SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL307

Predictive Modelling

Group-Based Assignment

January 2021 Presentation

Submitted by:

Name	PI No.
Choo Se Ying	CE211207
Han Ji Su	N1972389
Ho Zhong Ta Benjamin	W1711125

Tutorial Group: T02 / Group 7

Instructor's Name: Huang Bin

Submission Date: 23 February 2021

1. Introduction

Direct marketing campaigns are a common strategy used by companies to increase sales revenue. In this paper, we examine the use of telemarketing phone calls conducted in centralised contact centres for the sale of bank long-term deposits. Telemarketing calls in this context include both inbound and outbound calls which are differentiated by the party who triggered the call, i.e. customer or contact centre.

Optimising targeting for bank telemarketing is critical as the need to increase profits and reduce costs became increasingly important for banks after the 2008 financial crisis. In this paper, historical data from 2008 to 2013 was retrieved from a Portuguese bank to build predictive models to predict the result of a telemarketing phone call to sell long term deposits (Moro, Cortez, & Rita, 2014). The goal was to model the success of subscribing to the product using inputs that were known before the telemarketing call was made. An understanding of the characteristics of customers who are more likely to subscribe is valuable for bank campaign managers as fewer and more effective phone calls can be performed to help reduce time and costs for these marketing campaigns.

2. Data Understanding

Real data comprising 52944 phone records were collected from a Portuguese bank from May 2008 to June 2013. The dataset was unbalanced, as only 6557 calls were related to successes. The samples were divided into training and test data using a time ordered split. The training data was used for variable selection and included calls up to June 2012, in a total of 51651 samples. The test data was used for measuring the predictive performance of the examined models and included 1293 calls from July 2012 to June 2013. A total of 150 variables were initially identified and categorized into 4 types, namely telemarketing features, product information, client information, and social and economic influence factors.

Given that the training data involved a large number of records, a holdout method was used to further divide the training data randomly into training and validation sets to perform 20 runs. Using only the training data, the rminer package of the R tool was used to run four classification models – logistic regression (LR), decision trees (DT), neural network (NN) and support vector machine (SVM) – to select significant variables for testing. Business knowledge and a revised forward selection method were applied to shortlist a final set of 22 variables that were significant in improving the average AUC computed on the validation set. Results in this exploratory experiment showed that the NN model with 22 variables and H = 6 and N = 7 delivered the best AUC and ALIFT values of 0.929 and 0.878 respectively.

3. Base Model Illustration

Four classification models were tested for their predictive capabilities using a rolling window scheme that discards oldest data and performs model updates. Performance during the modelling and evaluation phases was compared using two metrics, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT). For both metrics and phases, the NN model with 22 variables and H = 6 obtained the best results and delivered an AUC of 0.794 and ALIFT of 0.672 during the evaluation phase. Furthermore, the cumulative LIFT analysis showed that 79% of successful sales could be achieved when contacting only half of the customers. A decrease in AUC and ALIFT was observed when comparing the NN model during modelling and evaluation phases, which was expected as the variable selection was tuned based on validation set errors during the modelling phase, but the best model was then fixed and tested on completely unseen data.

Despite accurate predictive performance from the NN model, the model's complexity can make it difficult to understand. Secondary methods such as decision tree and sensitivity analysis are required for knowledge extraction and measurement of predictor importance. In this case, sensitivity analysis of the NN states that the three-month Euribor rate as the most relevant variable.

For this paper, we will select the NN model as our baseline model for comparison.

4. Proposed Model

We propose the Decision Tree (DT) model as the new model. DT are machine learning models are used to solve classification and regression problems, which will be used in this case to predict the successful of telemarketing calls. Decision rules are charted out in a tree structure, which has multiple nodes that represent each input's probability of predicting an outcome (Kim, Cho, & Ryu, 2020). The resulting tree provides insights on the predictor's importance and is more explainable compared to the NN model. DT has managed to perform better than NN (Olson, Delen, & Meng, 2012) and costs lesser processing time without sacrificing its accuracy (Jin & HE, 2019). One issue with the NN model is that the flexibility of the model potentially causes overfitting as the model adjusts its parameters too close to the training data (Salman & Liu, 2019). This issue can be fixed by DT with the use of pruning (Bertsimas & Dunn, 2017)

Thus, using DT, we will be able to produce a model with better explain ability and significantly faster execution speed without sacrificing predictive capability.

