Predict-422
May 15, 2017

FINAL COURSE PROJECT

Classification and predictive models for Charity data

Nitin Gaonkar

INTRODUCTION:

There has been lot of interest lately among the charitable organization and nonprofits about statistical modelling, the right models could benefit the organization in identifying the prospects who are likely to contribute to the organization, Lot of fundraiser have successfully used the statistical models to improve their fundraising results. In this project, we are building various machine learning models for a charitable organization to improve their cost effectiveness of direct marketing campaigns to the previous donors.

The project is divided into two parts, the first part is to build the classification models using the data from the most recent campaign that can effectively captures likely donors so that the expected net profit is maximized. The second part is to develop a predictive model to predict expected gift amounts from donors. To achieve these objectives of the project we would be building various classifications models and predictive models on the training data and validate it with the validation set to come up with the most effective models which would improve the cost effectiveness and predict the expected gift amounts.

ANALYSIS:

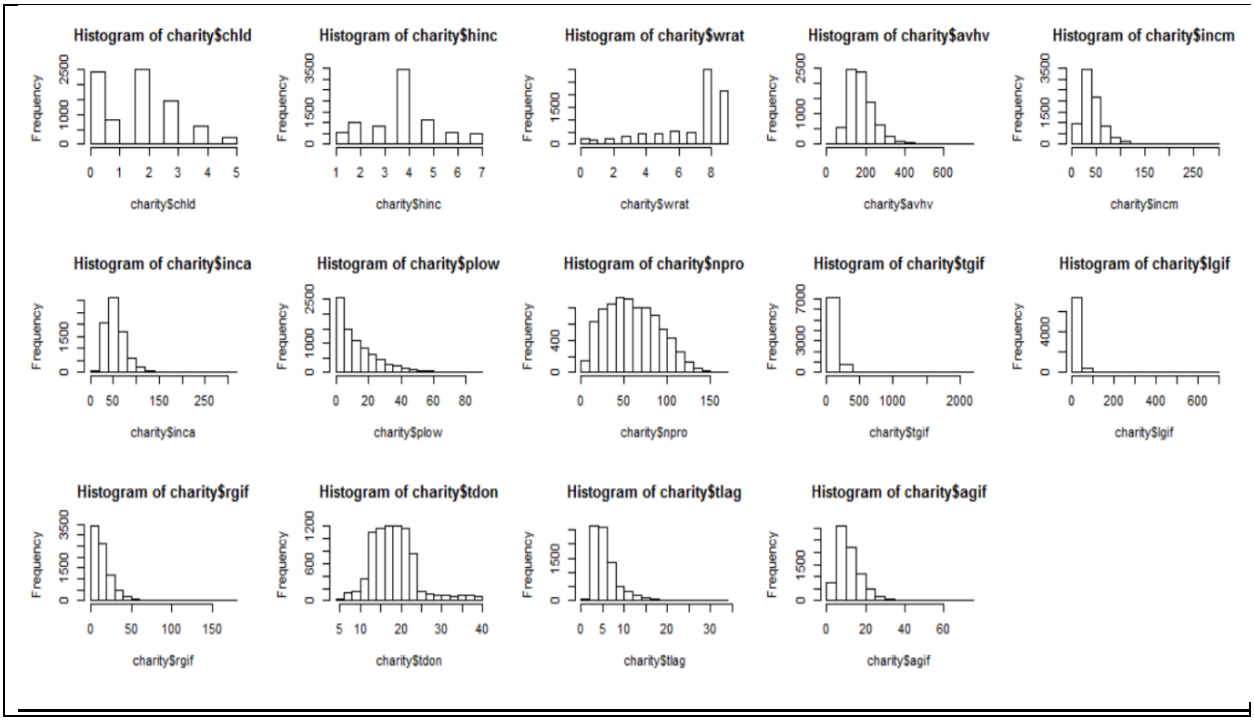
The entire data set has 8009 observations and for the analysis the data is spilt into training set comprising of 3984 observations, a validation set with 2018 observations and test dataset with 2007 observations. Weighted sampling has been used, over-representing the responders so that the training and validation samples have approximately equal numbers of donors and non-donors. The response rate in the test sample has the more typical 10% response rate. Below are the variables used for the analysis(Table1). DAMT and DONR are the dependent variables and the below shown variables are used as predictors

Fig (1) shows the histogram of all the predictor variables and shows the visualization of the continuous variables, Also the values were standardized in the training data such that each predictor variables has a mean of 0 and a standard deviation of 1, Missing data evaluation was done and found that there was no data missing in the data set.

