Group Project

Introduction

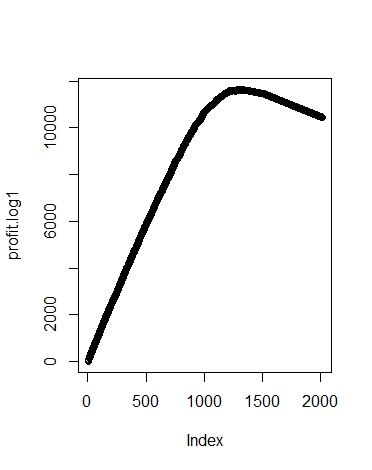
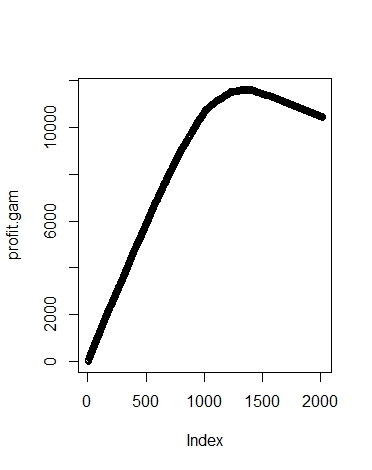
This research is based on the data generated from a charitable organization who wishes to develop a machine learning model to improve the cost-effectiveness of their direct marketing campaigns to previous donors. according to their recent mailing records, the typical overall response rate is 10%. Out 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.

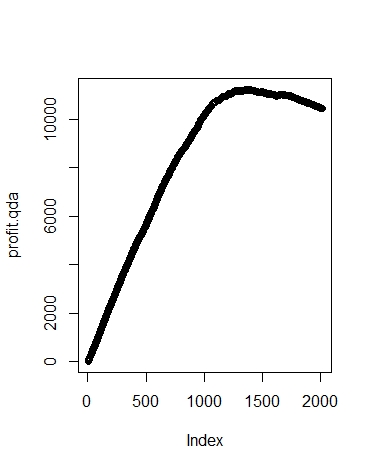
../../../../../Downloads/ggplotpdf.pdf

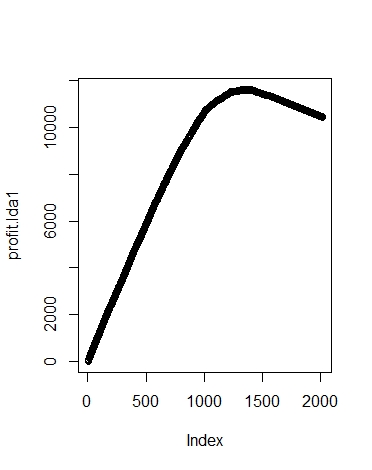
Part 1: Prediction modelling for DNOR

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.

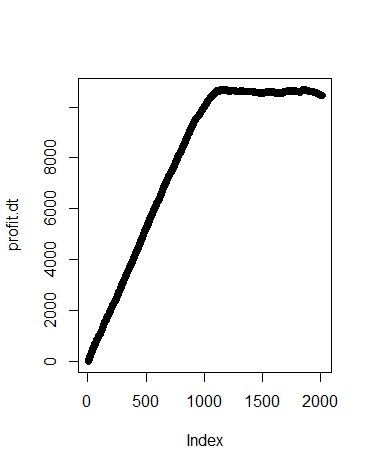
Logistic GAM Model: Maximum profit earned with Logistic GAM Model is 11624.5.



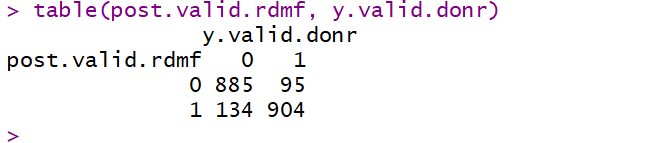


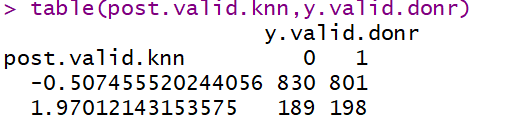
LDA Model: Maximum profit earned with LDA Model is 11624.5.

