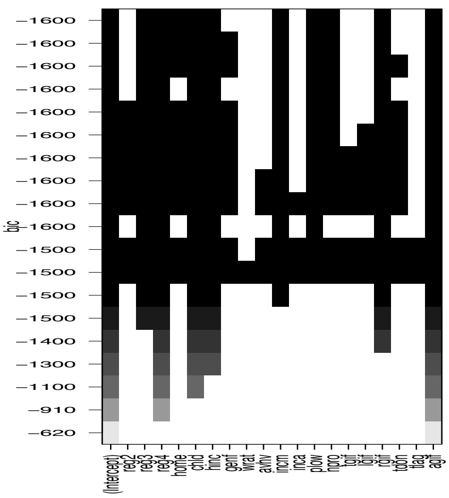
Part 2: Prediction modelling for DMAT

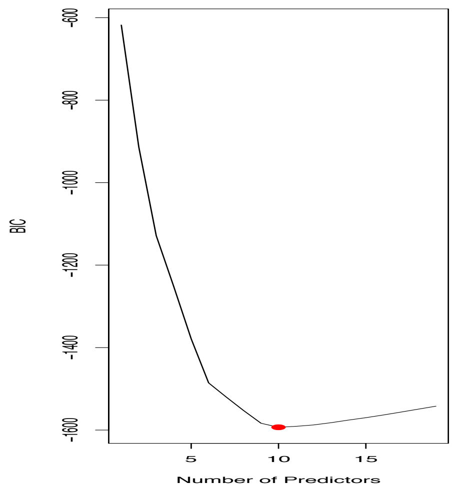
In this part, we Develop a prediction model for the DAMT variable using any of the variables as predictors (except ID and DONR). We fitted as many candidate models as we could generate, including least squares regression, best subset selection with bic, cp, adjusted R square, validation set approach and 5-fold cross-validation, principal components regression, partial least squares, ridge regression and lasso regression. We generate models with the training data and evaluate the fitted models using the validation data.

Least Square Regression

For the first model, we use OLS regression to fit a model using all 19 predictor variables. Eleven of them (agif, rgif, npro, plow, incm, hinc, chld, home, reg4 and reg3) are statistically significant with level for three stars. Then we check for the collinearity between predictors. The VIF result indicates that no obvious collinearity exists among variables, which can be interpreted as we need more data or we should try other models to select variables. We evaluated this model by calculating the validation data mean prediction error (1.8692).

Best Subset selection- using BIC model



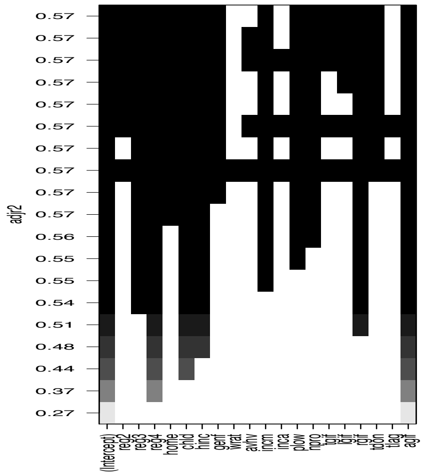


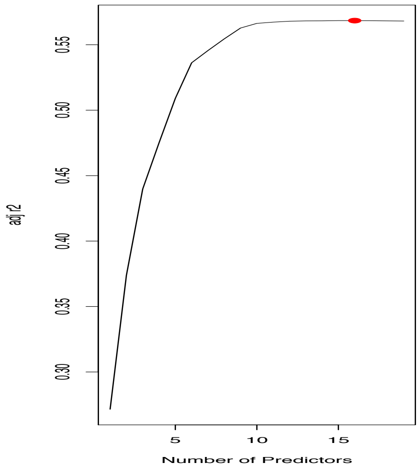
The BIC model suggests the optimal number of variables should be 10, as pictures above. The relevant variables are shown as below.

```
> coef(regfit.full, 10)
(Intercept)        reg3        reg4        home        chld        hinc        incm        plow        npro        rgif        agif
 14.1480296   0.3581844   0.6686119   0.2511736  -0.6297229   0.5012535   0.3166469   0.2578157   0.1856372   0.4926266   0.6552395
```

The mean prediction error of fitting this model with the validation data is 1.85794