Table.1

Vars.	Description	Vars.	Description
ID	Identification number	PLOW	% categorized as "low income" in potential donor's neighborhood
REG	5 regions indicator variables respectively called REG1, REG2, REG3 and REG4	NPRO	Lifetime number of promotions received to date
HOME	(1 = homeowner, 0 = not a homeowner	TGIF	Dollar amount of lifetime gifts to date
CHLD	Number of children	LGIF	Dollar amount of largest gift to date
HINC	Household income (7 categories)	RGIF	Dollar amount of most recent gift
GENF	Gender (0 = Male, 1 = Female)	TDON	Number of months since last donation
WRAT	Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative within each state. The segments are denoted 0-9, with 9 being the highest wealth group and 0 being the lowest	TLAG	Number of months between first and second gift
AVHV	Average Home Value in potential donor's neighborhood in $ thousands	AGIF	Average dollar amount of gifts to date
INCM	Median Family Income in potential donor's neighborhood in $ thousands	DONR	Classification Response Variable (1=Donor, 0 = Non-donor)
INCA	Average Family Income in potential donor's neighborhood in $ thousands	DAMT	Prediction Response Variable (Donation amount in $)

Fig (1)



MODELS CLASSIFICATION AND PREDICTION:

I have built the models for classification and prediction, I have used LDA, k nearest neighbors(KNN), decision tress, bagging, boosting, logistic regression SVM etc. Similarly, for prediction models I have used linear regression, decision trees, random forest, gradient boosting etc.
For the classification models, the models were compared based on error rates and profits, for prediction models the mean prediction error was chosen to decide on the models. Once I decided on the models and finalized the classification and prediction model, these models were applied on the test data and the results were saved in a csv file.

INDIVIDUAL MODELS AND RESULTS:

Classification models:

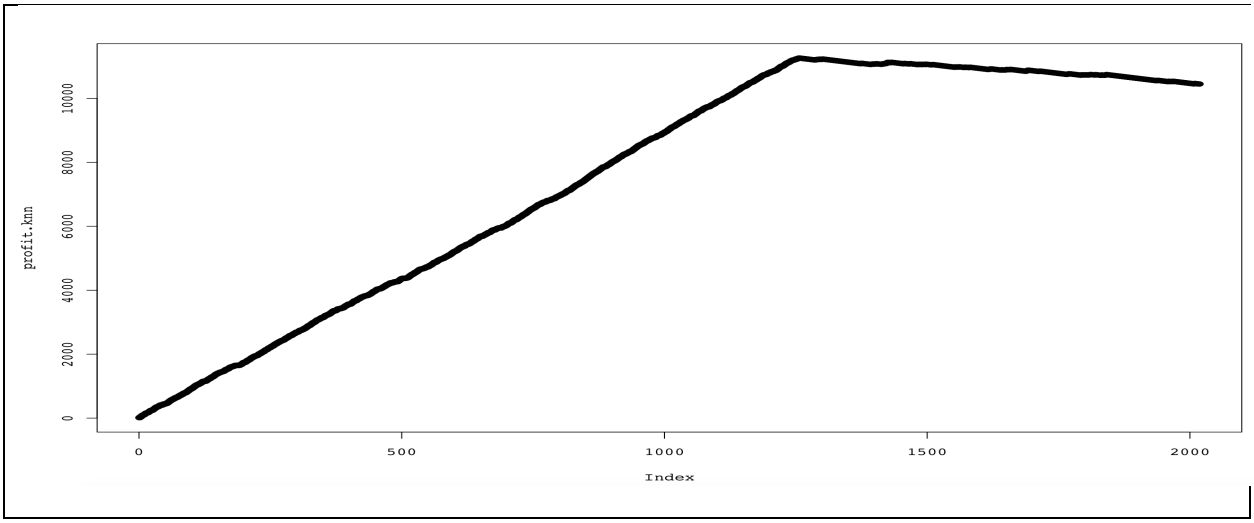
K-Nearest Neighbor(KNN):