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



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





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

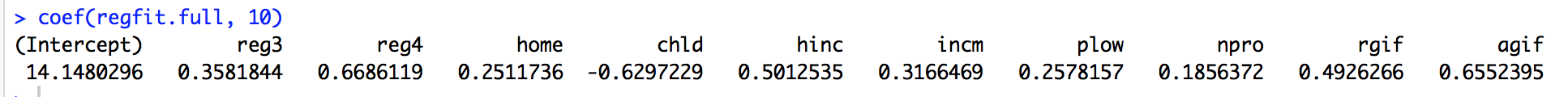
In this part, we Develop a prediction model for the DAMT variable using any of the variables as predictors (except ID and DONR). We fitted as many candidate models as we could generate, including least squares regression, best subset selection with bic, cp, adjusted R square, validation set approach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

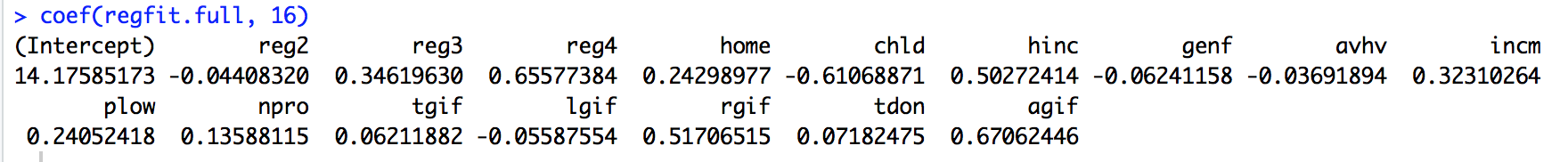
Best Subset selection- using BIC model

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



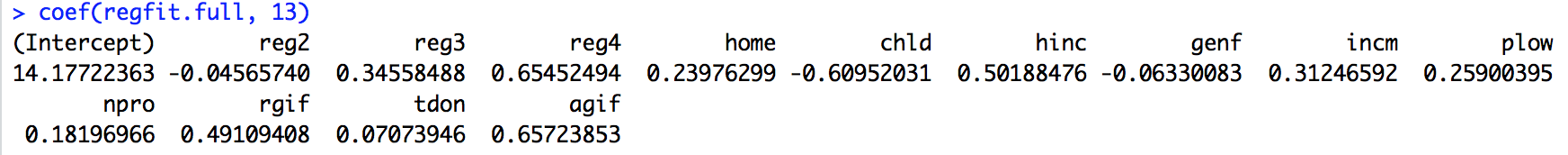
The mean prediction error of fitting this model with the validation data is 1.85794

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.

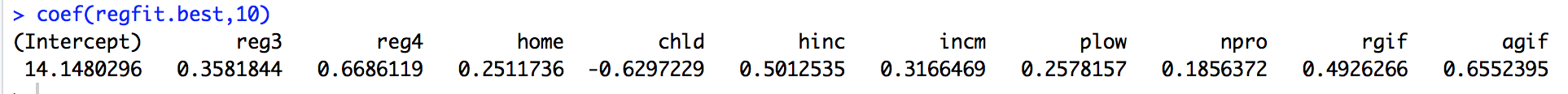
Best Subset selection- using Cp

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



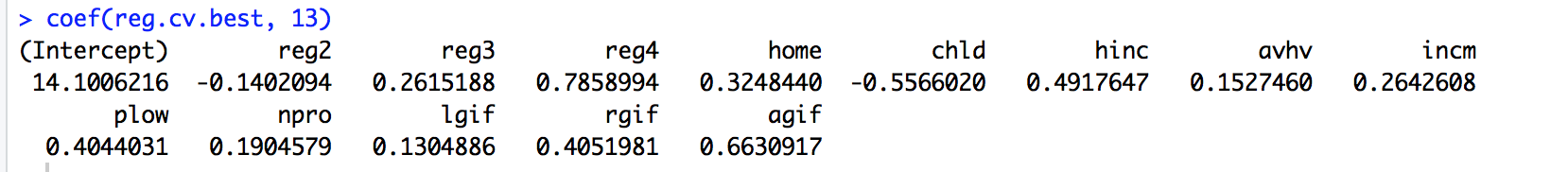
The mean prediction error of fitting this model with the validation data is 1.8599.