5. Data Preparation

The "bank-full.csv" dataset was selected as the dataset to be used. Initial studies of the dataset show that there are 45,211 records. There is a total of 17 fields, 16 independent variables and one target variable. Figure 1. below shows the fields and their individual purpose.

	Variable	Data Type	Purpose
1	age	Continuous	Age of individual
2	job	Nominal	Type of job
3	marital	Nominal	Marital status
4	education	Nominal	Education status
5	default	Flag	Whether individual has credit in default
6	balance	Continuous	Average yearly balance (Euros)
7	housing	Flag	Whether individual has housing loan
8	loan	Flag	Whether individual has personal loan
9	contact	Flag	Communication type
10	month	Nominal	Last contacted month
11	day_of_week	Nominal	Last contact day
12	duration	Continuous	Duration of last contact
13	campaign	Continuous	Number of contacts performed for individual
14	pdays	Continuous	Number of days that passed since last contact
15	previous	Continuous	Number of contacts performed prior this campaign
16	poutcome	Nominal	Outcome of previous marketing campaign
17	y (target)	Flag	Whether individual purchased a term deposit

Figure 1. Overview of the fields present in the dataset.

Field	Measurement	Values	Missing	Check	Role
age	Continuous	[18,95]		None	Input
job	Nominal	admin.,blue-collar,entrepreneur,housemaid,manageme...		None	Input
marital	Nominal	divorced,married,single		None	Input
education	Nominal	primary,secondary,tertiary,unknown		None	Input
default	Flag	yes/no		None	Input
balance	Continuous	[-8019,102127]		None	Input
housing	Flag	yes/no		None	Input
loan	Flag	yes/no		None	Input
contact	Nominal	cellular,telephone,unknown		None	Input
day	Continuous	[1,31]		None	Input
month	Nominal	apr,aug,dec,feb,jan,jul,jun,mar,may,nov,oct,sep		None	Input
duration	Continuous	[0,4918]		None	Input
campaign	Continuous	[1,63]		None	Input
pdays	Continuous	[-1,871]		None	Input
previous	Continuous	[0,275]		None	Input
poutcome	Nominal	failure,other,success,unknown		None	Input
y	Flag	yes/no		None	Target

Figure 2. Overview of the fields present in the dataset on SPSS Modeller

There are three fields with "unknown" values. The "unknown" values will be substituted with blank values and the data type will be changed to flag.

Figure 3. below shows the data quality report after the change in values.

Complete fields (%): 92.35% Complete records (%): 17.39%

Field	Measurement	Outliers	Extremes	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
age	Continuous	379	2	100	45211	0	0	0	0
job	Nominal	--	--	100	45211	0	0	0	0
marital	Nominal	--	--	100	45211	0	0	0	0
education	Nominal	--	--	95.893	43354	0	0	0	1857
default	Flag	--	--	100	45211	0	0	0	0
balance	Continuous	461	284	100	45211	0	0	0	0
housing	Flag	--	--	100	45211	0	0	0	0
loan	Flag	--	--	100	45211	0	0	0	0
contact	Nominal	--	--	71.202	32191	0	0	0	13020
day	Continuous	0	0	100	45211	0	0	0	0
month	Nominal	--	--	100	45211	0	0	0	0
duration	Continuous	763	200	100	45211	0	0	0	0
campaign	Continuous	509	331	100	45211	0	0	0	0
pdays	Continuous	1644	79	100	45211	0	0	0	0
previous	Continuous	397	185	100	45211	0	0	0	0
poutcome	Nominal	--	--	18.252	8252	0	0	0	36959
y	Flag	--	--	100	45211	0	0	0	0

Figure 3. Summary of the data quality on SPSS Modeller

For both fields "contact" and "education", the incomplete fields will be removed. For "poutcome", since there are only 8,252 valid records, which is only around 18.252% of the dataset, the field will be removed. The dataset is cleaned using the data audit node in IBM SPSS Modeler after removing the

"poutcome" field using the filter node. Figure 4. below shows the quality report after. The number of records in the final dataset is 30,907.