The KNN classifier first identifies the k points in the training data that are closest to x0, then estimates the conditional probability for class j as the fraction of points in observations, whose response values equal j, finally applies the Bayes classifier to the test observation x0 to the class with the largest probability, I used KNN classifier for building my first classification model, I started off by setting the value of k to 63 by thumb rule I obtained the value by (sqrt(3984)). Then I ran the KNN model for series of k values and for K=15, I got the 82 % of accuracy from the model and the max profit was 11263 with 1256 emails. In the below plot we can see that how profits changes based on the no of emails. Below are the few results for different values of K. for k=63 we are getting the highest profits, but the classification accuracy is only 80 %

Table.2

k	profits	mails	Accuracy
63	11304.5	1344	79%
10	11001	1242	81%
20	11194	1276	81%
5	10993	1224	81%
15	11263	1256	82%

Plot1(knn- profit/mails, k=15)



K FOLD CROSS VALIDATION

I used a K fold cross validation to obtain a reliable estimate of a model's out of sample predictive accuracy and to compare the different types of models and the tuning parameters, the cross-validation function can be used for any models to decide on the models and for tuning parameters.
Below models were run in the k fold cross validation and below are the average values of AUC for these models, I used this validation for improving the tuning parameters.

Table 3

Models	No of trees	Average AUC Value
GBM	1000	0.967426285
GBM	2500	0.968249474
GBM	3500	0.967796656
RandomForest	1000	0.959665695
RandomForest	2500	0.95950543
RandomForest	3500	0.960229954
RandomForest	5000	0.959777159

Decision Trees: classification tress, Random Forests, Bagging and Boosting Model:

Tree based models can be used for regression and classifications models, here I have used these models for classification, tree based models are useful and simple for interpretation. Firstly, I used the simple tree based model to find the accuracy and profits, my model yielded a misclassification error rate: 0.1426. With this model, I got a profit of 11149 with 1168 emails.
Bagging, the next decision tree model I built was bagging model, the accuracy for this model was around 88% for classification and a profit of 12099 with 1038 emails.
Random forest, the next decision tree model I built was random forest model, the accuracy for this model was around 89% for classification and a profit of 11180.5 with 1058 emails.
Next, I built a boosting model, I used adboost algorithm and the maximum profit was 11205 with 1024 emails.
Then I tried series of GBM models and the results were as below:
For GBM model with 1000 trees gave a profit of 11549 with 1374 emails

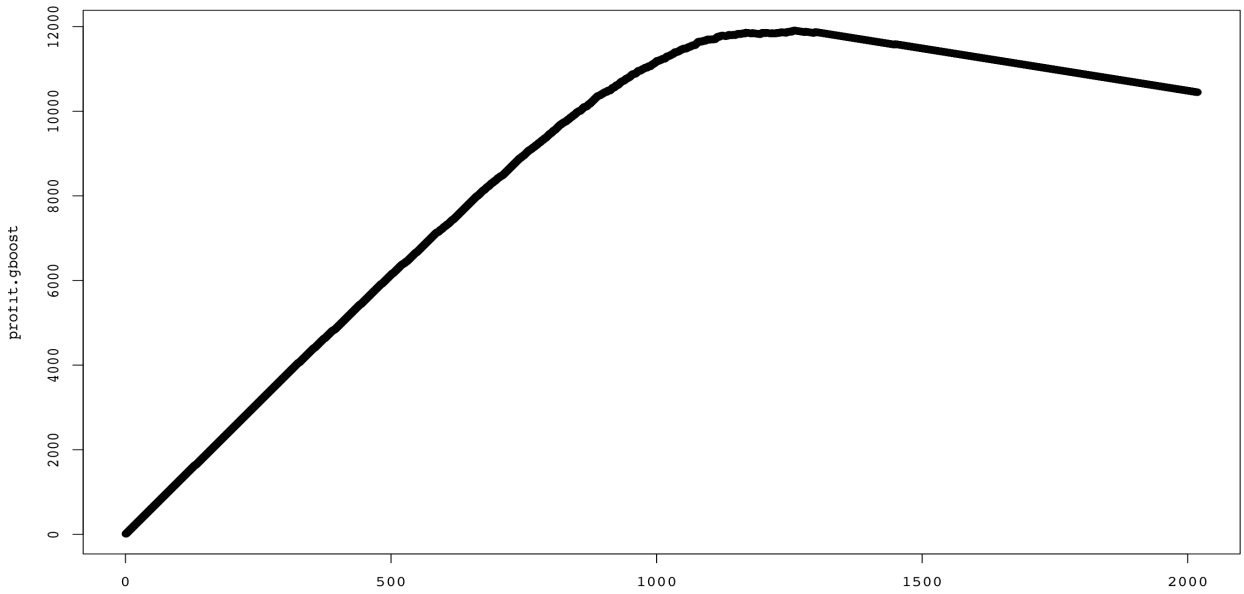
Next I ran a GBM model with 2500 trees, it gave a profit of 11831.5 with 1240.0 emails and an accuracy of 84 %
Then built a xgboost model, which gave a profit of 11856.5 with 1271 emails

