PREDICT 422 Course Project

Sri Seshadri

3/13/2018

Contents

1. Introduction **2**

2. About the data **2**

3. Exploratory Data Analysis (EDA) **2**
3.1 Data Transformation. 3

4. Feature selection **3**
4.1 Correlations among predictors 3

5. Classification Modeling **3**
5.1 Random forest 3
5.2 Logistic Regression 6
5.3 Linear Discriminant Analysis (LDA) 11
5.4 Quadratic Discriminant Analysis (QDA) 12
5.5 Partial Least Squares Discriminant Analysis (PLSDA) 13
5.6 Support Vector Machine 14

1. Introduction

A charitable organization is wanting to maximize the donations as a result of their direct mailing campaign. Historically they have had 10% response rate for their mailing campaigns, with an average of `$14.50` collected as donation. However it costs `$2.00` to produce and send out a mail. This results in an expected donation of 14.50 x 10% - 2 = -$0.55. The charitable organization is wanting to use to build predicitive models on the data that was collected recently. We will be helping the organization with building predictive models to classify who are the likely donors. Also we will model the likely donations from the predicted donors.

2. About the data

A weighted sampling method over representing the donors such that an equal of number non-donors and donors are represented was used. The data is split into two groups a) Training set; to train predictive models b) Validation set - to validate the predictive models. The model that maximizes the profit is used as the criteria for model selection for predicting potential donors. Another model is trained on to predict donation amount from the donors. The mean prediction error (MAE/MSE) is used as a criteria for model selection. The nomencalture of the features in the data are provided in Apendix A.

Below is the breakdown of the samples as training, validation and test (final prediction made on this set of the data):

Table 1: Samples breakdown

part	Samples
test	2007
train	3984
valid	2018

3. Exploratory Data Analysis (EDA)

Figure 1 shows the correlation matrix ofthe variables. There appears to be association between the number of children in the household and donors. As expected, the median and mean household incomes are positively correlated. The incomes are positively corrleated with home values and negatively correlated with percent categorized as kow income. The lifetime number of promotions received is positively correlated with the life time gifts. Average gifts are correlated with largest and recent gifts as expected. There are weak correlations between donor (donor amount) and all predictors except number of children.

The below tables show the percent donors by each of the levels in number of children, wealth rating, household income and home owner.

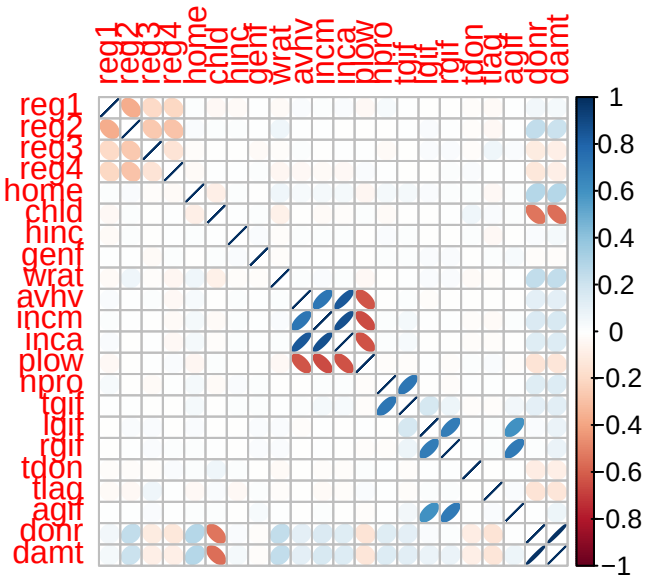


Figure 1: Correlation between variables

	No_Donor	Yes_Donor	%
0	194	1201	86.1
1	203	201	49.8
2	770	381	33.1
3	494	162	24.7
4	237	44	15.7
5	91	6	6.2

3.1 Data Transformation.

The features lifetime gifts to date, largest gift to date, recent gift and average home value are skewed to the right. We'll use log transformation to minimize the effect of extreme values in the model. Further transformation if needed would be made and noted.

4. Feature selection

4.1 Correlations among predictors

Refering figure 2, it c an be noted that there is over 80% correlation between average income and median income and log home values. Therefore average income is retained and remove home value and median income as predictors. Also the log of largest gift has over 80% correlation with recent gift amount and average gift amount. Hence largest gift is retained as predictors and the average and recent gifts are removed. Similarly the number of promotions received has 87% correlation with log of lifetime gifts to date. The number of promotions is retained as predictor.

The following are the predictors that are retained for model fitting.

Table 3: List of retained predictors

reg1	chld	plow
reg2	hinc	npro
reg3	genf	lgif
reg4	wrat	tdon
home	inca	tlag

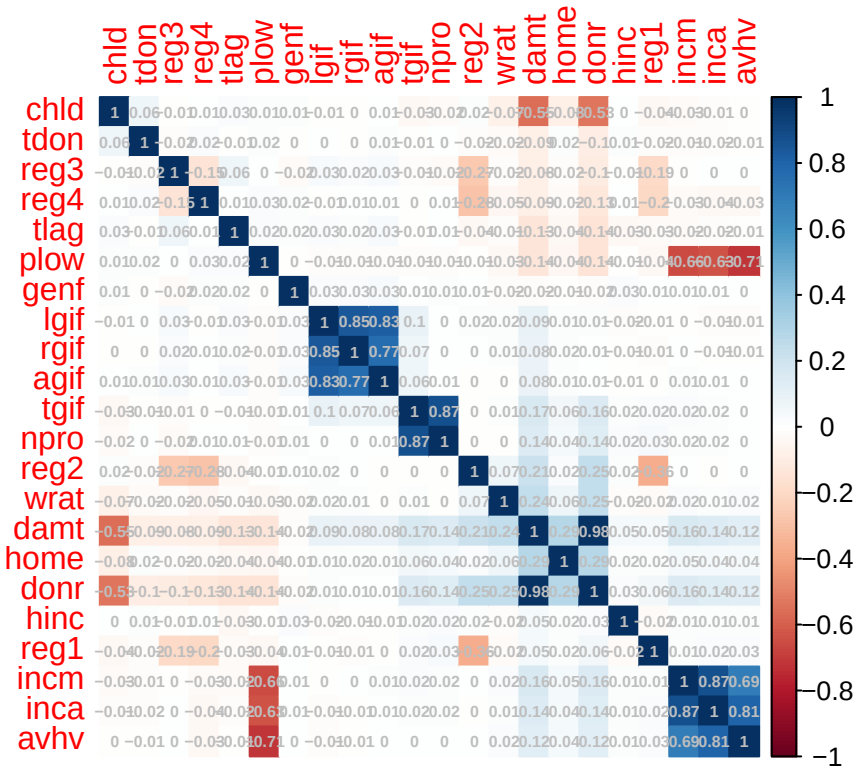


Figure 2: Correlation plot of transformed training data

It is seen from the variable importance plot in figure 3, that categorical variables like number of children, household income, wealth rating, home ownership and region 2 came up to be very important. The classification trees could be biased in choosing categorical variables for split. Other techniques would be explored in the following sections.