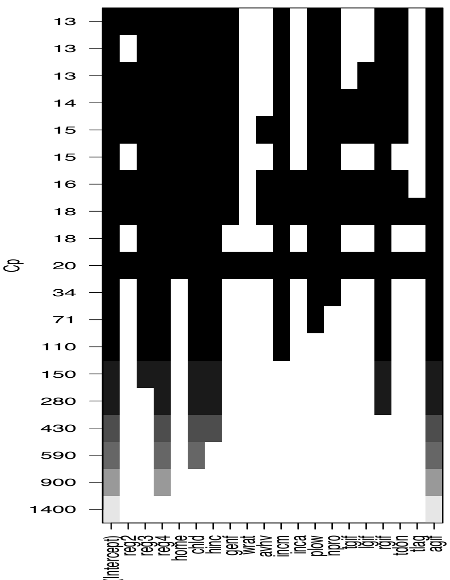
Best Subset selection- using Adjust R^2

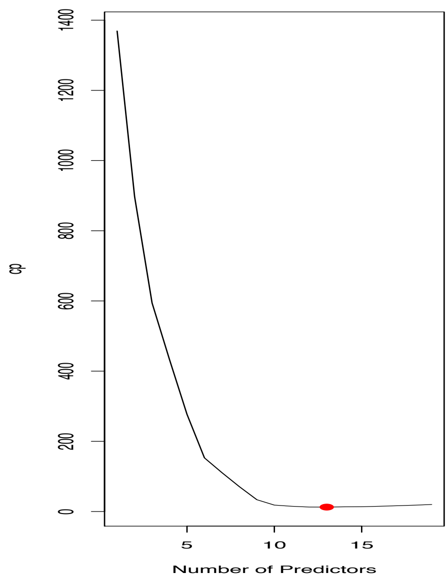




The adjusted R square model suggests the optimal number of variables should be 16, as pictures above. The relevant variables are shown as below. The mean prediction error of fitting this model with the validation data is 1.8680.

```
> coef(regfit.full, 16)
(Intercept)        reg2        reg3        reg4        home        chld        hinc        genf        avhv        incm
14.17585173 -0.04408320  0.34619630  0.65577384  0.24298977 -0.61068871  0.50272414 -0.06241158 -0.03691894  0.32310264
       plow        npro        tgif        lgif        rgif        tdon        agif
 0.24052418  0.13588115  0.06211882 -0.05587554  0.51706515  0.07182475  0.67062446
```





Best Subset selection- using Cp

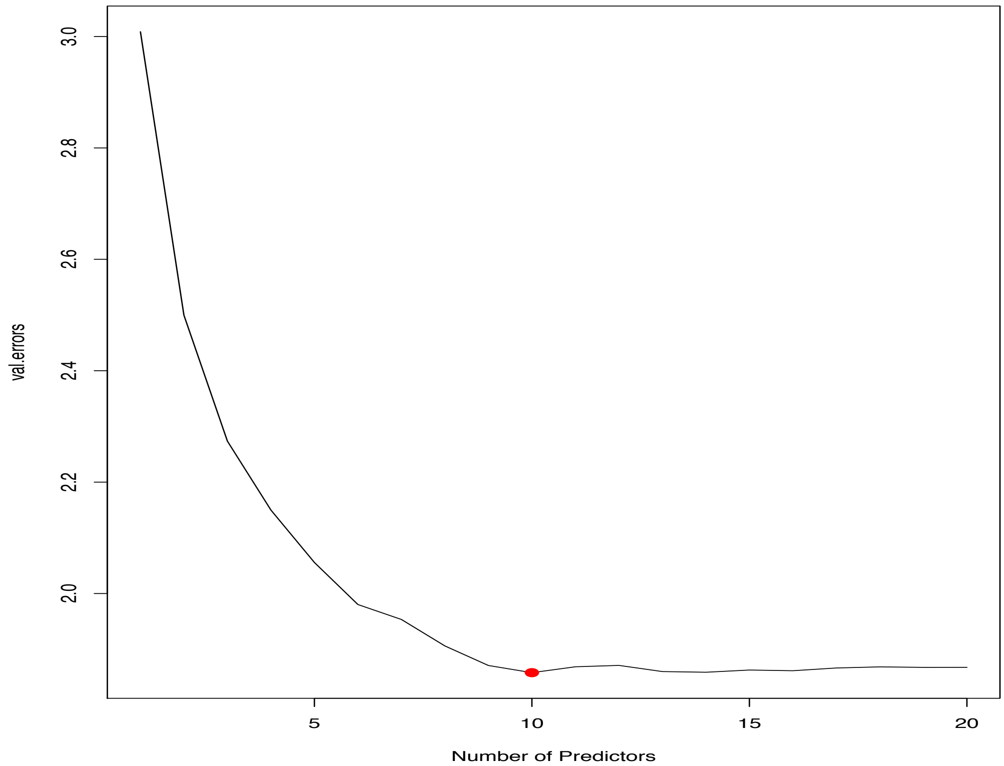
The Cp model suggests the optimal number of variables should be 13, as pictures above. The relevant variables are shown as below.

```
> coef(regfit.full, 13)
(Intercept)        reg2        reg3        reg4        home        chld        hinc        genf        incm        plow 
14.17722363 -0.04565740  0.34558488  0.65452494  0.23976299 -0.60952031  0.50188476 -0.06330083  0.31246592  0.25900395 
       npro        rgif        tdon        agif 
 0.18196966  0.49109408  0.07073946  0.65723853 
```