Decision trees model comparison:

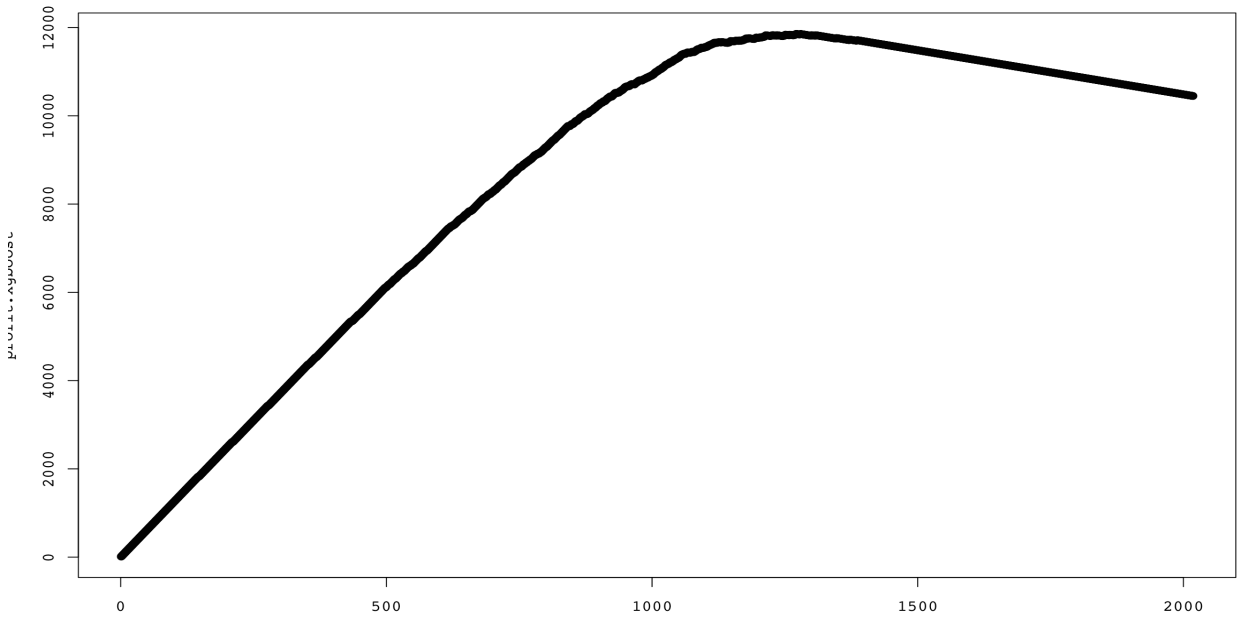
Table.4

Decision trees	Classification error rate	no of email	profit projected
Tree model	16%	1168	11149
Bagging	12%	1048	11171.5
Boosting	13%	1259	11909.5
Random Forest	11%	1058	11180
xgboost	10%	1271	11856.5

Plot2: gboost(gbm model with no of trees=2500)
gboost model, as we can see that around 1259 emails the profit is max and then it goes down gradually.



Plot3: XGBOOST model, as we can see that around 1217 emails the profit is max and then it goes down gradually.



Linear Discriminant Analysis (LDA) :

I built few models using LDA, but did not get yield great results, from the model we can see that the 49 % of the training observations corresponds to non-donor and below screenshot also shows the group means which are the average of predictors in each class.
This model gave us 82 % classification accuracy and a profit of 11624.5 with 1329 emails.

```
> model.lda1
Call:
lda(donr ~ reg1 + reg2 + reg3 + reg4 + home + chld + hinc + I(hinc^2) +
    genf + wrat + avhv + incm + inca + plow + npro + tgif + lgif +
    rgif + tdon + tlag + agif, data = data.train.std.c)

Prior probabilities of groups:
       0        1
0.499247 0.500753

Group means:
         reg1       reg2       reg3       reg4       home       chld       hinc I(hinc^2)       genf       wrat
0 -0.05653402 -0.2474198  0.1044670  0.1264554 -0.2894101  0.5315373 -0.02776724 1.4831064  0.01729059 -0.2496107
1  0.05636400  0.2466756 -0.1041528 -0.1260751  0.2885397 -0.5299386  0.02768373 0.5178453 -0.01723859  0.2488600
         avhv       incm       inca       plow       npro       tgif       lgif       rgif       tdon       tlag
0 -0.1247827 -0.1582579 -0.1395470  0.1434778 -0.1359105 -0.1160341 -0.02570108 -0.01491827  0.09640838  0.1413265
1  0.1244074  0.1577819  0.1391273 -0.1430463  0.1355018  0.1156851  0.02562379  0.01487340 -0.09611843 -0.1409015
          agif
0 -0.009475162
1  0.009446666
```

Logistics regression:

logistic regression models the probability that Y belongs to a category, here i achieved the best model by forward step elimination method. This method gave me a classification accuracy of 83 % and projected profit of 11642.5 with 1291 emails. Below is the summary of the model.

```
> summary(logistic_model3)

Call:
glm(formula = donr ~ reg1 + reg2 + reg3 + reg4 + home + chld +
    hinc + I(hinc^2) + genf + wrat + avhv + incm + inca + plow +
    npro + tgif + lgif + rgif + tdon + tlag + agif, family = binomial("logit"),
    data = data.train.std.c)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3544  -0.2717   0.0271   0.3846   2.8757
```

Quadratic Discriminant Analysis (QDA)

LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class- specific mean vector and a covariance matrix that is common to all K classes. Quadratic discriminant analysis (QDA) provides an alternative approach. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to per- form prediction. With my QDA model I got a classification accuracy of 79 % and a projected profit of 11219.5 with 1372 emails.

```
> model.qda
Call:
qda(donr ~ reg1 + reg2 + reg3 + reg4 + home + chld + hinc + I(hinc^2) +
    genf + wrat + avhv + incm + inca + plow + npro + tgif + lgif +
    rgif + tdon + tlag + agif, data = data.train.std.c)

Prior probabilities of groups:
       0        1
0.499247 0.500753
```

SVM (support vector machine)

The support vector machine is a generalization of a simple and intu- itive classifier called the maximal margin classifier. I ran the SVM model with various cost values ranging from 0.001, 0.01, 0.1, 1,5,10,100. Out of these models selected the best model based on cross validation error rate.

```
- best parameters:
 cost
   10

- best performance: 0.1573683

- Detailed performance results:
   cost     error  dispersion
1 1e-03 0.1588834 0.007805793
2 1e-02 0.1581227 0.013825032
3 1e-01 0.1581202 0.014193017
4 1e+00 0.1578702 0.013363852
5 5e+00 0.1576189 0.014240381
6 1e+01 0.1573683 0.013859874
7 1e+02 0.1573683 0.013859874
```

Based on the above stats the model with cost 10 was selected and prediction were made on the validation set. The profit projected by this model is 10403.5 with 1055 emails and the classification accuracy for the model is 83 %.