Best Subsets using validation set approach

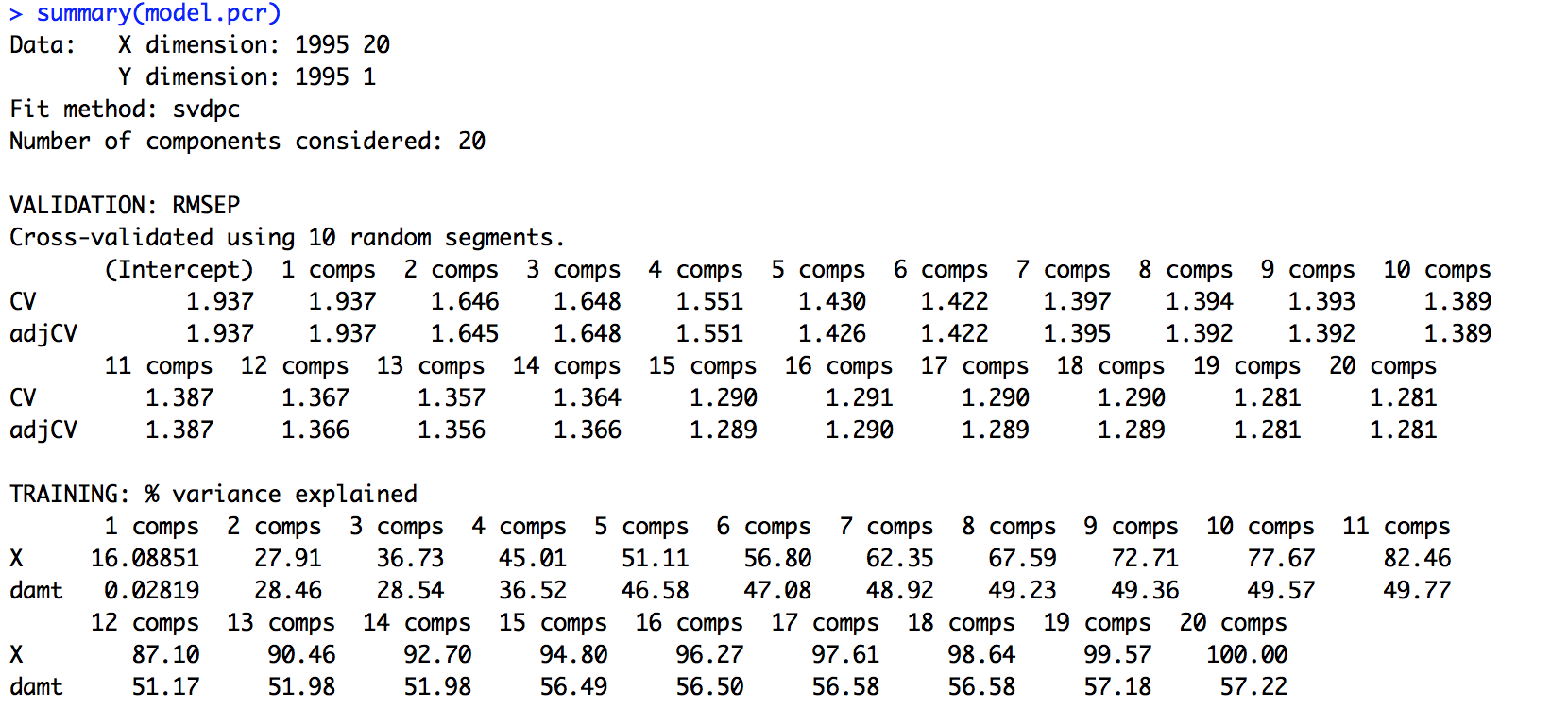
The validation set approach suggests the optimal number of variables should be 10, as pictures below. The relevant variables are shown following. The mean prediction error of fitting this model with the validation data is 1.8579.

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.



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

Principle Components Regression and Partial Least Regression

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

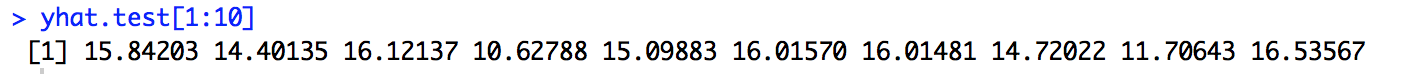
The figures below provide the estimated lasso regression coefficient values and CV errors for altering values of lamda. The best lambda is 0.0088 and the MPE for this model is 1.8613.

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: