Final Project: PA422 Machine Learning
Jack Doig
PREDICT 422 Section 55

Contents

1. Introduction

Project goals

The aim of this project is to use machine learning in order to optimize the investments made by a charitable organization in their mail-outs to potential donors.

The specific goals of the project are to determine which donors to target and then calculate the expected donation from those donors, then factor in costs, and maximize the resulting net income from a direct marketing campaign.

Data Set

The dataset contains 20 continuous and categorical predictors, which cover a range of demographic and historical behaviour information with respect to the donors. There are 2 response variables: the binary DONR variable representing whether an individual will donate, and DAMT, representing the amount that a person will donate. There are 3984 training observations, 2018 validation observations and 2007 test observations.

Process

The process undertaken is firstly to explore the data so as to understand its structure and relationships. A number of different techniques are used so as to reveal information which may be useful in the data preparation and modelling stages.

Secondly data is prepared, which involves any imputation of variables, correction of missing variables and the derivation of any new variables through variable transformation.

The modelling stage attempts to fit a number of different modelling, variable selection and ensemble methods to the data, and finally those models are judged against the validation set.

Outcome

A final model will be selected and predictions made and submitted against the test set for assessment. The measure of success for the classification model is profit on the known donations, whereas the measure of success for predicting the donor amount is the mean squared error.

2. Data Exploration

The purpose of Exploratory Data Exploration is to examine the data, its internal structure, unusual values and the relationships between variables so as to provide clues as to how best transform and model the data.

Missing Values

All columns were explored for the existence of missing values as these values may need to be handled so as to enable more effective modelling. This was tested using the R *is.na* function as well as testing for the empty string.

Fortunately, there were no missing values in the data set according to these criteria, outside of ‘na’ values in the response variables (donr, damt) for the test set, which is expected, as these are the values to be predicted.

Basic Statistics of continuous variables

In order to get a sense of the continuous variables, a set of descriptive statistics was generated using the BasicStats package on the full data set.

	damt	chld	hinc	wrat	avhv	incm	inca	plow	npro	tgif	lgif	rgif	tdon	tlag	agif
nobs	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00	8009.00
NAs	2007.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Minimum	0.00	0.00	1.00	0.00	48.00	3.00	12.00	0.00	2.00	23.00	3.00	1.00	5.00	1.00	1.29
Maximum	27.00	5.00	7.00	9.00	710.00	287.00	305.00	87.00	164.00	2057.00	681.00	173.00	40.00	34.00	72.27
1 Quartile	0.00	0.00	3.00	6.00	133.00	27.00	40.00	4.00	36.00	63.00	10.00	7.00	15.00	4.00	6.97
3 Quartile	14.00	3.00	5.00	9.00	217.00	54.00	68.00	21.00	82.00	137.00	25.00	20.00	22.00	7.00	14.80
Mean	7.21	1.72	3.91	6.91	182.65	43.47	56.43	14.23	60.03	113.07	22.94	15.66	18.86	6.36	11.68
Median	0.00	2.00	4.00	8.00	169.00	38.00	51.00	10.00	58.00	89.00	16.00	12.00	18.00	5.00	10.23
Sum	43269.00	13753.00	31304.00	55375.00	1462812.00	348185.00	451933.00	113994.00	480790.00	905575.00	183730.00	125435.00	151079.00	50963.00	93550.41
SE Mean	0.10	0.02	0.02	0.03	0.81	0.28	0.28	0.15	0.34	0.96	0.33	0.14	0.06	0.04	0.07
LCL Mean	7.02	1.69	3.88	6.86	181.05	42.93	55.88	13.94	59.37	111.20	22.28	15.39	18.74	6.28	11.54
UCL Mean	7.40	1.75	3.94	6.97	184.24	44.02	56.97	14.53	60.70	114.94	23.60	15.93	18.99	6.44	11.82
Variance	54.19	1.96	2.15	5.90	5288.20	610.41	615.91	179.87	920.83	7305.99	896.88	154.62	33.45	13.72	43.12
Stdev	7.36	1.40	1.47	2.43	72.72	24.71	24.82	13.41	30.35	85.48	29.95	12.43	5.78	3.70	6.57
Skewness	0.12	0.27	0.01	-1.35	1.54	2.05	1.94	1.36	0.31	6.55	7.81	2.63	1.10	2.42	1.78
Kurtosis	-1.83	-0.80	-0.09	0.79	4.49	8.31	7.87	1.89	-0.62	107.52	110.38	13.92	2.12	8.41	6.02

This indicates the quality of the continuous data is generally very good, and reveals some interesting points about the data:

- It reconfirms that there are no NAs amongst the continuous predictors
- Some variables (tgif, lgif) have a strong skewness (>5) amongst the distribution
- Some variables (incm, inca, tgif, lgif, rgif, tlag, agif) have a strong (>5) kurtosis
- With a median wealth rating (wrat) of 8, and average of 6.91, as well as a low percentage of low income in neighbourhood (plow), this indicates a high wealth amongst the donor population
- The median (2) and mean number (1.72) of children are consistent with national averages, indicating good data
- Columns indicating gift size (largest gift, most recent gift, average gift) are generally believable, with the range of largest gifts from $3 to $681, and the median average gift being $11.68, and the median most recent gift being $12

Basic Statistics and Outliers in the categorical variables

The categorical variables were tested for low representation in a particular class, or less than 1% of the values falling into that class.

Among the categorical variables of region, gender, and home ownership, there is a balanced representation across the different classes:

region	count	%-age
0	1661	20.74%
1	1605	20.04%
2	2555	31.90%
3	1071	13.37%
4	1117	13.95%

Genf	count	%-age
0	3161	39.47%
1	4848	60.53%

home	count(id)	%-age
0	1069	13.35%
1	6940	86.65%

There is no reason to believe anything is incorrect about this distribution of categorical data. The observations are:

- There is a high representation of female donors in the data
- There is a very high representation of home ownership in the data

Outliers in the continuous variables

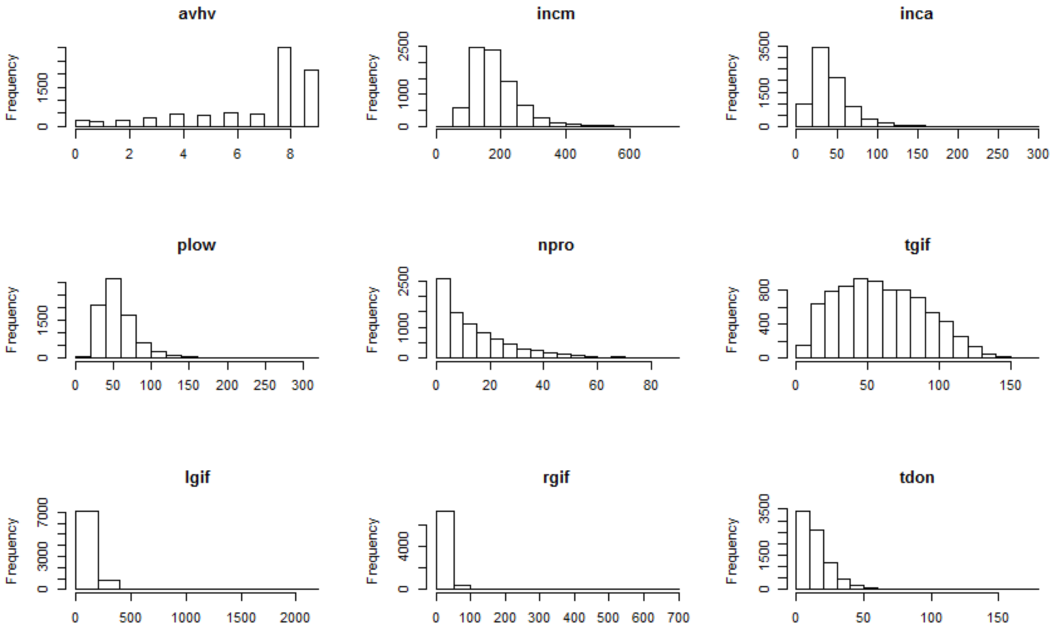
There are a high number of outliers in the continuous variables, when assuming a normal distribution of the raw data, here is a summary of points sitting outside of various multiples of standard deviation:

Predictor	>3 std deviations	>5 std deviations	>7 std deviations
wrat	97	0	0
avhv	61	7	1
incm	61	9	4
inca	64	10	2
plow	52	1	0
npro	6	0	0
tgif	51	12	5
lgif	68	25	13
rgif	64	11	4
tdon	95	0	0
tlag	79	15	3
agif	56	9	2

While there are a large number of values which sit outside of 3 standard deviations and above, there is no reason to believe this is bad data. This may suggest that we need to employ transformations that bring normality to the distributions, as this is required by several of the models.

Distribution of continuous variables

Exploring the distributions of the continuous variables reveals s a range of distributions. Below is a sample of the distributions of continuous predictors:

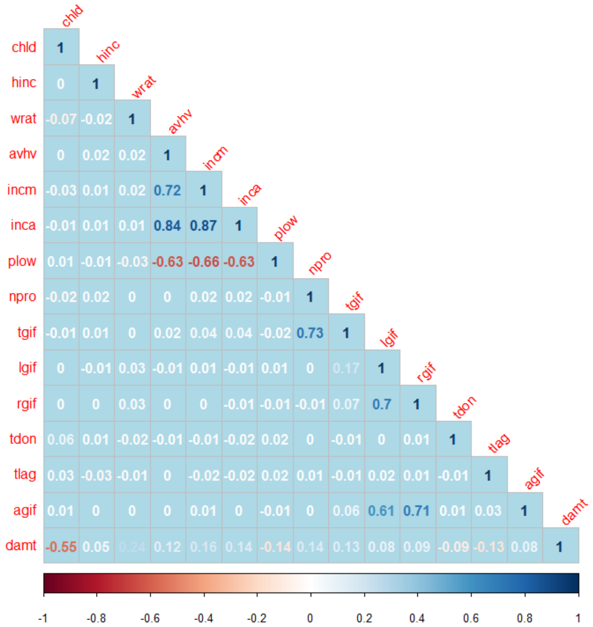


As can be seen in the picture above, the distributions for some of the variables are non-normal, which may cause some undesired behaviour when modelling. This could be addressed during the data preparation stage.

Correlations

Exploring correlations amongst the continuous predictor variables is useful as it may to possible to reduce dimensionality in the predictor data set. This might improve the model by increasing the parsimony, and preventing the problems of caused by multi-collinearity in some data models.

Below is a chart of correlations amongst the continuous predictors.



	chld	hinc	wrat	avhv	incm	inca	plow	npro	tgif	lgif	rgif	tdon	tlag	agif	damt
chld	1														
hinc	0	1													
wrat	-0.07	-0.02	1												
avhv	0	0.02	0.02	1											
incm	-0.03	0.01	0.02	0.72	1										
inca	-0.01	0.01	0.01	0.84	0.87	1									
plow	0.01	-0.01	-0.03	-0.63	-0.66	-0.63	1								
npro	-0.02	0.02	0	0	0.02	0.02	-0.01	1							
tgif	-0.01	0.01	0	0.02	0.04	0.04	-0.02	0.73	1						
lgif	0	-0.01	0.03	-0.01	0.01	-0.01	0.01	0	0.17	1					
rgif	0	0	0.03	0	0	-0.01	-0.01	-0.01	0.07	0.7	1				
tdon	0.06	0.01	-0.02	-0.01	-0.01	-0.02	0.02	0	-0.01	0	0.01	1			
tlag	0.03	-0.03	-0.01	0	-0.02	-0.02	0.02	0.01	-0.01	0.02	0.01	-0.01	1		
agif	0.01	0	0	0	0.01	0	-0.01	0	0.06	0.61	0.71	0.01	0.03	1	
damt	-0.55	0.05	0.24	0.12	0.16	0.14	-0.14	0.14	0.13	0.08	0.09	-0.09	-0.13	0.08	1

There is a small amount of correlation amongst the raw predictors, which could be dealt with at the data preparation stage.

Observations:

- There are correlations among the average, largest and most recent gift amount, suggesting people are consistent in donation behavior
- There is an unexpected strong negative correlation between the chld variable and the damt, or people with fewer children donate more
- Finally, the predictors related to a general measure of wealth are correlated.

Data consistency

Studying the internal consistency, or making judgments about the correctness of the data as it relates to the real-world or other data points in the data set, can uncover data which is wrong and allow for imputation of that data.

Several areas were found to have suspicious internal consistency; they are outlined here:

Area/Rule	Statement of correctness (variable affected)	# rows violating correctness in full data set	% rows affected
Gifts	Average Gift (agif) Cannot be higher than largest gift (lgif)	1010	12.6%
Gifts	Largest gift (lgif) cannot be lower than most recent gift (rgif)	1638	20.45%
Gifts	If largest gift (lgif) and most recent gift (rgif) are not equal, they cannot sum to greater than total gift (tgif)	760	9.48%
Gifts	Average Gift (agif) cannot be larger than total Gift (tgif)	1	0.01%
Gifts	If recent gift (rgif) = total gift (tgif), then this must also be equal to average gift (agif) and largest gift (lgif)	14	0.17%
Gifts	If largest gift (lgif) = total gift (tgif), then this must also be equal to average gift (agif) and recent gift (lgif)	322	4.02%

Note the gift issues outlined above are not mutually exclusive, or several observations are affect by two or more inconsistency problems. The above findings indicate there is some inconsistent data in the gift information, which may need to be imputed at the data transformation stage.

3. Data Preparation

Derived New Variables, interaction terms and dropped terms

A large number of interaction variables were tested against a simple model so as to determine if they improved performance on the MSE and the Profit calculation for simple models.

The interactions were based on the knowledge of the domain, and the understanding of the collinearity in the model. The interaction terms that were used in the final models are below:

Variable	Notes	Outcome
Number of Donations	tgif/agif dervies the number of donations made	Not used: did not improve either model type

incm*inca	This improved performance on the prediction model	Used in prediction model only
wrat*hinc	This improved performance on the prediction model	Used in prediction model only
avhv	Removing this variable improved performance on some of the prediction models	Removed from prediction modelling
wrat^2	Improve Classification model performance	Used in classification model only
incm^2	Improve Classification model performance	Used in classification model only
hinc^2	Improve Classification model performance	Used in classification model only
tdon^2	Improve Classification model performance	Used in classification model only

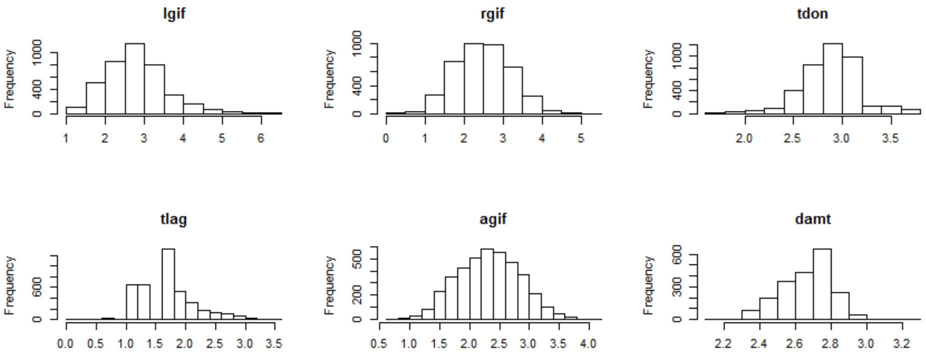
Transforms

A number of standard transforms were explored and examined to see if they increased correlation with the response variable damt, and improved the out-of-sample performance in the CV set for both the classification and prediction models.

Transform Type	General Notes	Outcome
X^2	This improved performance on the classification model significantly for some variables Little impact on the prediction model	Apply to selected classification variables
Square Root(X)	Increase correlation the most Reduced performance on Classification and Prediction models	No action
Ln(X)	Increased performance on prediction model significantly	Apply to continuous variables
X^3	Degraded performance of all models	No action

Given the distributions of the categorical data as outlined in section 1.6 above, a number of log transformations were applied to the predictors so as to create normality among the predictors before modelling.

A sample of the distributions after this transformation are below:



The distributions above following the log transforms follow a much clearer normal distribution as compared to those before in the EDA section above.

Imputing inconsistent data

The inconsistent data around the gift amounts (recent, largest, average and total) is dealt with as follows:

Area/Rule	Statement of correctness (variable affected)	% rows violating	Action taken
Gifts	Largest gift (lgif) cannot be lower than most recent gift (rgif)	20.45%	Overwrite the largest gift with the most recent gift
Gifts	Average Gift (agif) Cannot be higher than largest gift (lgif)	12.6%	Overwrite the average gift with the most recent gift
Gifts	If largest gift (lgif) and most recent gift (rgif) are not equal, they cannot sum to greater than total gift (tgif)	9.48%	No Action
Gifts	If largest gift (lgif) = total gift (tgif), then this must also be equal to average gift (agif) and recent gift (lgif)	4.02%	No Action
Gifts	If recent gift (rgif) = total gift (tgif), then this must also be equal to average gift (agif) and largest gift (lgif)	0.17%	No Action
Gifts	Average Gift (agif) cannot be larger than total Gift (tgif)	0.01%	No Action

In each instance of inconsistent data, the rows were examined, and the most inconsistent of the four variables was imputed so as to create consistency across the rows. Generally, the largest gift column was the most inconsistent with the other three predictors. Note the first two transforms made a marginal improvement to the prediction modelling testing.

3. Model Building

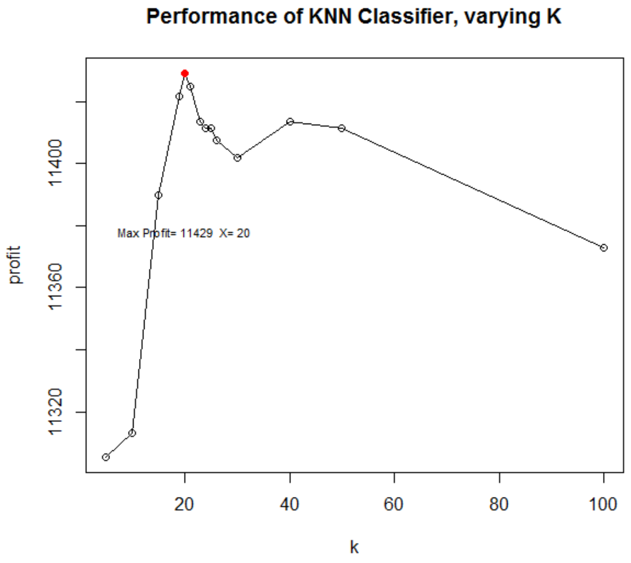
Classification Models

KNN

A KNN model was run with the transformed variables. A search was performed to optimize K, and below are the results the cross validation set, for the optimal k:

Model Type	Best K	Number of Mailings	Profit
KNN	20	1376	$11,429

The chart below shows how KNN performs as we vary K:



Logistic/Probit Models

Logistic and Probit models were run with the transformed variables. The results are as follows on the cross validation set:

Model Type	Number of Mailings	Profit
Logistic	1306	$11,685
Probit	1304	$11,689

Generalized Additive model (GAM)

A GAM model was run with the transformed variables. The results are as follows on the cross validation set:

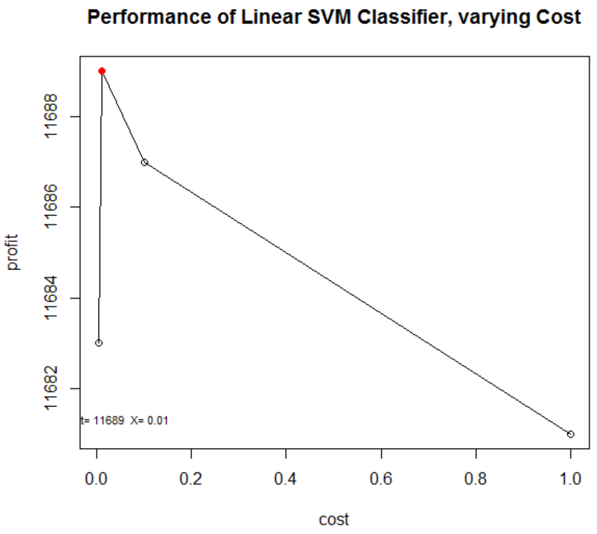
Model Type	Number of Mailings	Profit
GAM	1306	$11,685

Simple Vector Machine (SVM) for Classification

A SVM model was run with the transformed variables. The results are as follows on the cross validation set:

Model Type	Best Cost	Number of Mailings	Profit
SVM (Linear)	20	1376	$11,429

Below the graph shows the varying performance of the model against the cross validation set:



Linear/Quadratic Discriminant Analysis

A LDA model was run with the transformed variables. The results are as follows on the cross validation set:

Model Type	Number of Mailings	Profit
LDA	1300	$11,682.5
QDA	1357	$10,959.5

Continuous Prediction Models

As a general note, the performance of the models is validated against the validation set, and the measure is Mean Squared Error

Ordinary Least Squares

A standard OLS regression was run against the transformed variables:

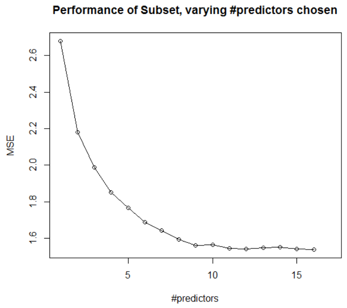
Model Type	MSE
OLS	1.538899

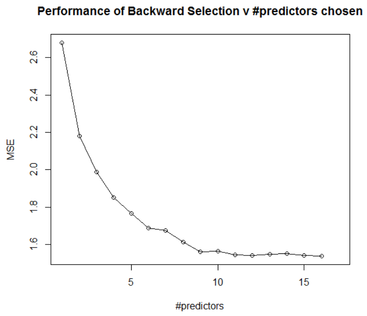
Subset, Forward and backward Stepwise Regression

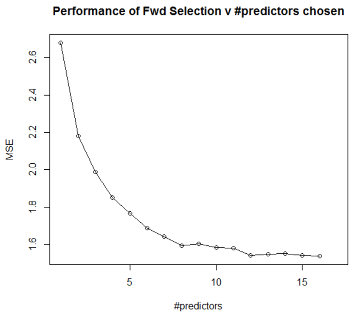
A subset regression was run against 17 transformed variables:

Model Type	Number of variables	MSE
Subset	17	1.540677
Forward	17	1.540677
Backward	17	1.540677

The below chart shows the performance of the CV sample as the number of predictors in the subset changes for each of the three types of variable selection processes:





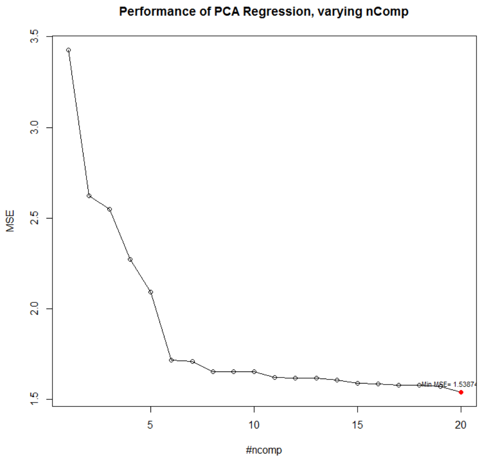


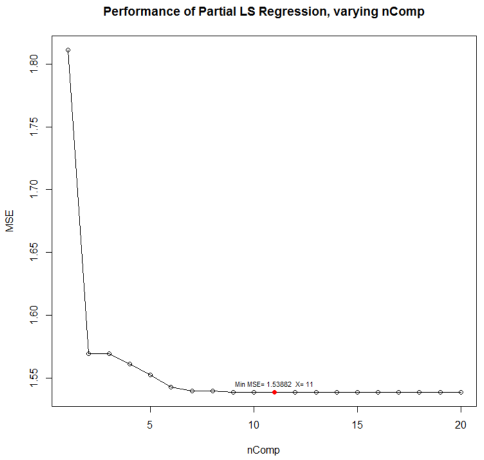
Principal Components/Partial Least Squares Regressions

Below is the outcome for the best performing

Model Type	nComp	MSE
PCA Regression	20	1.538744
PLS Regression	11	1.538824

The image below shows how the MSE changes as components are added for PCA and PLS regressions respectively:

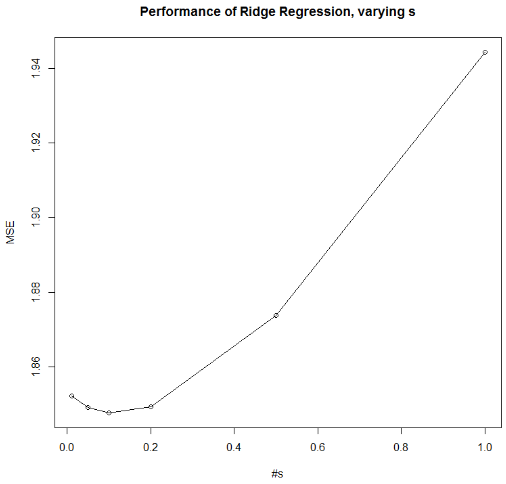


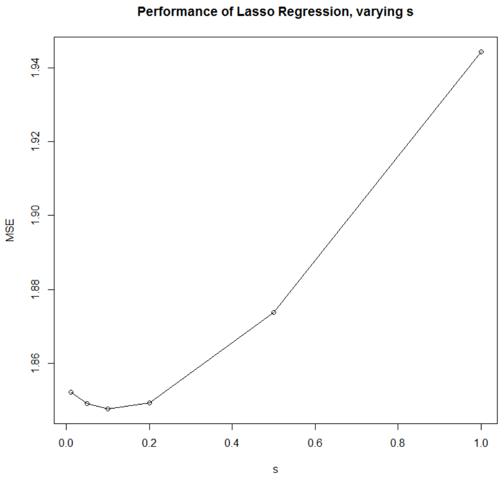


Ridge and Lasso Regression

Below is the outcome for the best performing

Model Type	s	MSE
Ridge Regression	0.1253756	1.84749
Lasso Regresssion	0.007592436	1.852077





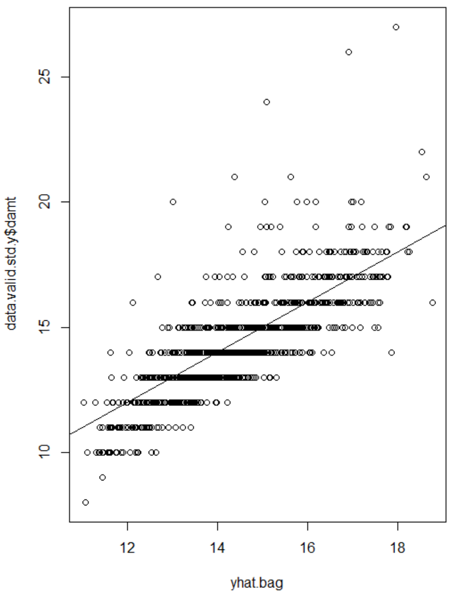
Boosted Trees and Bagging

Below is the outcome of a boosted tree applied to the transformed variables

Model Type	#Trees	MSE
Boosted Tree	5000	1.534127
Bagged Tree	–	1.664953

The boosted Tress model allows us to understand the relative importance of the variables. Whil we can get a sense of how well the data is fitting the CV set (bagged example on left):

	rel.inf
lgif	22.94354638
rgif	17.84320016
agif	14.58355678
reg4	12.69923025
chld	9.484831371
hinc	5.374311204
wrat	4.897415937
reg3	3.292889058
tgif	2.258113125
plow	1.649814975
incm	1.629396714
reg2	1.546612884
inca	0.676566346
home	0.404373315
npro	0.243817218
tdon	0.214036424
reg1	0.188881136
tlag	0.061330445
genf	0.008076276