The mean prediction error of fitting this model with the validation data is 1.8599.

Best Subsets using validation set approach

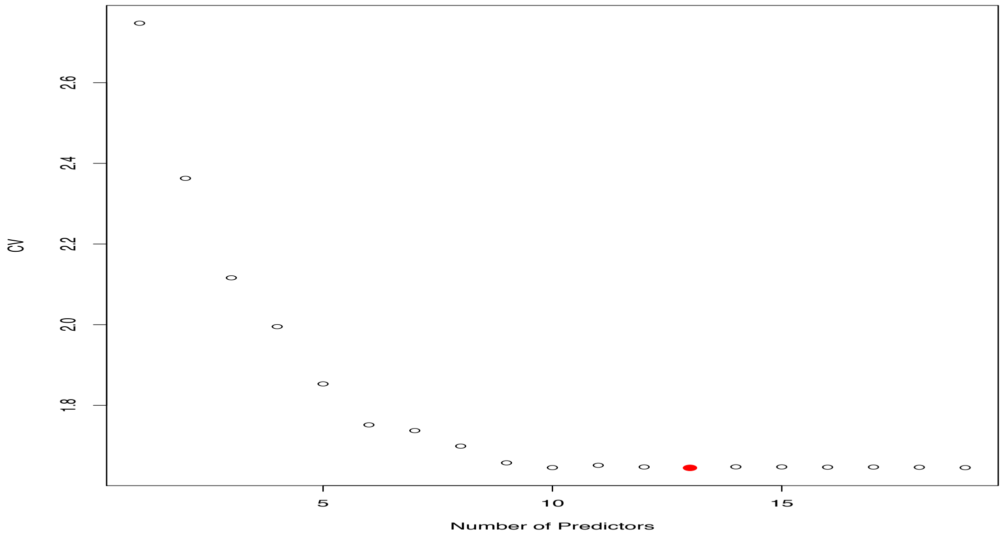
The validation set approach suggests the optimal number of variables should be 10, as pictures below. The relevant variables are shown following. The mean prediction error of fitting this model with the validation data is 1.8579.




```
> coef(regfit.best,10)
(Intercept)        reg3        reg4        home        chld        hinc        incm        plow        npro        rgif        agif 
 14.1480296   0.3581844   0.6686119   0.2511736  -0.6297229   0.5012535   0.3166469   0.2578157   0.1856372   0.4926266   0.6552395 
```

Best Subsets with 5-folds cross-validation

The validation set approach suggests the optimal number of variables should be 13, as pictures below. The relevant variables are shown following.



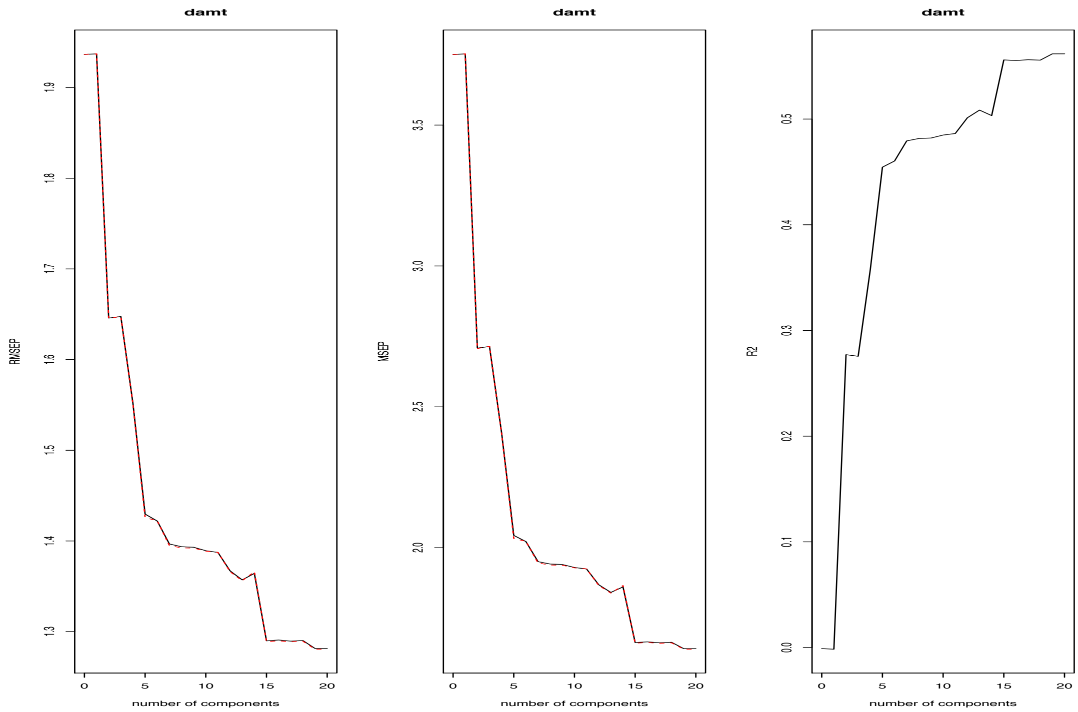

```
> coef(reg.cv.best, 13)
(Intercept)        reg2        reg3        reg4        home        chld        hinc        avhv        incm
 14.1006216  -0.1402094   0.2615188   0.7858994   0.3248440  -0.5566020   0.4917647   0.1527460   0.2642608
       plow        npro        lgif        rgif        agif
  0.4044031   0.1904579   0.1304886   0.4051981   0.6630917
```

The mean prediction error of fitting this model with the validation data is 1.8599.