Summary results of all classification model:

Classifications models	Classification error rate	No of Projected mailing	Projected profits($)
Tree model	16%	1168	11149
Bagging	12%	1038	11171
GBM	13%	1259	11831.5
RandomForest	11%	1058	11180
xgboost	10%	1271	11856.5
KNN	18%	1256	11263
Logistic regression	17%	1291	11642.5
QDA	21%	1372	11219.5
svm	17%	1055	10403

PREDICTION MODELS

The other part of the project is to predict the amount of donation based on the characteristic. I developed various prediction models and few will be discussed below and the summary of all the models will be provided at the end.

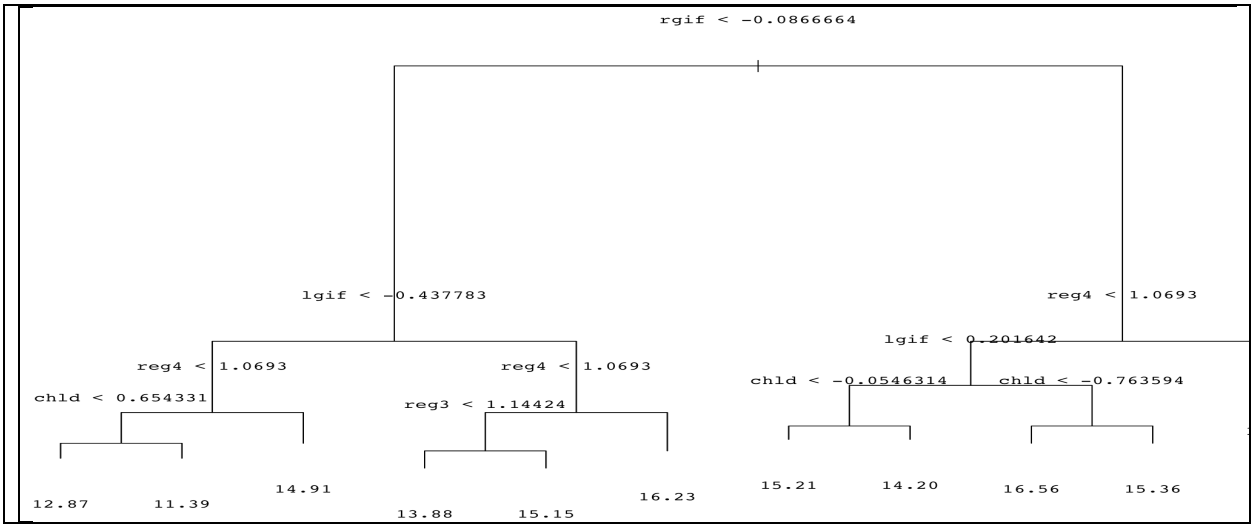
Multi regression model

Multiple regression is a logical extension of the principles of simple linear regression to situations in which there are several predictor variables, some benefits of linear regression models are that they have low bias which makes them less prone to overfitting versus more flexible methods and they are also highly interpretable. The multiregression model was built with all the variables chosen from forward and backward stepwise regression, the best mean prediction error I got from the multi regression model is 1.870095 and standard error of 0.168765