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
age	Continuous	297	0	100	30907	0	0	0	0
job	Nominal	--	--	100	30907	0	0	0	0
marital	Nominal	--	--	100	30907	0	0	0	0
education	Nominal	--	--	100	30907	0	0	0	0
default	Flag	--	--	100	30907	0	0	0	0
balance	Continuous	309	189	100	30907	0	0	0	0
housing	Flag	--	--	100	30907	0	0	0	0
loan	Flag	--	--	100	30907	0	0	0	0
contact	Nominal	--	--	100	30907	0	0	0	0
day	Continuous	0	0	100	30907	0	0	0	0
month	Nominal	--	--	100	30907	0	0	0	0
duration	Continuous	531	135	100	30907	0	0	0	0
campaign	Continuous	391	231	100	30907	0	0	0	0
pdays	Continuous	168	27	100	30907	0	0	0	0
previous	Continuous	305	125	100	30907	0	0	0	0
y	Flag	--	--	100	30907	0	0	0	0

Figure 4. Summary of the data quality on SPSS modeller after data cleaning

A distribution node is used to study the distribution of the target variable "y" as shown in Figure 5. below.

Value	Proportion	%	Count
no		85.4	26394
yes		14.6	4513

Figure 5. Distribution of the target variable "y" from the distribution node

From Figure 5., it can be noted that the data is imbalanced. Thus, the dataset requires either undersampling or oversampling to balance the dataset. Overfitting refers to the process of replicating or synthetically creating more records of the minority class to alleviate the imbalance problem. There will be no loss of data as more data is created. However, increased dataset will increase the execution time and potentially cause overfitting problems (Kaur & Gosain, 2018). Undersampling does the opposite, records from the majority class are removed randomly to balance the classes. This results in the loss of data from the majority class and the overall amount of data will be reduced. As we focus on the prediction of the minority class, undersampling will be used. The proportion of the minority class is 14.6%, which is significantly smaller than the majority. If oversampling is used, there will be an addition of approximately 20,000 minority records which will potentially cause overfitting and increase the execution costs. Thus, undersampling is selected for this model. The undersampled dataset will have 9,004 records, which is sufficient for model training. Figure 6. shows the distribution after undersampling.

Value	Proportion	%	Count
no		50.22	4553
yes		49.78	4513

Figure 6. Distribution of the target variable "y" from the distribution node after undersampling

The dataset will be partitioned 80% for training dataset and 20% for test dataset.

6. Predictive models on SPSS Modeller

Three DT algorithms will be used to train the data and the different models will be compared, with the best model being the proposed model. The algorithms used are Classification and Regression Tree (CART), Chi-squared automatic interaction (CHAID) and C5.0.

For CART, the stopping rule used is shown in Figure 7. below. The maximum tree depth is set as five, with the pruning function activated with the maximum surrogates of five. The rest of the settings are kept as default.

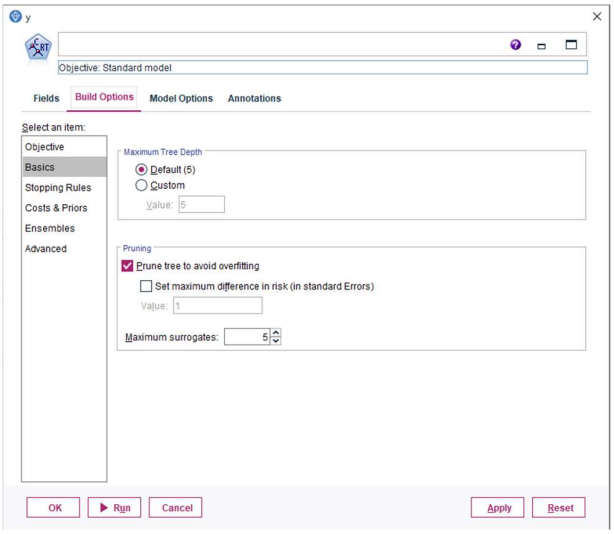

Figure 7. Build options of the CART model on SPSS Modeller

For CHAID, the stopping rule used is shown in Figure 8. below. The rest of the settings are kept at default.

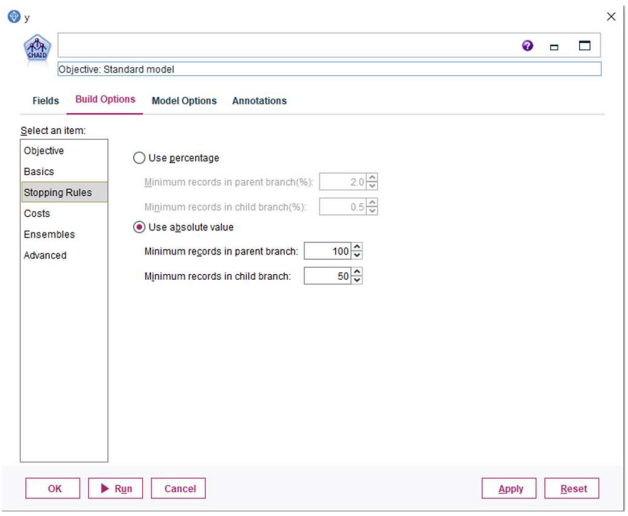

Figure 8. Build options of the CHAID model on SPSS Modeller

For CART and CHAID, the stopping rule was changed to absolute value from percentage as it provided better accuracy.

The default settings for C5.0 produced a huge decision tree with eight levels of data. Post-pruning is required to prevent overfitting and to produce a DT that is more usable (Dinov, 2020). The C5.0 node is set on expert mode, with cross-validation set as 10, pruning severity at 90 and minimum records per child branch is set as 100 as shown in Figure 9. below.

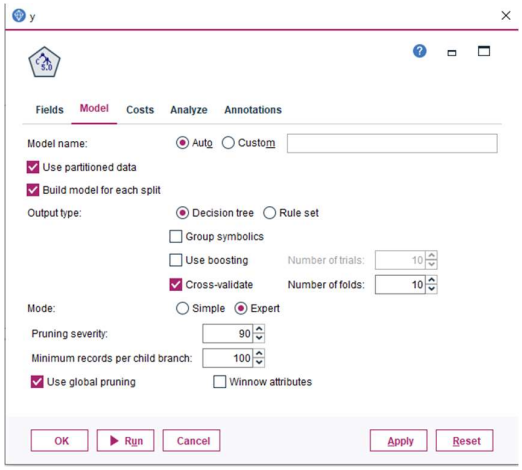

Figure 9. Build options of the CHAID model on SPSS Modeller

For the baseline comparison model, a NN with default settings is used with the dataset and partition settings.

7. Modelling Results and Discussion

The Area under ROC curve (AUC), accuracy, hit rate and the lift charts will be used to compare the models. The AUC measures the overall performance of a binary classifier and it involves all possible classification models thresholds (Melo, 2013). The AUC ranges from 0 to 1. When the AUC is 1, it indicates that there is a perfect prediction by the model. An AUC of 0.5 or less means that the model is unable to discriminate and has poor accuracy (Lan, 2019).

	AUC (Training)	AUC (Test)
CART	0.857	0.856
CHAID	0.877	0.887
C5.0	0.809	0.797
ANN	0.876	0.883

Figure 10. Summary of the AUC results (Training and Testing) of all predictive models

Figure 10. above shows the respective AUC of all models. The CHAID model has the highest AUC, higher than the NN model.

The accuracy and hit rate of the test data is used as another comparison factor. Four confusion matrices are created by applying the test dataset on all models. Figure 11. below shows the respective accuracy and hit rates of each model. The full confusion matrices can be found in Appendix A, B, C, D.

	CART	CHAID	C5.0	NN
Overall Accuracy	80.68%	80.68%	78.76%	80.51%
Accuracy for "yes"	81.42%	84.75%	74.08%	83.60%
Accuracy for "no"	80.00%	76.96%	83.04%	77.70%
Hit Rate for "yes"	78.80%	77.06%	79.95%	77.39%
Hit Rate for "no"	82.51%	84.68%	77.82%	83.84%

Figure 11. Summary of the accuracy and hit rate of all predictive models

From Figure 11. the CHAID model appears to perform the best on the test dataset, with the highest overall accuracy, "yes" accuracy and "no" hit rate. All three DT performed better than the NN model in this aspect.

Lastly, the lift measures the effectiveness of a predictive model by calculating the ratio between from the results with and without the predictive model (Glen, 2019). Thus, the lift value can show how good the hit rate is by comparing with a random hit rate. The lift charts are shown as a cumulative form. Additionally, when the lift is further up from the baseline, it indicates that the model is better.

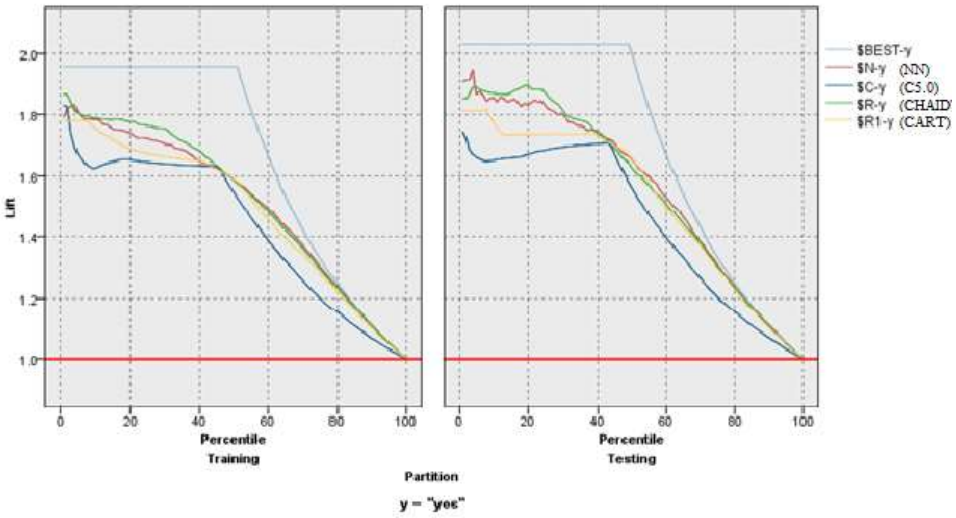

Figure 12. Lift charts of all predictive models

Figure 12. shows the lift chart of all models. The CHAID model performed consistently better in the training dataset. In the testing dataset, NN's lift started off better before dropping below CHAID. Thus, CHAID is the best model based on the lift chart.

Of all the models, the CHAID DT performed consistently better than NN in all comparison methods, thus the CHAID DT model is the best model.

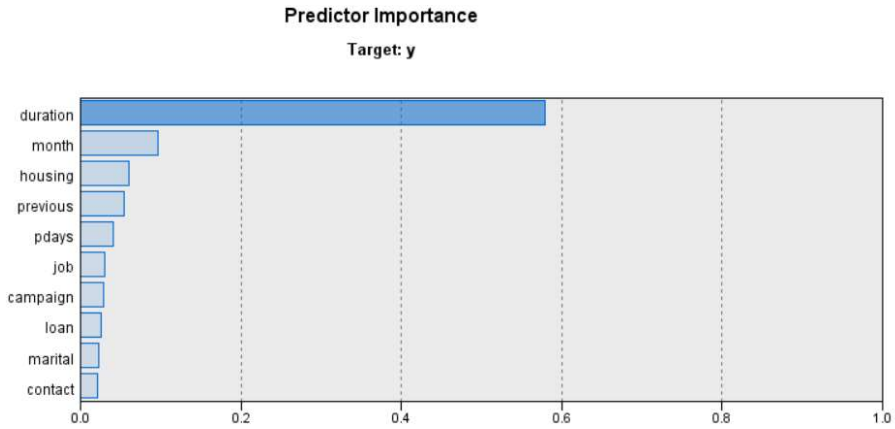

Figure 13. Predictor importance of the CHAID model

Figure 13. above shows the predictor importance of the model. The duration of call is the most important, followed by month and whether the client has housing loan. The decision tree created can be found in Appendix F. Figure 14. shows six rules that were extracted from the decision tree.

Node 26	For call duration of 202s to 252s, when contacted in December, January, March, Oct or September, client is 87.805% more likely to purchase
Node 33	For call duration of 410s to 561s, clients with no housing loan are 78.876% more likely to purchase
Node 56	For call duration of 202s to 252s, when contacted in April, August, February and is working as a blue-collar, housemaid or technician, clients are 73.267% least likely to purchase
Node 57	For call duration of 202s to 252s, when contacted in January, July, May, or November and has no previous contact (cold call), clients are 83.425% least likely to purchase
Node 83	For call duration of 202s to 252s, when contacted in January, July, May, or November and has previous contact (not cold call), and has average yearly balance of more than 924 euros, are 72% more likely to buy
Node 86	For call duration of 252s to 410s, when contacted in April, February, or January, has housing loan, and has average yearly balance of less than 924 euros are 72.603% more likely to buy

Figure 14. Six rules from the decision tree

From the rules, clients with no housing loan, and has previous contact, with higher yearly balance are more likely to purchase the term deposit. While occupations such as blue-collar, housemaid or technicians are least likely to buy. The bank can consider looking into clients that match these descriptions to better improve their efficiency.