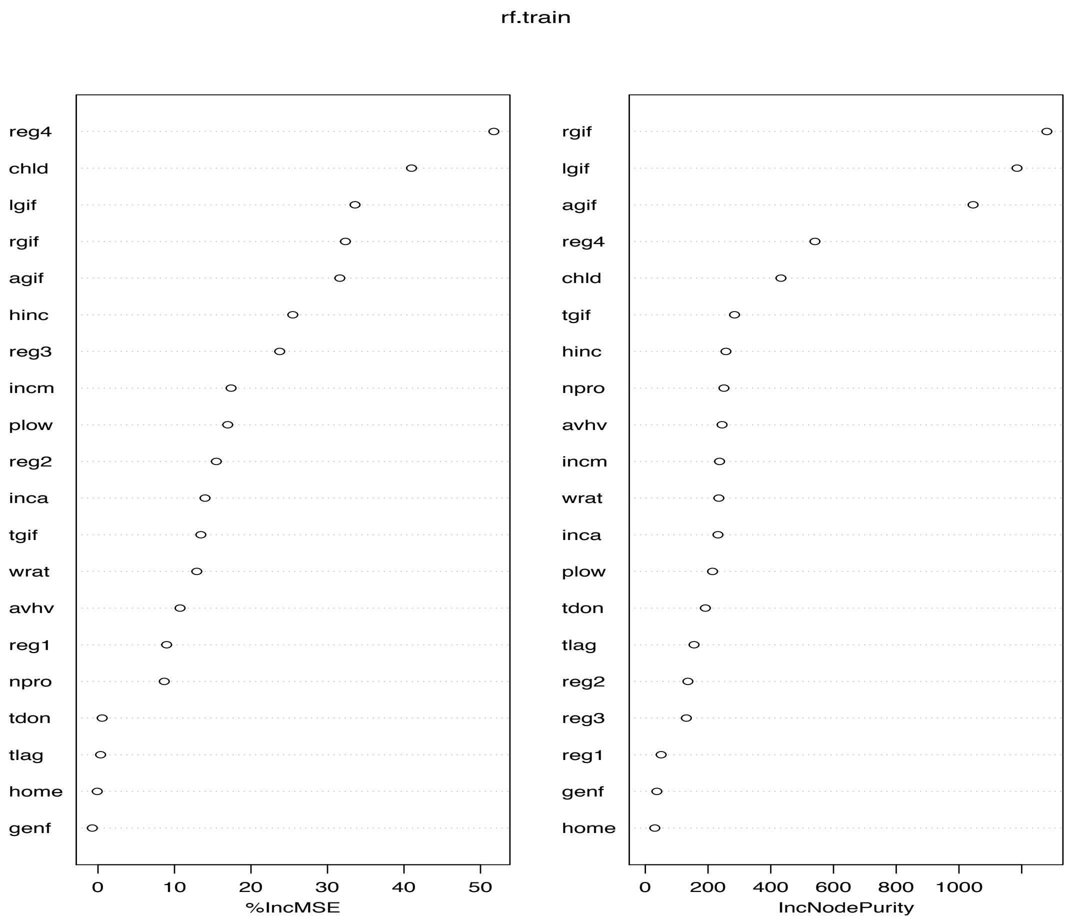
Principle Components Regression and Partial Least Regression

```
> summary(model.pcr)
Data: 	X dimension: 1995 20 
	Y dimension: 1995 1
Fit method: svdpc
Number of components considered: 20

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps
CV           1.937    1.937    1.646    1.648    1.551    1.430    1.422    1.397    1.394    1.393     1.389
adjCV        1.937    1.937    1.645    1.648    1.551    1.426    1.422    1.395    1.392    1.392     1.389
       11 comps  12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
CV        1.387     1.367     1.357     1.364     1.290     1.291     1.290     1.290     1.281     1.281
adjCV     1.387     1.366     1.356     1.366     1.289     1.290     1.289     1.289     1.281     1.281

TRAINING: % variance explained
      1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X    16.08851    27.91    36.73    45.01    51.11    56.80    62.35    67.59    72.71     77.67     82.46
damt  0.02819    28.46    28.54    36.52    46.58    47.08    48.92    49.23    49.36     49.57     49.77
      12 comps  13 comps  14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
X        87.10     90.46     92.70     94.80     96.27     97.61     98.64     99.57    100.00
damt     51.17     51.98     51.98     56.49     56.50     56.58     56.58     57.18     57.22
```

The validation set approach suggests the optimal number of variables should be 20, as picture above. In Pictures above, we can observe that the RMSEP and MSEP is decreasing with number of components. Under such circumstances, Partial Least Regression would only perform worse than PCR. The mean prediction error is 1.8675.