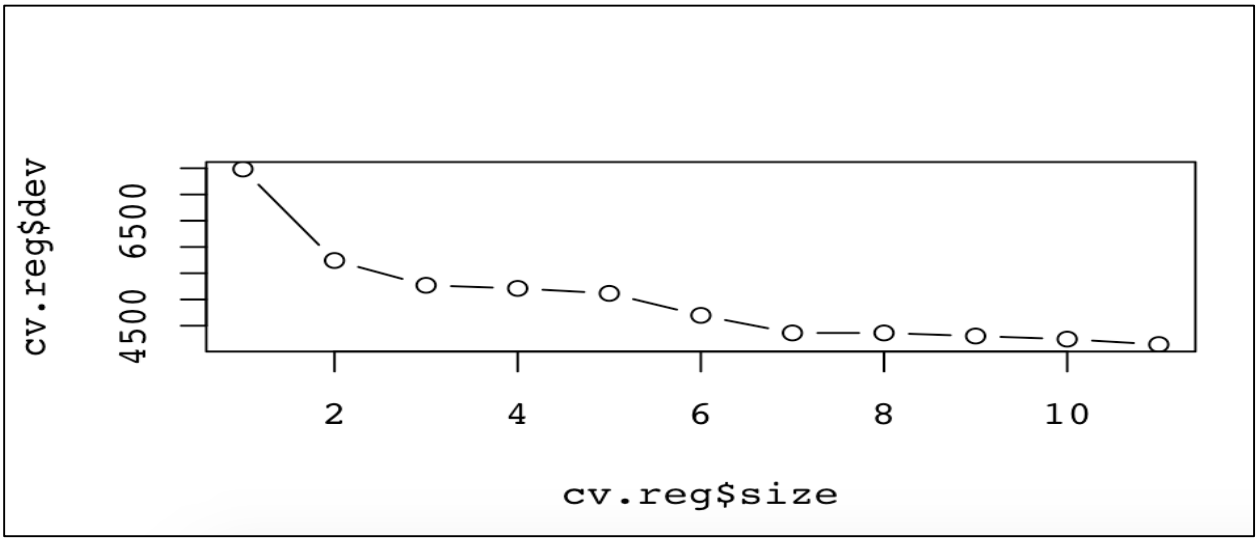
Regression trees:

Here I fit a regression trees to the dataset, the model gave a prediction error of 2.241075 and standard error of 0.1920681, I used the cv function to see if the pruning the trees would help to improve the prediction error, but it did not help much, the cross-validation function chose the most complex tree. Below plots shows the trees for the regression model and the output of the CV function.

Plot 4:



Plot 5- cross validation model vs size



Decision trees, Bagging, RandomForest, Gradient Boosting machine, xgboost:

Tree based models can be used for regression and classifications models, here I have used these models for regression, tree based models are useful and simple for interpretation. For simple pruned tree based model gave a prediction error of 2.241075 and standard error of 0.1920681, whereas the bagging model with 500 trees gave the prediction error of 1.723826 and standard error of 0.1920681. With RandomForest model I got a prediction error of 1.665959 and standard error of 0.1728008.
(GBM)Apart from being used in classification problems, GBM models can also be used for prediction. GBM models that were composed of 2500 trees appeared to perform well in the classification setting and so I considered to use it for prediction with 2500 trees, after series of iteration I found that with 2000 trees and shrinkage value of 0.001 for prediction gave the higher performing model, the prediction error came down to 1.37 and the standard error was 0. 1728008.This GBM model had the lowest mean prediction error considered thus far.

Support Vector Machine

Support vector machines are called Support Vector regressions (SVR) when used in the prediction setting. It contains tuning parameters such as cost, gamma and epsilon. First a SVM model was fit with a cost as 1.5 which resulted in prediction error of 1.685081 and standard error of 0.182173
Then I built a model with a fixed gamma value of 0.5 and performed 10-fold CV to find useful values for the cost and epsilon parameters. The potential epsilon values I considered in the CV process were 0.1, 0.2,0.3,1.5,2 along with potential cost values of 0.01, 1 ,5,10,50. After performing 10-fold cross validation, it appeared that 0.2 and 1 were promising values for epsilon and cost respectively. Using a cost value of 1, epsilon value of 0.2 and a gamma value of 0.5, i obtained a support vector regression model with 1,385 support vectors. When this was applied to the validation set, it resulted in a mean prediction error of 1.669018 and a standard error of 0.1828826

Below screen shot contains the details of the best model for SVM.'

```
> svm.bst.model

Call:
best.tune(method = svm, train.x = damt ~ ., data = data.train.std.y, ranges = list(epsilon = c(0.1,
    0.2, 0.3, 1.5, 2), cost = c(0.01, 1, 5, 10, 50)), kernel = "radial")


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.05
    epsilon:  0.2


Number of Support Vectors:  1385
```

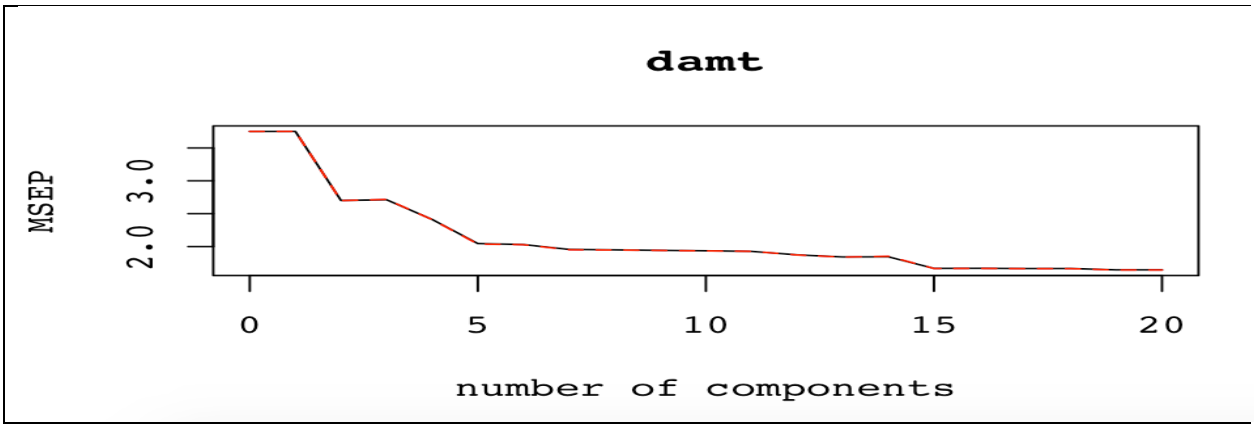
RIDGE and LASSO:

Ridge regression is like least squares, this model creates shrinkage of the predictors by using a tuning parameter λ to obtain a set of coefficients estimates. I built a ridge model and the mean prediction error I got from the model was 1.873231 and prediction error of 0.1711169.
I fitted the data for the lasso regression model and the prediction error was like the ridge 1.86133 and standard error was 0.1694185

PCR

For PCR model uses clustering to decrease the dimensionality of the problem space. From the plot 6 we can see the around 15 components reduce the mean squared error to the lowest point. This model gave me a prediction error of 1.865497 and standard error of 0.1698902.

Plot 6: Mean Standard Error of Prediction for models with increasing number of components.



Summary of prediction model:

Prediction models	Mean prediction error	Standard error
Least Squares Regression	1.87	0.16
SVM	1.66	0.18
Ridge Regression	1.87	0.17
Lasso Regression	1.86	0.16
tree model	2.24	0.19
Bagging	1.72	0.19
randomForest	1.66	0.17
GBM	1.37	0. 17
Principal component Regression	1.86	0.16

Conclusion and Results:

The use of predictive modeling to create informative, statistically supported business decisions has existed for years within many industries. Although the process has been slightly slower in adoption within the field of philanthropic giving, the plethora of research and substantial availability of modeling programs specifically for giving offices indicate that it is a beneficial tool in non-profit organizations, in this project the organization is trying to maximize the net profit by capturing the likely donors, this project would benefit the organization and save the funds by targeting only the likely donors.

As far the data goes, the data used for this project was clean and did not have any missing data or any erroneous data, because of this no much time was spent in the cleansing and imputing, After EDA, we transformed few of the predictors in the data.

The whole purpose of this project was to come up with good classification and predictive models. I fitted various classification and predictive models to the data, only few are highlighted and used in the report. For both classification and predictive models, the GBM gave the best results with maximum profit and less mean prediction error, Various other models like xgboost, random Forest were promising for the classification models.

Finally, I chose GBM models with (2500 trees and shrinkage of 0.05) for the classification model as it gave me the maximum profit.
And for prediction model, I chose GBM model with (2000 trees, shrinkage=0.01) and it gave me the minimum prediction error.

REFERENCES

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R

Mount J and Zumel N (2014). Practical Data Science with R. Manning Publication Co.

http://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1003&context=honors_mathematics
http://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html
http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html