8. Limitations and Conclusion

DT model performed better than NN model with better explainability. The success is heavily influenced by the duration spent on call, possibly due to time spent enquiring about the products. Thus, "duration" might be a redundant input and can be removed. Another limitation is the lack of "yes" data that cause

the loss of "no" data due to undersampling. If possible, more "yes" data should be collected to better improve the model's accuracy.

9. References

Bertsimas, D., & Dunn, J. (2017). Optimal classification trees. *Mach Learn 106*, 1039-1082. Retrieved from https://doi.org/10.1007/s10994-017-5633-9

Dinov, I. (2020, September). Data science and predictive analytics (UMich HS650). Retrieved February 22, 2021, from https://www.socr.umich.edu/people/dinov/courses/DSPA_notes/08_DecisionTreeClass.html

Glen, S. (2019, July 23). Gain and lift chart: Definition, example. Retrieved February 22, 2021, from https://www.statisticshowto.com/lift-chart-gain/

Jin, W., & He, Y. (2019). Three data mining models to predict bank telemarketing. IOP Conference Series: Materials Science and Engineering.

Kaur, P., & Gosain, A. (2018). Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise. doi:10.1007/978-981-10-6602-3_3

Lans, D. (2019, July 1). Illustrating predictive models with the ROC curve. Retrieved February 22, 2021, from https://towardsdatascience.com/illustrating-predictive-models-with-the-roc-curve-67e7b3aa8914

Melo, F. (2013). Area under the ROC Curve. Retrieved from https://doi.org/10.1007/978-1-4419-9863-7_209

Salman, S., & Liu, X. (2019). *Overfitting Mechanism and Avoidance in Deep Neural Networks*. Retrieved from https://arxiv.org/pdf/1901.06566.pdf

Kim, H., Cho, H., & Ryu, D. (2020). Corporate Default Predictions Using Machine Learning: Literature Review. Sustainability 12(16), 6325.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, Volume 52, Issue 2, 464-473.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.

10. Appendix A

Confusion Matrix for Neural Network Model

Results for output field y

Individual Models

Comparing $N-y with y

'Partition'	1_Training		2_Testing	
Correct	5,883	80.52%	1,471	80.51%
Wrong	1,423	19.48%	356	19.49%
Total	7,306		1,827	

Coincidence Matrix for $N-y (rows show actuals)

'Partition' = 1_Training	no	yes
no	2,818	847
yes	576	3,065
'Partition' = 2_Testing	**no**	**yes**
no	742	213
yes	143	729

Confusion Matrix for C5.0 Model

Comparing $C-y with y

'Partition'	1_Training		2_Testing	
Correct	5,794	79.3%	1,439	78.76%
Wrong	1,512	20.7%	388	21.24%
Total	7,306		1,827	

Coincidence Matrix for $C-y (rows show actuals)

'Partition' = 1_Training	no	yes
no	3,077	588
yes	924	2,717
'Partition' = 2_Testing	**no**	**yes**
no	793	162
yes	226	646

Confusion Matrix for CHAID Model

Comparing $R-y with y

'Partition'	1_Training		2_Testing	
Correct	5,871	80.36%	1,474	80.68%
Wrong	1,435	19.64%	353	19.32%
Total	7,306		1,827	

Coincidence Matrix for $R-y (rows show actuals)

'Partition' = 1_Training	no	yes
no	2,783	882
yes	553	3,088
'Partition' = 2_Testing	**no**	**yes**
no	735	220
yes	133	739

Confusion Matrix for CART Model

Comparing $R1-y with y

'Partition'	1_Training		2_Testing	
Correct	5,864	80.26%	1,474	80.68%
Wrong	1,442	19.74%	353	19.32%
Total	7,306		1,827	

Coincidence Matrix for $R1-y (rows show actuals)

'Partition' = 1_Training	no	yes
no	2,926	739
yes	703	2,938

'Partition' = 2_Testing	no	yes
no	764	191
yes	162	710

Evaluation Metrics of all predictive models

Evaluation Metrics

'Partition'	1_Training		2_Testing	
Model	AUC	Gini	AUC	Gini
$N-y	0.876	0.753	0.883	0.765
$C-y	0.809	0.618	0.797	0.595
$R-y	0.877	0.755	0.887	0.774
$R1-y	0.857	0.715	0.856	0.712

CHAID Decision Tree