Team -4

Punya Swaroop Sirigiri

Vibhor Kaushik

Executive Summary

Analyzing Marketing Trends of a Clothing Store

Team 4

OPIM5604 Predictive Modeling

Contents

The work presented is our team's work and our team's work alone

EXCECUTIVE SUMMARY

Abstract

The primary purpose of this document is to help the marketing management of a clothing store in predicting which customers are most likely to respond to direct mail marketing promotions. It also provides an estimate to the management on reducing the cost of mailing promotions to the customers who may not respond to the mails.

Background

The clothing store chain in New England, who provided the customer data wants to **identify the potential customers who would respond to the direct mail marketing.**

In order to help the clothing store chain to identify the potential customers, we built predictive models using various modeling techniques like Logistic Regression (LR), Decision Tree (DT) Analysis, Neural Network (NN), Discriminant Analysis (DA) and Ensemble Modeling (Refer Methodology Section of this report for detailed explanation). The models were built using the variables from the data set on a direct mail marketing campaign conducted in the previous year. We used this data to develop best performing prediction model for future marketing campaigns.

For the marketing management team of clothing store, the primary **objective is to decrease the cost of mailing to unresponsive customers**. Prediction models are evaluated based on their accuracy rates, error rates, false negative rates, and false positive rates by comparing to the actual results in the data set. In most marketing problems, a false negative decision error is worse. **False negative in this case means that we failed to contact a customer who would have responded positively to the promotion.** Had this customer been contacted, he or she would have responded, adding revenue to the clothing store. False Negative is very expensive because the company losses the opportunity to earn the revenue from those customers, and modelers should endeavor to minimize the probability of this error.

The methodology section of this report discusses the general SEMMA (*Sampling, Exploration, Modification, Modeling and Assessment*) approach the team followed, the challenges it faced and decisions made to meet our principal objective. In brief, team studied the data, redefined the data types, identified insubstantial variables (based on high correlation, encrypted variables etc.) and critical variables which are used to build models for predicting customer response to the direct mail marketing promotion.

Finally, team chose **28 variables** among 50 to build models with a data set of **21,740 observations.** All the steps are described in detail, with screenshots in the appendix.

The details of the best performing models are summarized in the below graphs using training and validation data.

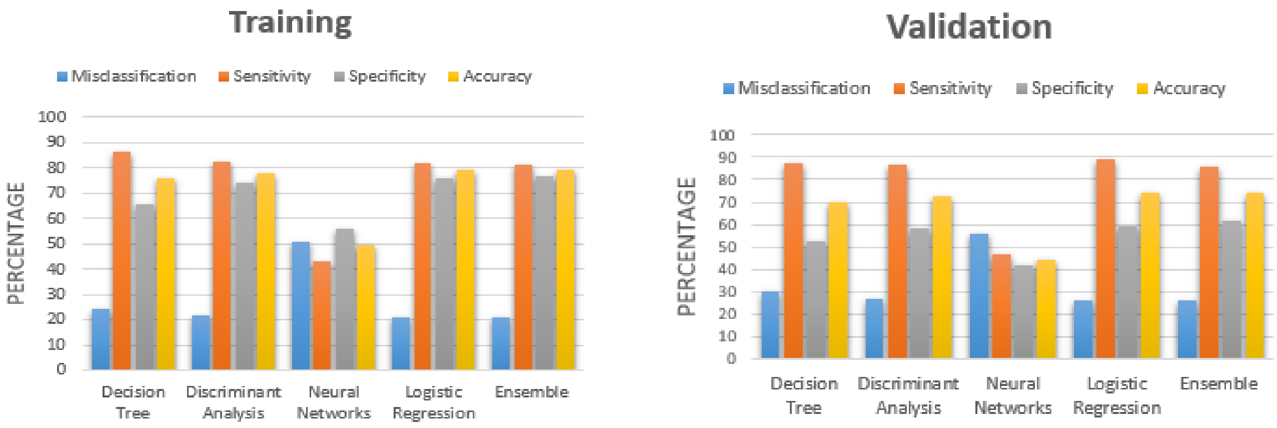


Figure 1: Comparison of various Model Performances

From the above graphs we can see that the Logistic Regression and Ensemble models outperform other models. These two models have a relatively low value of misclassification rate with a reasonably high sensitivity and specificity values (Refer Appendix for definitions). Owing to simplicity and easier understanding of Logistic Regression Model, team chose to go ahead with Logistic Regression Model over Ensemble Model.