5.1.1 Random forest model assessment.

Figure 4 shows that the random forest fit on the training data is fit perfectly with 100% Area Under the Curve (AUC). There is slight decrease in performance on the validation set; with 96% AUC. The model shows that maximum profit of `$11,791.50` can obtained by mailing 61% of the people. the accuracy of the model is shown below.

Table 4: Mails for campaign & profitability

Mails	Profit	Model
1231	11791.5	RandomForest

```
##            Reference
## Prediction   0   1
##          0 771  17
##          1 248 982
```

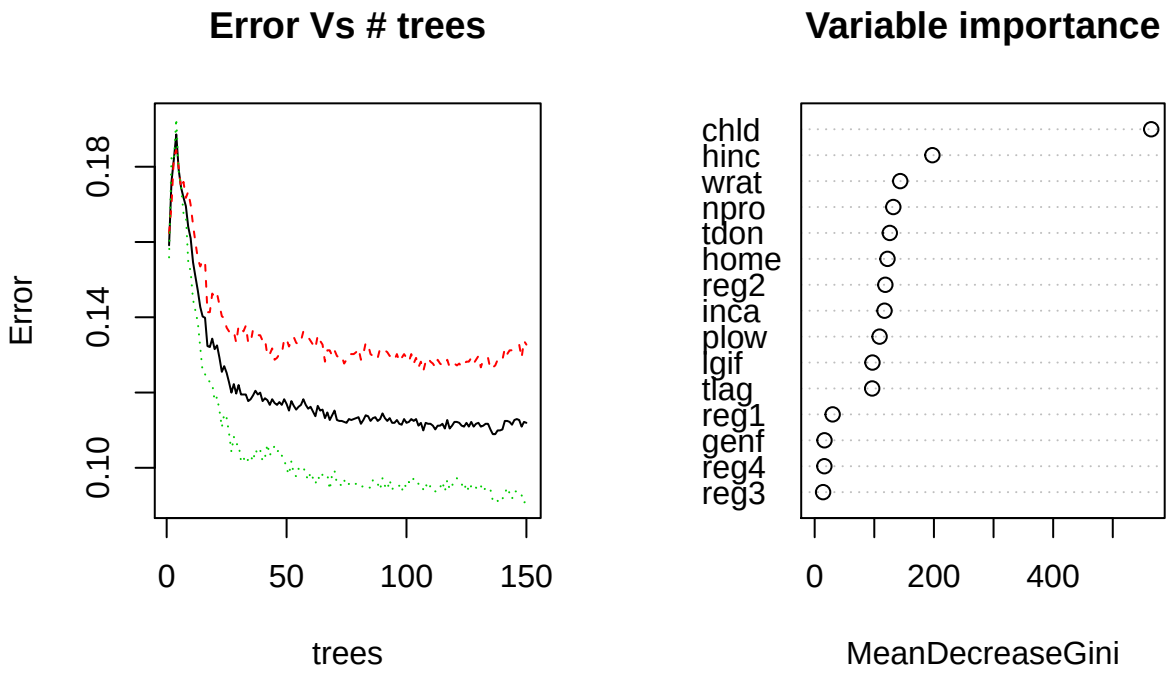


Figure 3: Variable importance plot

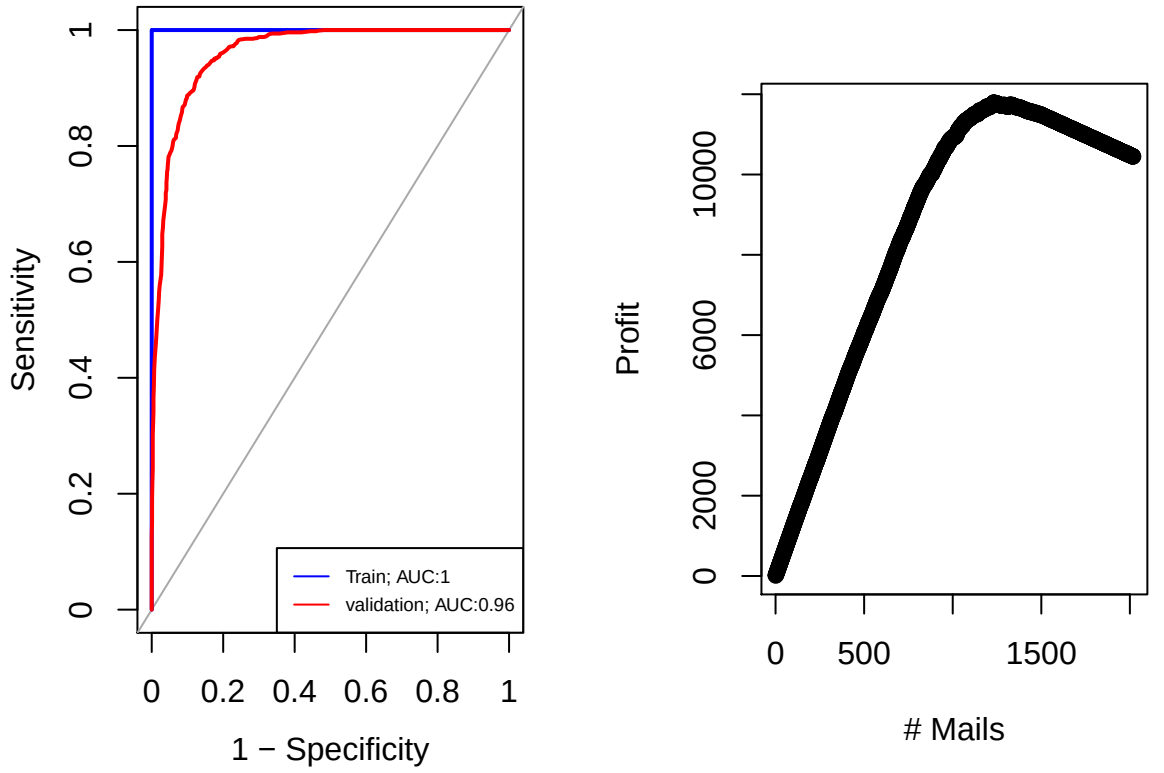


Figure 4: Left: ROC curves of random forest fit comparing the training and validation data. Right: Cumulative profit vs mails sent to potential donors based on Random forest classification

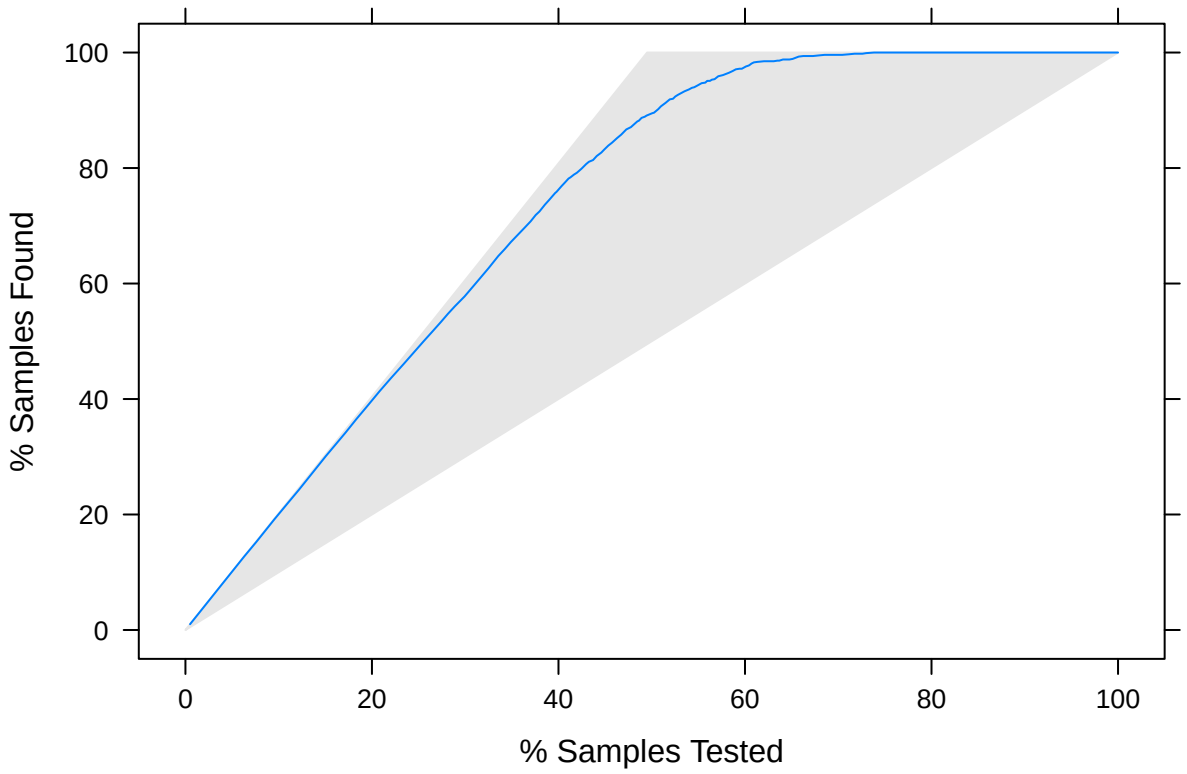


Figure 5: Lift chart of random forest model on validation data

5.2 Logistic Regression

In this section, two logistic regression models are attempted. - A full model, using all features that were deemed useful after removing correlated variables. - A reduced model using the features that were important in the variable importance plot.

It is seen that the validation ROC tracks the training ROC. The models does not provide a huge improvement over the random forest.

```
##
## Call:
## glm(formula = donr ~ . - damt, family = binomial, data = trainTransformed[,
##     c(Predictors, Response)])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.14634  -0.45682   0.05849   0.52330   2.84652
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.769959   0.479730  -9.943  < 2e-16 ***
## reg11        1.356178   0.152279   8.906  < 2e-16 ***
## reg21        2.545993   0.147987  17.204  < 2e-16 ***
## reg31        0.074432   0.171987   0.433    0.665
## reg41       -0.042279   0.170127  -0.249    0.804
## home1        3.463752   0.215297  16.088  < 2e-16 ***
## chld        -1.380894   0.047548 -29.042  < 2e-16 ***
```

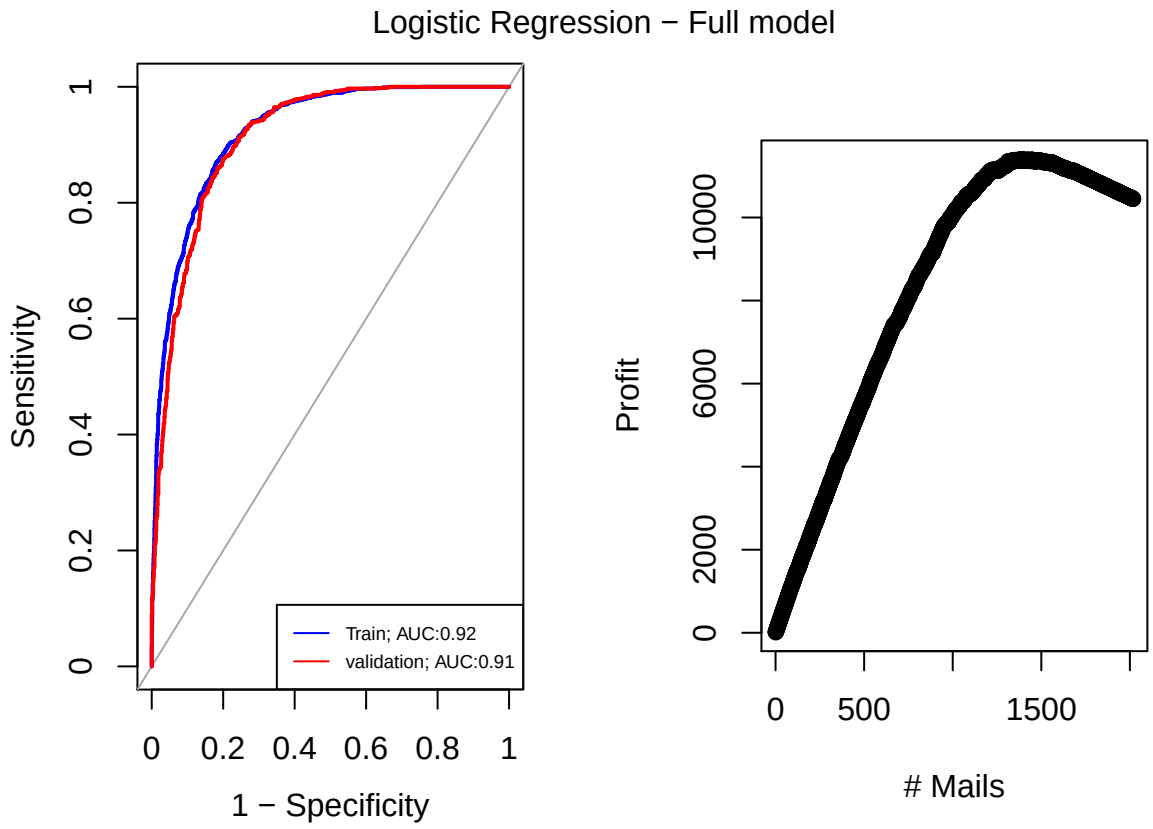


Figure 6: Left: ROC Curve of logistic regression Right: Profitability curve vs number of mails

```
## hinc         0.058932   0.037018   1.592    0.111
## genf1       -0.072181   0.098478  -0.733    0.464
## wrat         0.353208   0.024259  14.560  < 2e-16 ***
## inca         0.012920   0.002638   4.898 9.69e-07 ***
## plow        -0.019853   0.004968  -3.996 6.43e-05 ***
## npro         0.015636   0.001627   9.613  < 2e-16 ***
## lgif        -0.008980   0.062597  -0.143    0.886
## tdon        -0.040874   0.009011  -4.536 5.74e-06 ***
## tlag        -0.128993   0.014334  -8.999  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5523.0  on 3983  degrees of freedom
## Residual deviance: 2780.1  on 3968  degrees of freedom
## AIC: 2812.1
##
## Number of Fisher Scoring iterations: 6
```

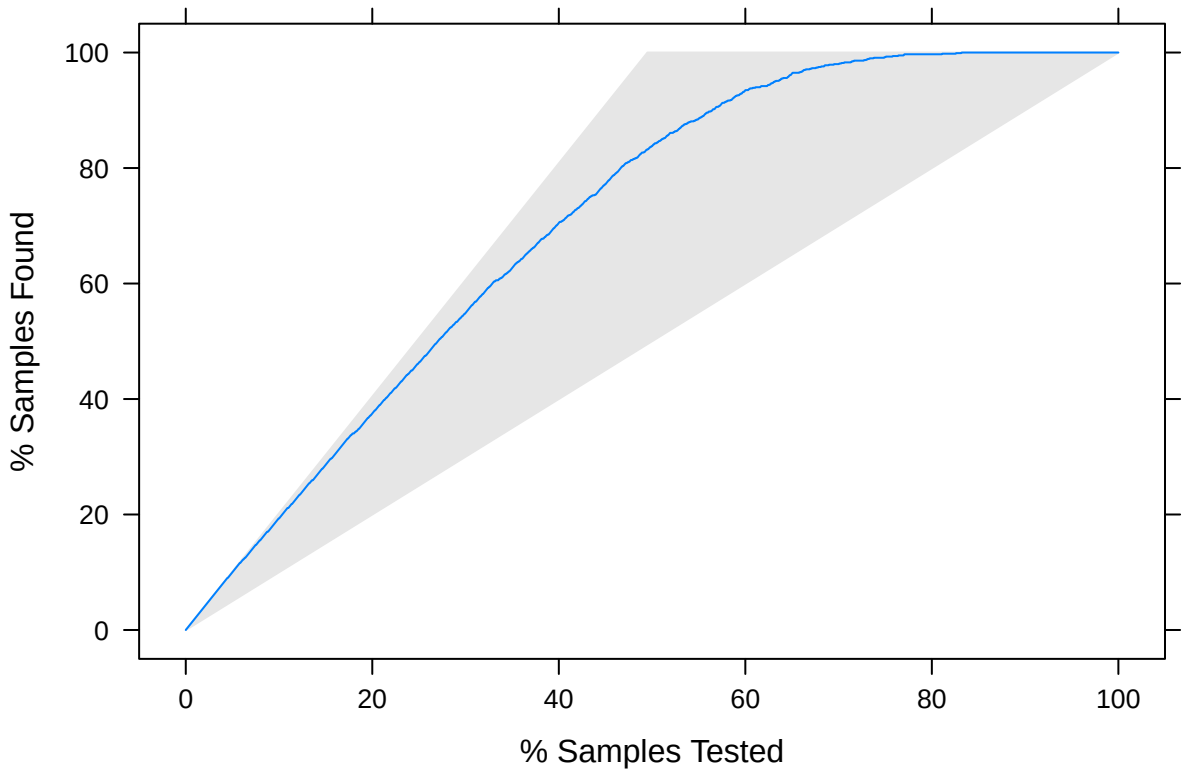


Table 5: Mails for campaign & profitability

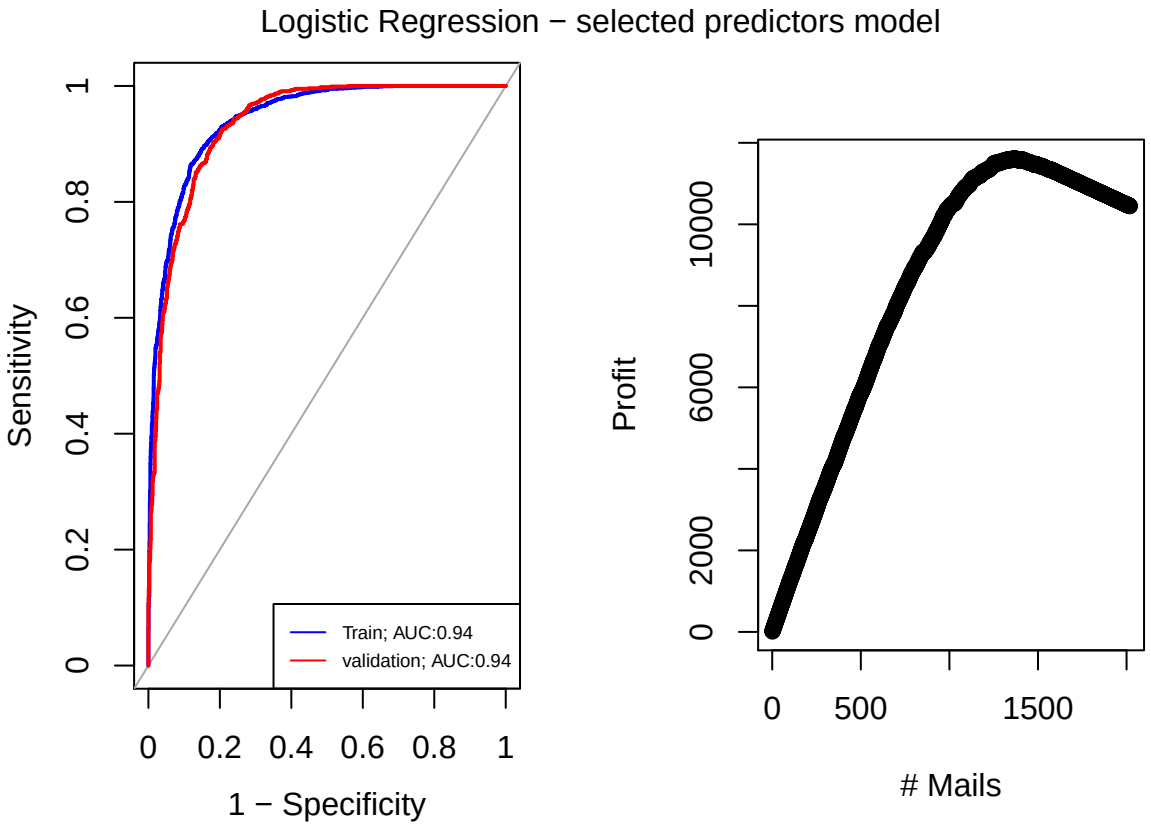
Mails	Profit	Model
1231	11791.5	RandomForest
1383	11400.5	Logistic

```
##            Reference
## Prediction   0   1
##          0 613  22
##          1 406 977
```

In the first logistic model, the residuals when plotted against house hold income category had a curve. Figure not shown for brevity of the report. So quadratic term in "hinc" feature was attempted.

```
##
## Call:
## glm(formula = donr ~ chld + wrat + I(hinc^2) + I(hinc) + npro +
##     tdon + home + reg2 + inca - damt, family = binomial, data = trainTransformed[,
##     c(Predictors, Response)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1819  -0.3464   0.0233   0.4281   2.9816
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -13.447832   0.626745 -21.457  < 2e-16 ***
## chld         -1.544023   0.054429 -28.368  < 2e-16 ***
## wrat          0.384739   0.026593  14.468  < 2e-16 ***
```

```
## I(hinc^2)    -0.504251   0.024881 -20.266  < 2e-16 ***
## I(hinc)       4.017005   0.201495  19.936  < 2e-16 ***
## npro          0.017237   0.001766   9.760  < 2e-16 ***
## tdon         -0.050733   0.010036  -5.055  4.3e-07 ***
## home1         3.913253   0.239629  16.330  < 2e-16 ***
## reg21         2.447509   0.126924  19.283  < 2e-16 ***
## inca          0.023395   0.002236  10.463  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5523.0  on 3983  degrees of freedom
## Residual deviance: 2414.8  on 3974  degrees of freedom
## AIC: 2434.8
##
## Number of Fisher Scoring iterations: 6
```



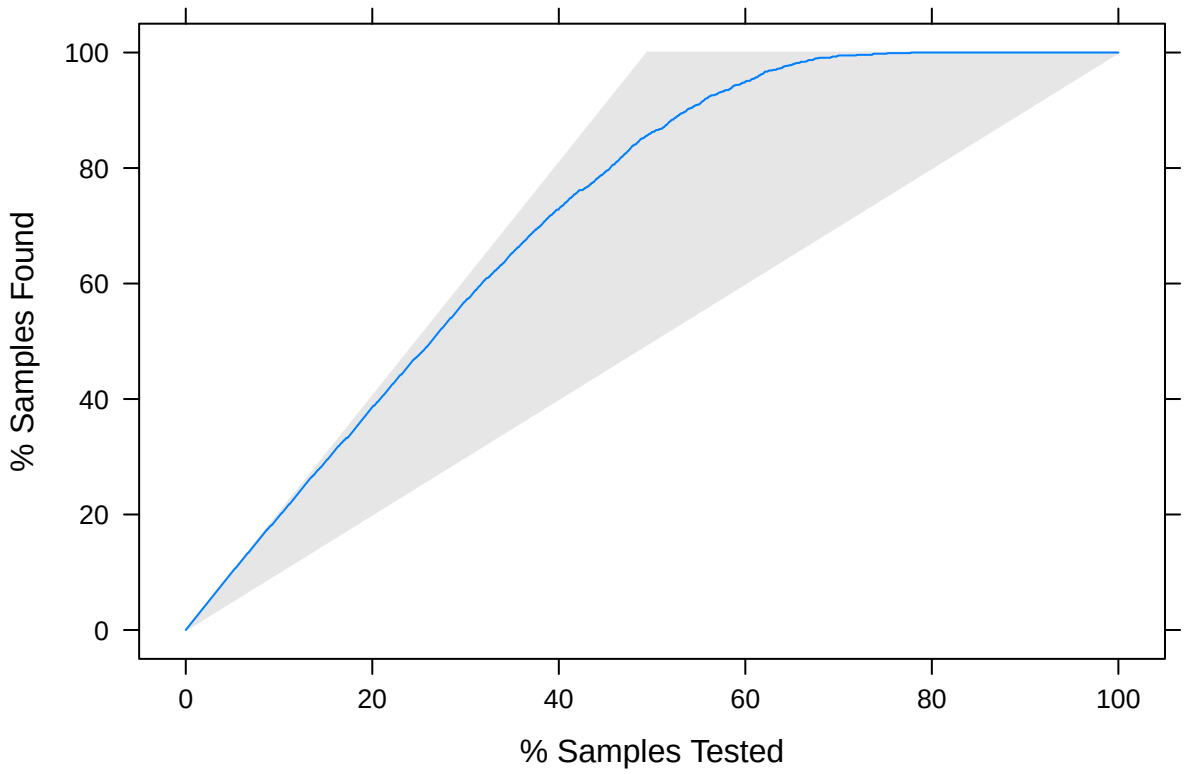
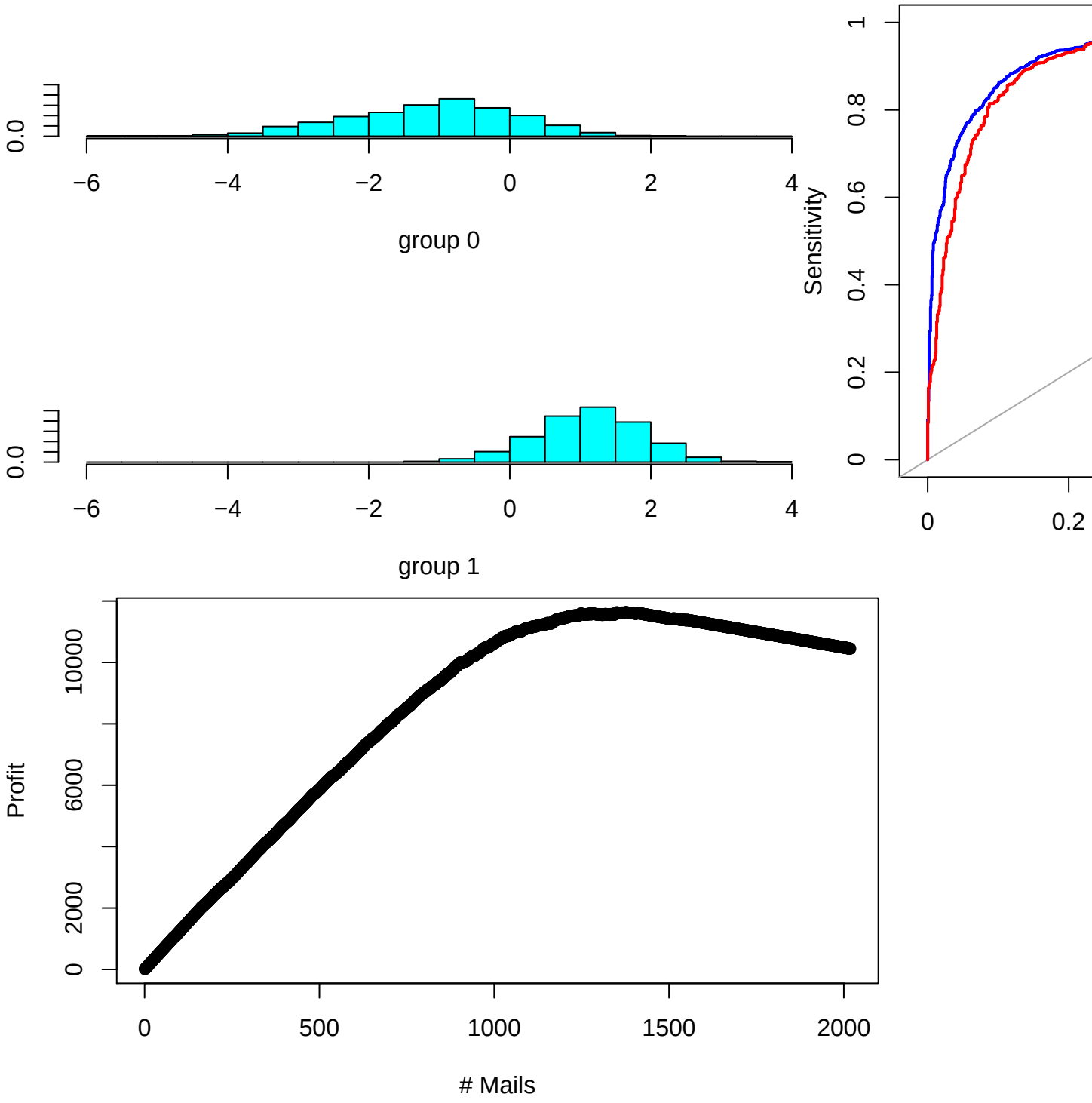