Bagging and Random Forest

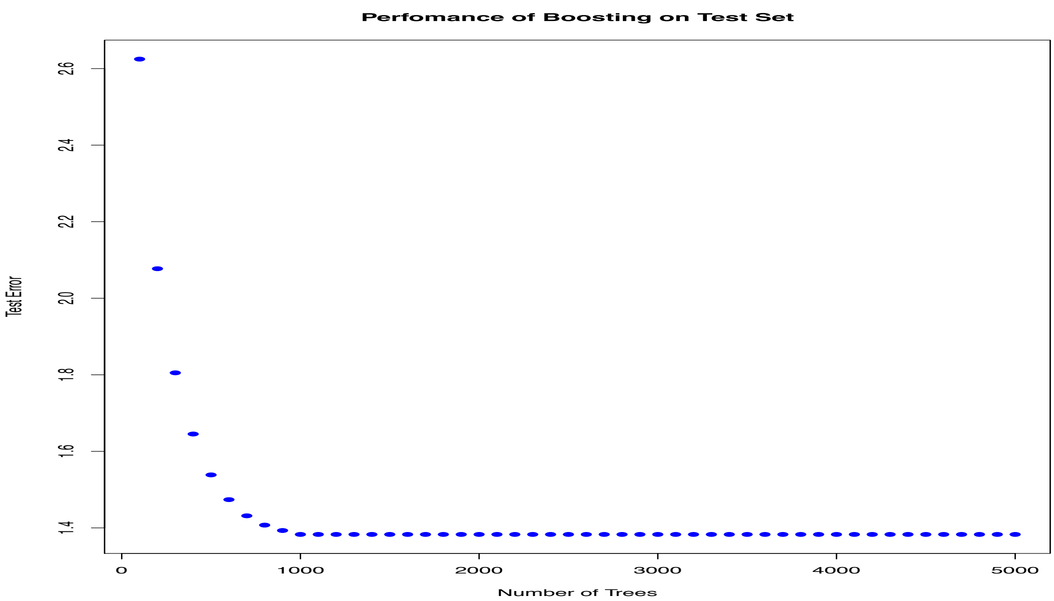


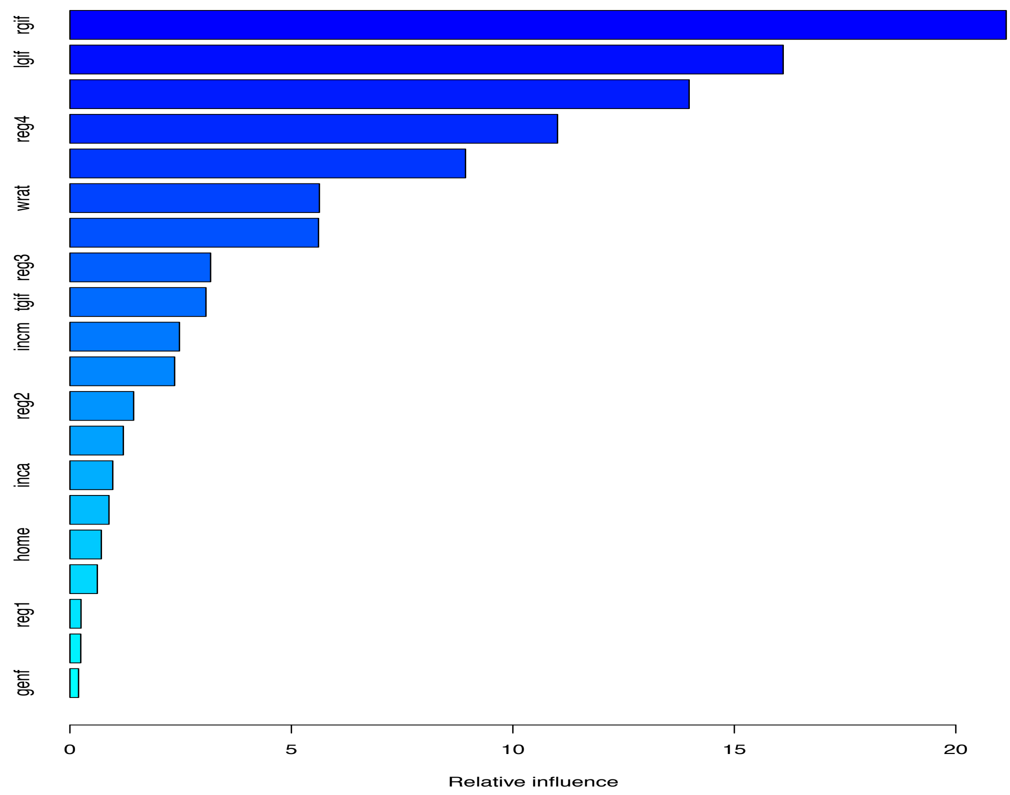
bagging is used to reduce variance, but main idea is to create an aggregate fitted value based of large number of bootstrap samples. However, random forest is used to lower variance among our models. averaging over a large amount of trees helps us reduce the variance. In this sense, random forest is said to have good predictive accuracy and bagging may have highly correlated predictors.

The bagging model starts with the mtry of default number and ntree of 500. The MPE of this model is 1.6721. The random forest model checks with the mtry of 20, which adversely define its performance with percentage of variance explained decreased from 60.6% to 60.35%. Then, we tried the mtry of 6 and 7. Among these trials, 6 mtry has the highest performance. The MPE under this model is 1.6752. The importance of variables shows in the picture above.