Deduction:

The decision outcomes of the Logistic Regression model when tested on real-world data set, are as shown in Table 1.1

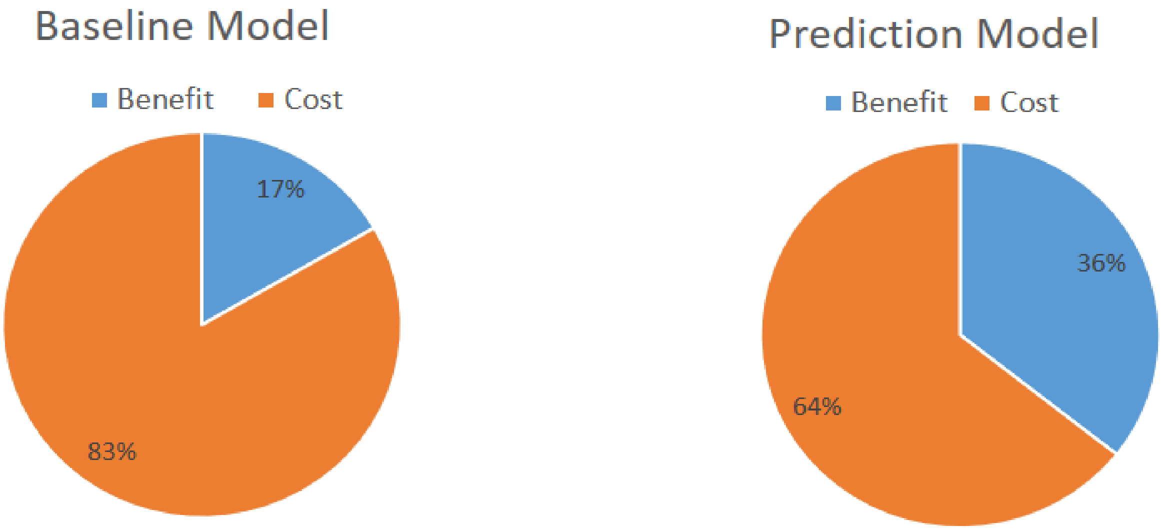
Confusion Matrix		**Predicted**	
		0	1
Actual	0	True Negative	False Positive
		LR Model = 7359	LR Model = 2737
	1	False Negative	True Positive
		LR Model = 495	LR Model = 1516

Table 1.1 Confusion Matrix of Prediction Model

Comparing the Predection Model with the baseline (that is, send mail to everyone) model,

Baseline model: Benefit of sending the promotion to 2011customers who would respond *minus* Cost of sending the promotion to 10096 customers who would not respond

Prediction Model: Benefit of sending the promotion to 1516 customers who would respond *minus* Cost of sending the promotion to 2737 customers who would not respond.



Pie Charts of Baseline and Prediction Models

Cost of mailing to unresponsive customers decreases by 18% while there is a 4% loss of oppurtunities with potential customers who may have responded if mailed. For example, let us assume that cost of mailing each customer is $ 2, then without any model if the management decides to send mails to all the 12,107 customers, then total cost spent of mailing would be $ 24,214 and only 16% of them respond to it. Where as with the suggested model, management would send mails only to 4,253 of the total customers with a response rate of 35.65% thus saving a mailing cost of $ 15,708. But the team also warns about losing oppurtunities with 4% of potential customers who might have responded if mailed. Marketing Management should consider this factor before making a final decision of mailing promotions to customers.

METHODOLOGY

The team has used the SEMMA approach to clean and remove inconsistencies in the data which could have led to inaccurate analysis thereby undermining the whole effort of this analytics project. This report covers the elements of the SEMMA approach:

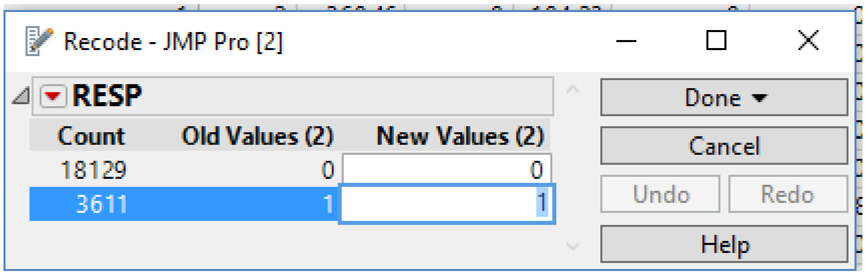
- **Sampling:** The dataset used for analysis must be large enough to contain the significant information but small enough to process[1]. The data set provided by the Clothing Store is large enough to do analysis with. As a result, team agreed to take all the observations and further divide the data set into Training, Validation and Testing for modeling.
- **Exploration:** This phase consists of searching for anticipated relationships, unanticipated trends, and anomalies to gain understanding and ideas[2]. Here, the team categorized the various variables (continuous, nominal, ordinal), looked for inconsistencies, missing values and outliers. Team also classified variables as potential predictors for the target variables by observing their distribution graphs and correlation matrixes.
- **Modification:** This phase consists of creating, selecting, transforming the variables to focus on the model selection process[3]. The team performed the necessary steps to remove the inconsistencies and issues of missing data and extreme values. The team performed Transformation, Standardization to handle these outliers in variables. The outcomes of the modification process are described in the methodology section of this document.
- **Modeling and Assessment:** This phase consists of modeling the data by using analytical tools to search for a combination of the data that reliably predicts a desired outcome. Assess competing predictive models by building charts to evaluate the usefulness and reliability of the findings from the data mining process[4]. Team built prediction models using Classification models in SAS JMP tool, with potential predicting variables identified in Exploration phase. The process of building the model is described in the methodology section of this document. Screenshots and detailed descriptions can be found in the appendix.

[1] From Course Materials: "Updated Intro to Predictive Modeling, Slide 31"
[2] From Course Materials: "Updated Intro to Predictive Modeling, Slide 31"
[3] From Course Materials: "Updated Intro to Predictive Modeling, Slide 31"
[4] From Course Materials: "Updated Intro to Predictive Modeling, Slide 31"

Data Preparation

As part of the Data Preparation, there were issues such as missing values and outliers which need to be handled. It is really important that the data is cleaned in the initial phase as it does effect the further modeling process and lead to wrong statistical inferences.

1. **Extract the Data:** The Clothing-Store data set was obtained from book series *Website*.[5]
2. **Initial Exploratory Analysis:** The raw data consisted of **21,741 observations with 51 variables**. The data collected was related to the clothing store based on a *direct marketing campaign* conducted the previous year. Initial Analysis of the data showed that only 3611 out of 21,741 customers, or 16.60%, responded to last's year campaign as shown in below Figure 2. Customers response is re-coded in terms of:
    - **1** - Indicating Response
    - **0** - Indicating No Response



Recode - JMP Pro [2]

RESP

Count	Old Values (2)	New Values (2)
18129	0	0
3611	1	1

Figure 2: Customer Resonse to direct marketing campaign conducted last year

3. **Data Modeling Types:** Out of 51 variables, 46 variables were considered as continuous variables, because they are numeric and have distance measures that allow to perform calculations on them (or they were calculated themselves). The target variable(***RESP***), the Flag Varibles (***CC_CARD, WEB, VALPHON***) are considered to be nominal as they have values of either "0" or "1". The variable **CLUSTYPE** represents Microvision Lifestyle cluster types[6] is also considered to be nominal as it has values between 1 and 50 and is not a measure.
4. **Encrypted Values:** Customer ID (***HHKEY***) is unique to every customer and is encrypted; Collectible Lines(**PCOLLSPND**) depicts the brand of choice and is encrypted, thus contains information that is not helpful to predict customers who respond to the direct mail marketing promotion. These variables are hence omitted in further analysis.

[5] http://dataminingconsultant.com/

[6] MicroVision is a revolutionary micro-geographic consumer targeting system created by Claritas Inc. It was designed and developed to create homogeneous segments which display different lifestyles and purchasing behaviors according to 1990 census data.

5. **Missing Values:** Only one out of 21,741 observations is missing in the given data set. Apart from this, the data set seems pure. We did not see any redundancies and inconsistencies in data. Since the missing data is very small and negligible(less than 1%).It has no critical functionality in the business objective, the team decided to remove this observation and go ahead further.
6. **Duplicate Observations:** Initial Exploratory Analysis showed that no duplicates were present in the data set.
7. **Data Redundancy:** After the initial analysis of the data**,** we concluded that given data is free from redundant values and hence no imputing/editing of the data is done.
8. **Transformation:** During the initial analysis of the data, we observed highly skewed data in some of the variables. It is advised not to work on skewed data in its raw form as it reduces the impact of low frequency values when calculating distances in modeling. At times skewness is influenced by presence of Outliers. To deal with this, we chose to transform the continuous variables. Transformation helped in making the variables normalised as shown in below Figure 3. We performed transformation for the 6 continuous variables, namely *Total Money Spent (MON), Amount Spend at Franchise CC(CCSPEND), Gross Margin Percentage (GMP), Number of Days between purchases (FREDAYS), Number of different product classes purchased(CLASSES), Lifetime Average time between visits(LTFREDAY).* Transformation helped us to deal with the data which is normal distributed.

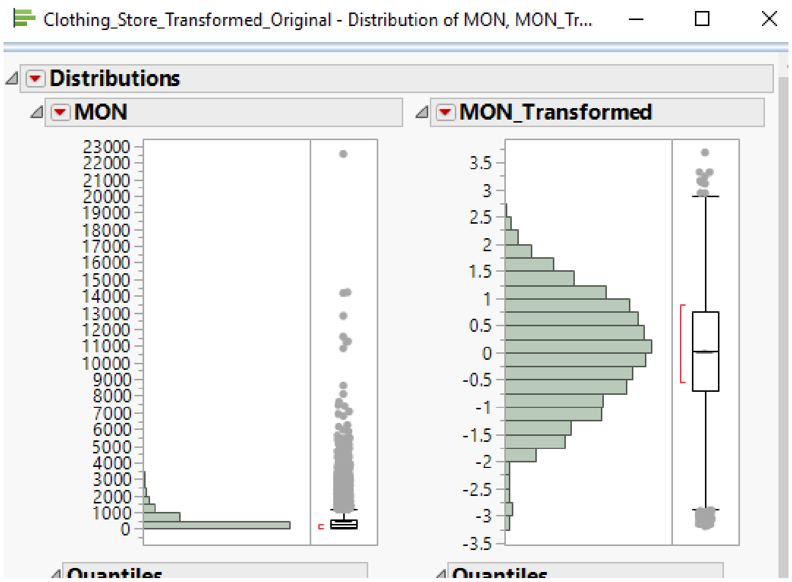
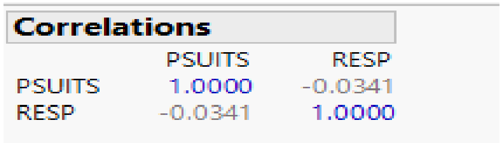


Figure 3: Transformation of highly skewed data gives a normal distribution curve

9. **Standardisation:** After transforming the data, standardisation of variables helped us to bring all the transformed variables onto same scale, making it advantageous if the model uses distance formula for prediction.
10. **Principal Component Analysis (PCA):** Having a large number of the data variables (27), we tried to perform the Principal Component Analysis, to reduce and check the dimensionality of the variables. PCA resulted in linear combination of 19 principal components which accounted 95 % of variance in the original data (Refer appendix for more details). Since it is difficult to explain the model to business in terms of principal components, the team chose not to use the PCA while building the model .

Modeling Phase

After the initial analysis of the data, we tried to reduce the number of the data variables by finding the correlation between them. Additionally, we considered the correlation between the input variables and the target variable (RESP) to get a better understanding of the data. Based on our analysis, we removed **24 variables**, where some of them had high correlation to the other independent variables used in our model and others had very less correlation with the target variable as shown in below Figure 4.(Refer to Appendix for more details).



Correlations

	PSUITS	RESP
PSUITS	1.0000	-0.0341
RESP	-0.0341	1.0000

Figure 4: Low correlation between the Target Variable(RESP) and input Variable (PSUITS)

Since the proportion of customers responded is very small, we decided to balance the data prior to modeling . We splited 7.5% of the total data as the training data(balanced), 7.5% as the validation data(balanced) and 55% as the testing data(unbalanced or real-world) using the **stratified sampling technique**. Thus we could use only 70% of the data in the given data set to Train, Validate and Test the model (Refer to Appendix for more details).

Training Phase

Training data set was used to build and train the predictive model. Various Classification models such as - Decision Trees, Logistic Regression, Neural Networks, Discriminant Analysis have been implemented to develop the best model which predicts whether customer responds to Direct Mail Marketing Promotion or not.

Partition using decision tree method with different combinations, maximum branches and maximum depth were used. Nominal Logistic Regression, derived from the Nominal target variable, was built. Neural Networks was built using the multilayer Perceptron architecture. Discriminant Analysis model was built using the discriminant function. Model performance on Training Data Set is tabulated in Tabel 1.1 *Comparision of Models.*

Ensemble Model

Ensemble modeling is the process of modeling the results of two or more analytical models and then synthesizing those results to improve the accuracy. After building the models and validating them, we built Ensemble model using the actual response variable(*RESP)* as Target variable and the Most Likely(or Probable) Responses from each of the model as Predictors. Since Logistic Regression model relatively outperformed other models in training and validation phases, we chose to use Logistic Regression to build the Ensemble Model.(Refer to Appendix for more details)

Validation Phase

We validated all the five trained predictve models using Validation Data Set. The validation data set was also used for model fine-tuning and to check the over-fitting of models. Contingency plot provided the confusion matrix for each model respectvely which is used to evaluate the performance of each model on validation data set. Results in Table 1.3. The team tried to minimize the **False Negatives in our confusion matrix.** (According to our Business Objective, False negative means that we failed to contact a customer who would have responded to the promotion).