Table 6: Mails for campaign & profitability

Mails	Profit	Model
1231	11791.5	RandomForest
1383	11400.5	Logistic
1370	11615.0	Logistic2

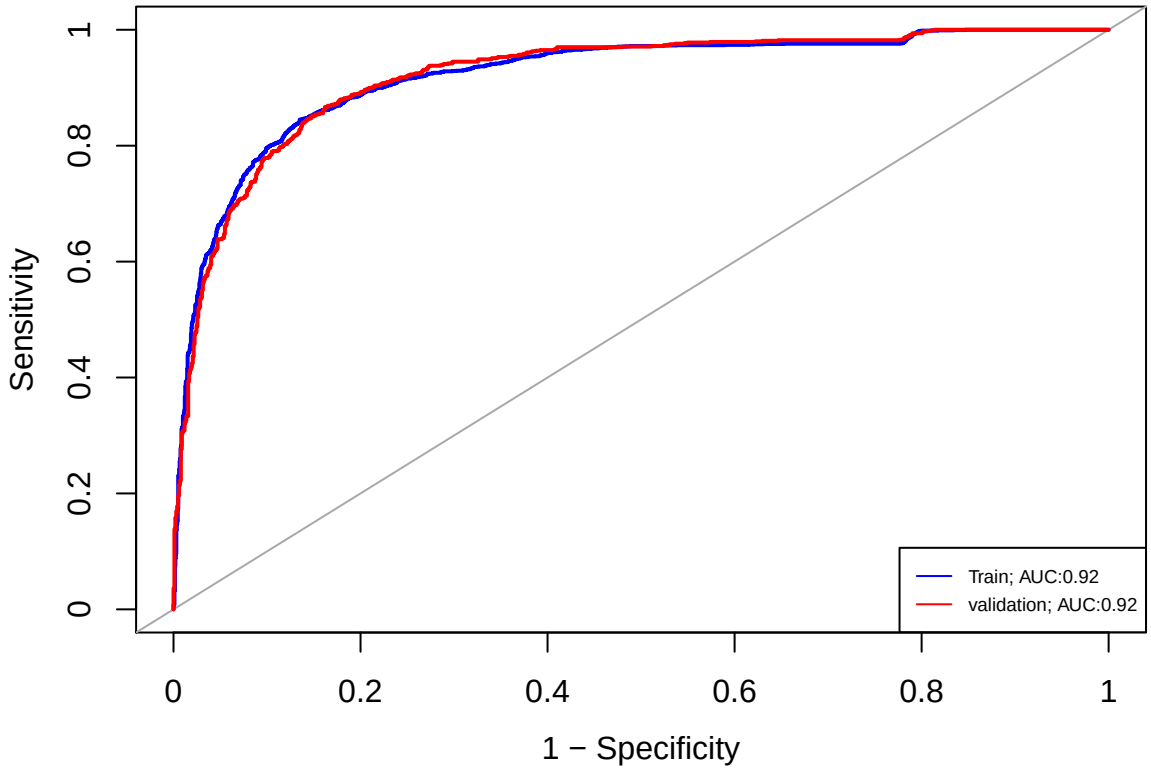
```
##            Reference
## Prediction   0   1
##          0 639   9
##          1 380 990
```

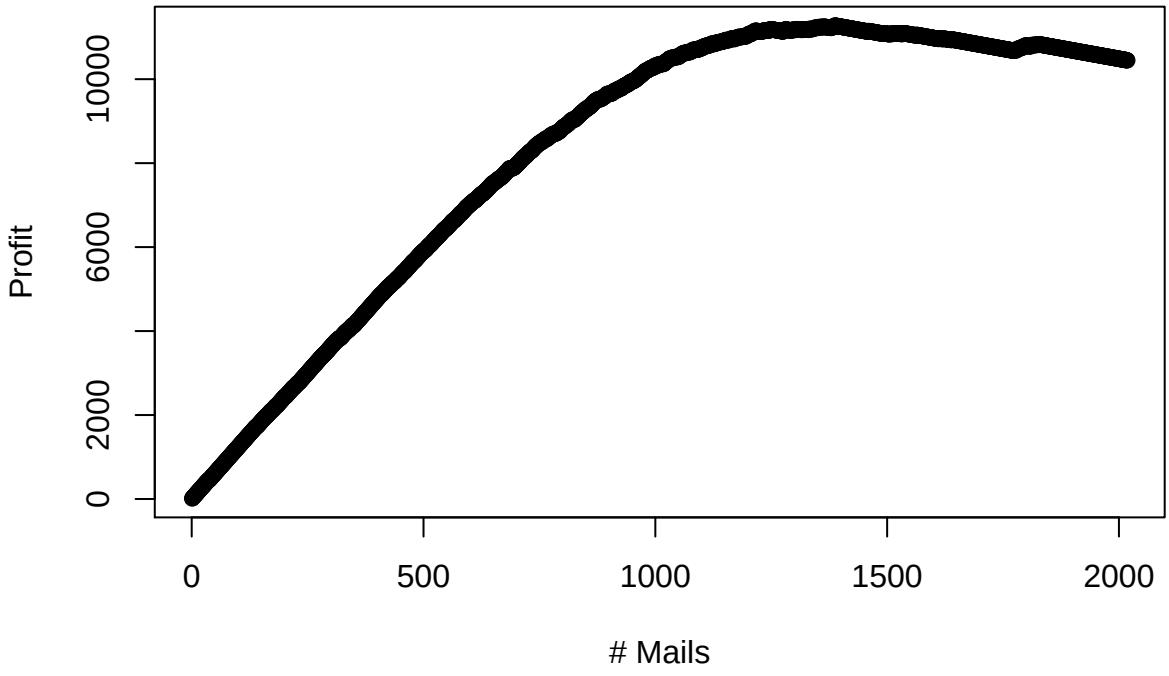
5.3 Linear Discriminant Analysis (LDA)




```
##   Mails Profit Model
## 1  1376  11632   lda
```

5.4 Quadratic Discriminant Analysis (QDA)



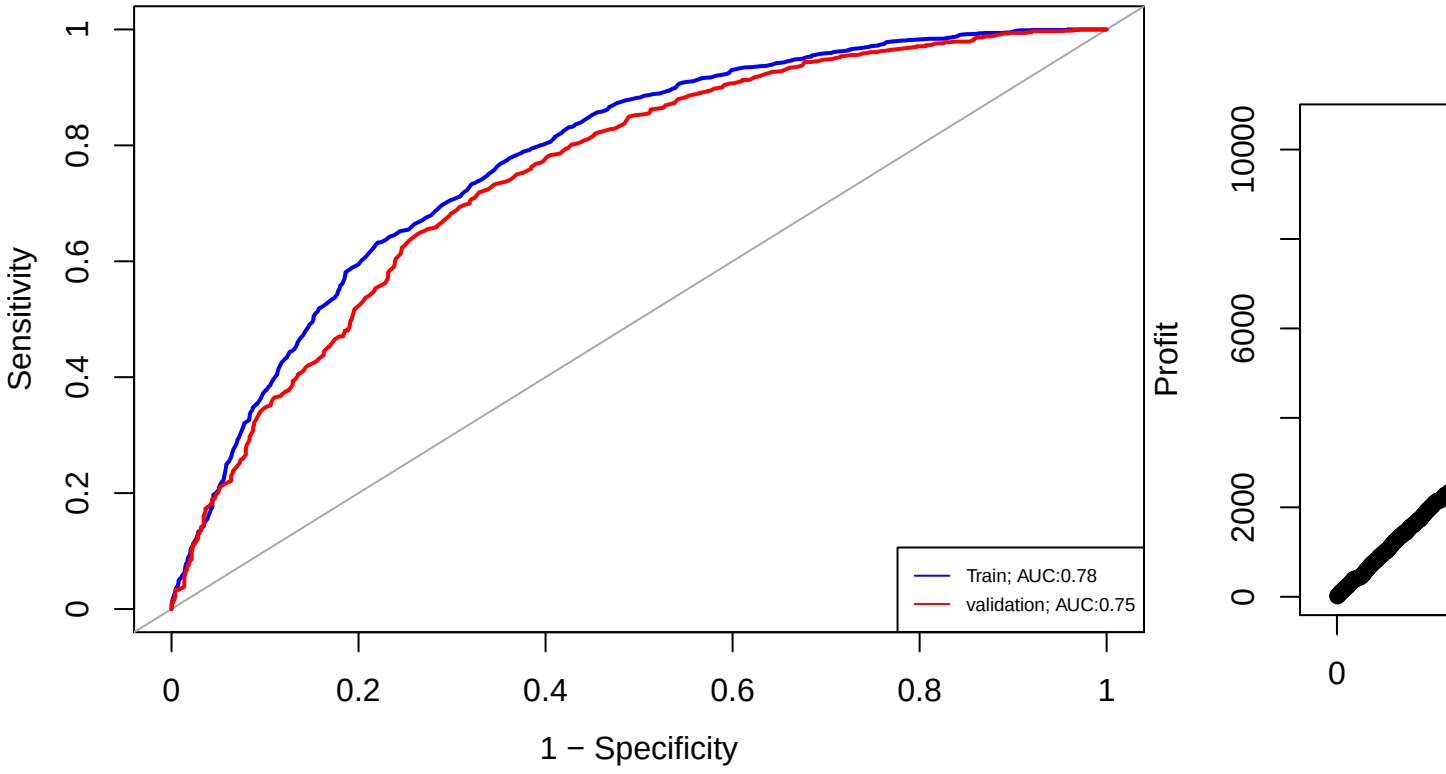



```
##   Mails  Profit Model
## 1  1387 11276.5   qda
```

```
## Confusion Matrix and Statistics
##
##           Reference
```

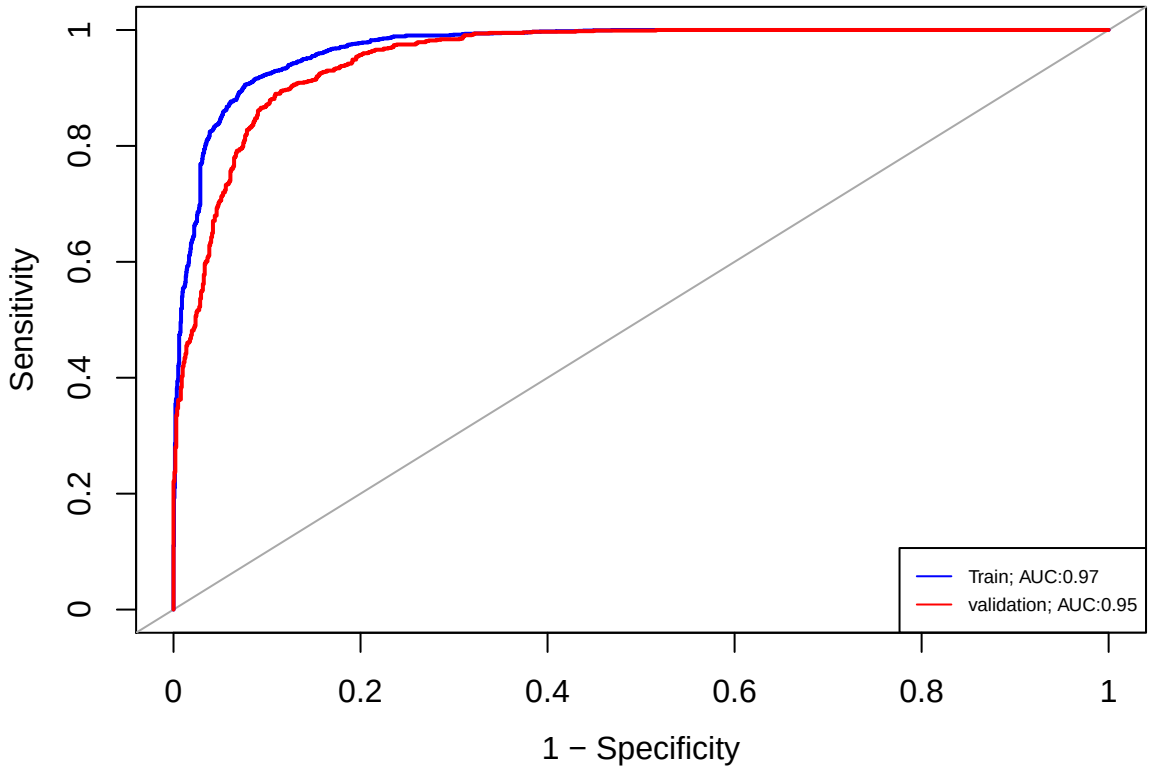
```
## Prediction   0   1
##          0 601  30
##          1 418 969
##
##                Accuracy : 0.778
##                  95% CI : (0.7592, 0.796)
##     No Information Rate : 0.505
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5576
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9700
##             Specificity : 0.5898
##          Pos Pred Value : 0.6986
##          Neg Pred Value : 0.9525
##              Prevalence : 0.4950
##          Detection Rate : 0.4802
##    Detection Prevalence : 0.6873
##       Balanced Accuracy : 0.7799
##
##        'Positive' Class : 1
##
```

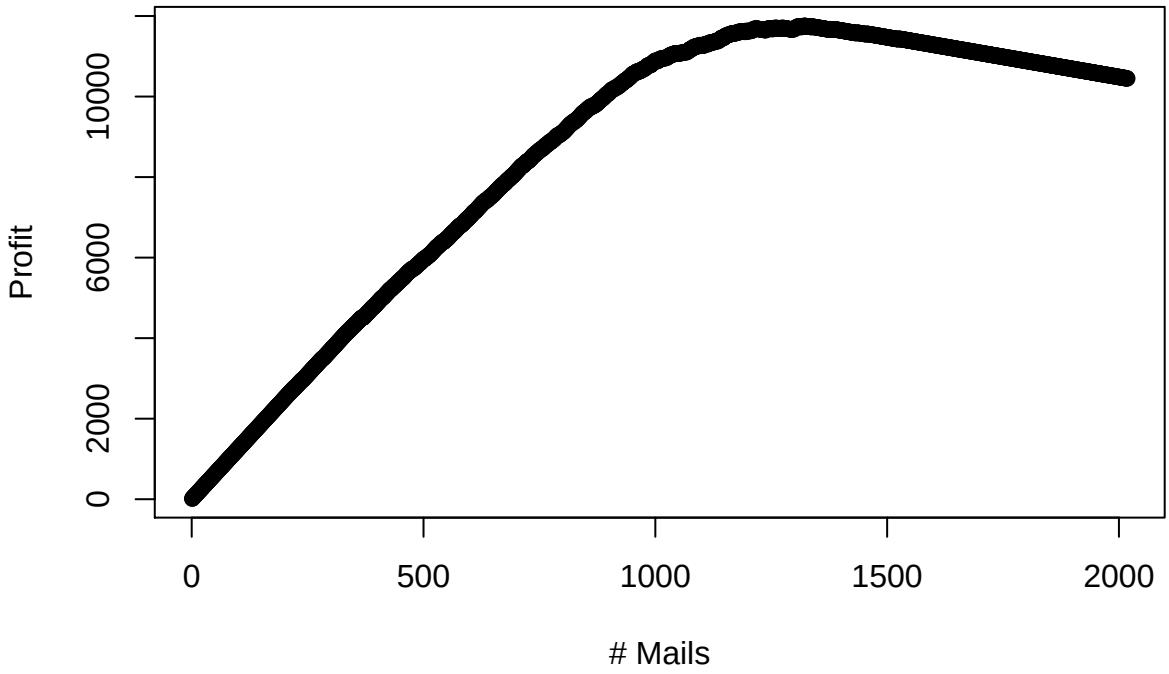
5.5 Partial Least Squares Discriminant Analysis (PLSDA)




```
##   Mails  Profit  Model
## 1  1906 10586.5 plsqda
```

5.6 Support Vector Machine






```
##   Mails  Profit Model
## 1  1321 11756.5   SVM
```