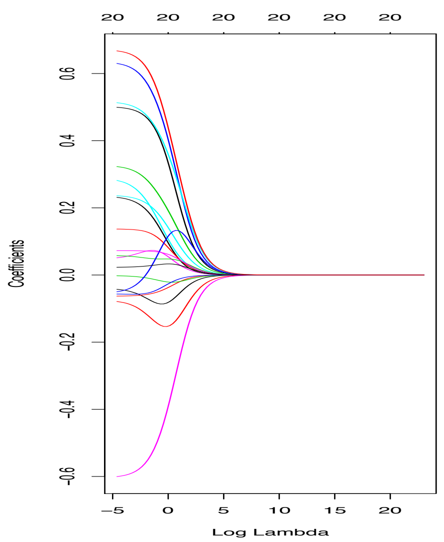
Shrinkage Parameter

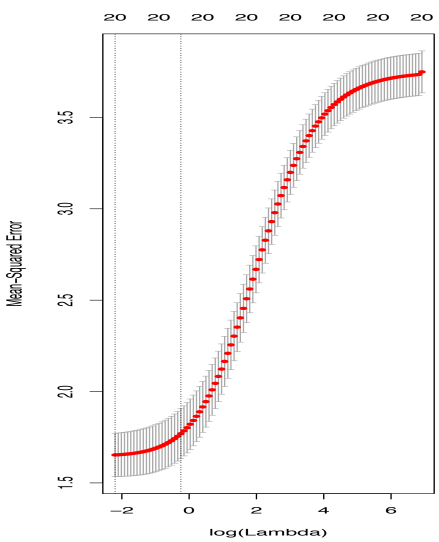
We first tried with 10000 trees and 0.1 shrinkage in order to find the optimal number of trees. As plot shows, 1000 is the most efficient selection and thus most ecological. Based on the 1000 trees, we tried the shrinkage parameter with 0.01 and 0.001. The mean prediction error is smallest with 0.01 as 1.3818. Additionally, we attached the plot of the relative importance among variables.





Ridge Regression

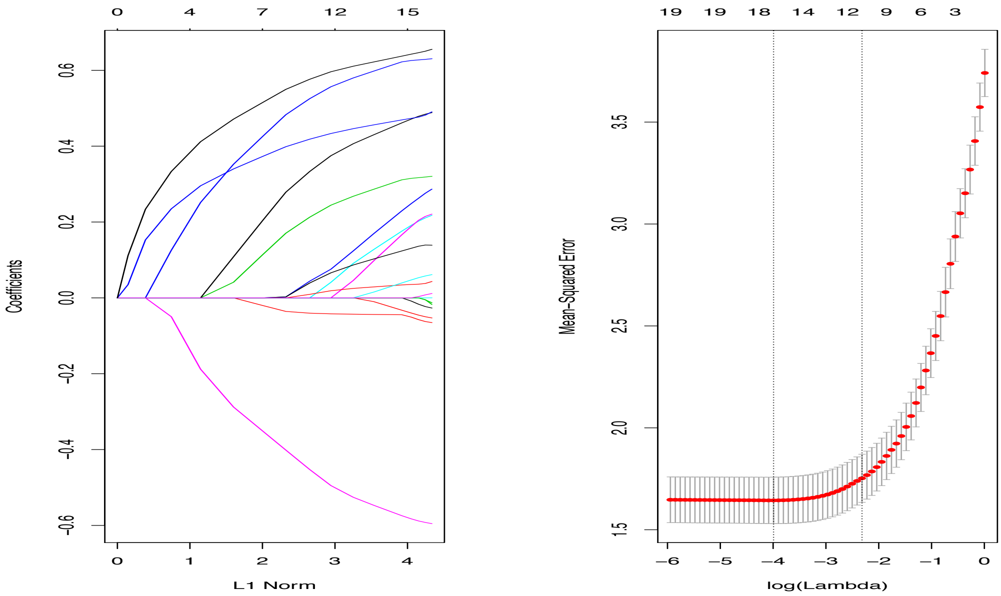




The figures above provide the estimated ridge regression coefficient values and CV errors for altering values of lamda. The best lamda is 0.1107 and the MPE for this model is 1.8732.

Lasso Regression

The figures below provide the estimated lasso regression coefficient values and CV errors for altering values of lamda. The best lambda is 0.0088 and the MPE for this model is 1.8613.



Result

For part 1: By comparing the largest profit of each model, we selected the best model which is Logistic regression with a largest profit of $11640.5.

Model	Logistic	Logistic GAM	LDA	QDA	Decision Tree
Num. of Mailing	1321	1329	1329	1377	1863
Max profit	11640.5	11624.5	11624.5	11224	10687

Then we used test data to fit in the Logistic model and try to predict number of mail for maximize total profit. As we calculated, by sending 349 mails to donors with highest posterior probabilities, we can achieve the largest profit.

For part 2: the shrinkage parameter with 0.01 has the least MPE. Thus we use this model to predict the DMAT response in test data. The result present as follow:

```
> yhat.test[1:10]
 [1] 15.84203 14.40135 16.12137 10.62788 15.09883 16.01570 16.01481 14.72022 11.70643 16.53567
```