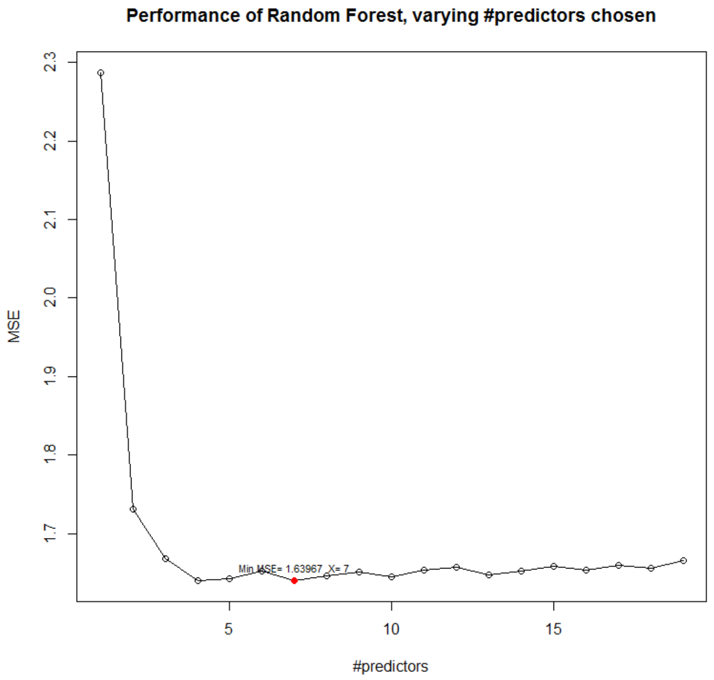


Random Forests

Below is the outcome for the best performing Random Forest

Model Type	# Predictors	MSE
Random Forest	7	1.639673

Below we show how the performance of the Random Forest changes as we vary the number of predictors contributing to the tree:

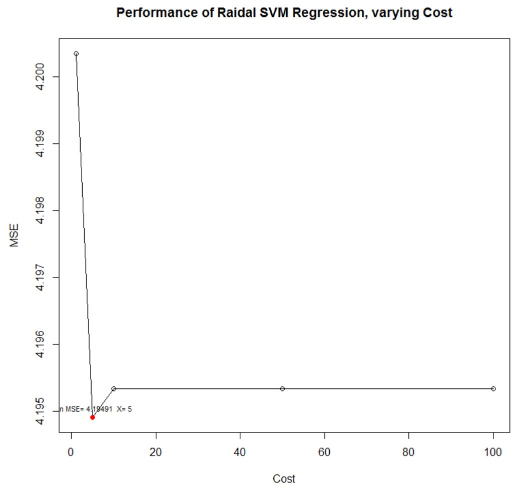


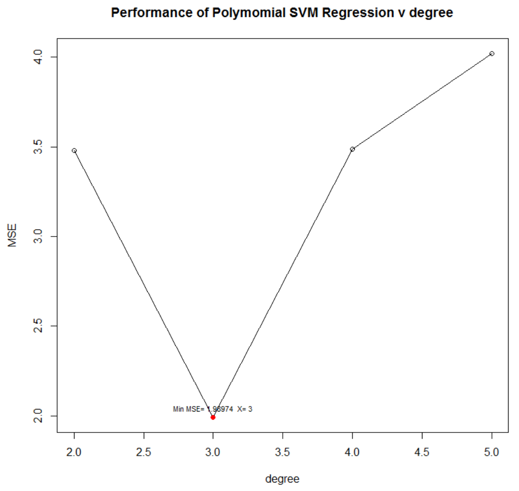
Simple Vector Machine (SVM) for Regression

Below is the outcome for the best performing SVM for regression

Model Type	Cost / Degree	MSE
SVM (Radial)	5	4.194906
SVM (Polynomial)	3	1.989741

The graph below shows the performance of the Radial and Polynomial Kernel SVM model as we vary the Cost parameter and the degree respectively:





4. Select Models

Given the model is for prediction, and not for inference, there is no serious need to make a highly interpretable model, therefore I am comfortable with choosing the best performing models on the metrics stated at the outset.

The best performing classification model is the Probit model applied to the transformed variables is below. The profit margin was used exclusively to judge the models.

Model Type	**Number of Mailings**	**Profit**
Probit	1304	$11,689

The best performing continuous prediction model is the Boosted tree, again applied to the transformed variables. The performance of this model on the cross-validation (out of sample) set is below. The MSE means the average square difference between the prediction and the actual value.

Model Type	**#Trees**	**MSE**
Boosted Tree	5000	1.534127

5. Conclusion

I was able to make some small gains against the sample model provided by trialling different transformations, data corrections, interaction terms, and models.

The highest value investment of time to improving the performance of the models was to rather focus on variable transformations and interaction terms; The log transform was by far the most important performance gain in the prediction model, and adding squared terms was the most important for the classification model.

While I tried many different model types, and tuned those requiring tuning, it didn’t feel like this was helping improve predictive performance. There were only marginal improvement, generally speaking, across the model types.

There seemed to be only a marginal improvement in finding and imputing logically incorrect gift data. This was disappointing as ~20% of the data had a data consistency problems, so this had seemed like an appropriate place to invest time.

There was very little difference in performance of the classification models, while the best performing prediction models were the ensemble methods, the PCA/PLS methods and simple OLS regression. The non-linear models, KNN and SVM, performed poorly.