	Misclassification		**Sensitivity**		**Specificity**		**Accuracy**	
Model	**Training**	**Validation**	**Training**	**Validation**	**Training**	**Validation**	**Training**	**Validation**
Decision Tree	24.07	29.87	86.25	87.63	65.62	52.63	75.93	70.13
Discriminant Analysis	21.82	27.25	82.62	87	73.75	58.5	78.18	72.75
Neural Networks	50.75	55.69	42.87	47	55.62	41.63	49.25	44.31
Logistic Regression	**21.07**	**26.06**	**81.75**	**88.75**	**76.12**	**59.13**	**78.93**	**73.94**
Ensemble	21.06	26.19	81.37	86	76.5	61.5	78.94	73.81

Table1.3: Comparison of various Model Performances

From the above Table we see that the Logistic Regression and Ensemble models outperform other models. These models have a relatively low value of misclassification rate with a reasonably high sensitivity and specificity values (Refer Appendix for definitions). Owing to simplicity and easier understanding of Logistic Regression Model, team chose to go ahead with Logistic Regression Model over Ensemble Model.

Testing Phase

After we trained the models and validated its performance, we tested the Logistic Regression model with Real-world or Unbalanced Data test. The decision outcomes of the Logistic Regression model when tested on real-world data set, is shown in Table 1.3. As discussed earlier, threshold value for False Negative should be as low as possible in our business setting, the model predicted that 35% of the customers would most likely respond to the mails, with 4% of customers not being contacted who might have responded to the promotion mails.

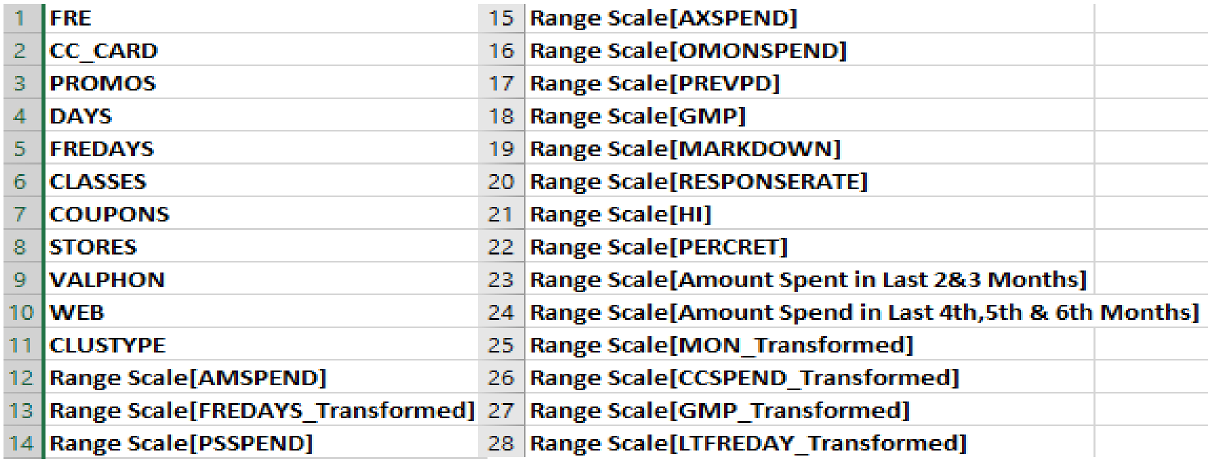
We observed that Accuracy of the Nominal Logistic Regression Model on the Testing Data is 73.3%

Confusion Matrix		**Predicted**	
		0	1
Actual	0	True Negative	False Positive
		LR Model = 7359	LR Model = 2737
	1	False Negative	True Positive
		LR Model = 495	LR Model = 1516

Table 1.3 Confusion Matrix *(Contingency Tabel for Logistic Regression)*

Important Findings:

1. The Nominal Logistic Regression Model relatively outperformed other models in the validation phase. Consequently, for predicting the Response or Target variable, the teamchose logistic regression model on the Testing Data set(real time data).
2. The final Logistic Regression Model has the following significant predictors:



1	FRE	15	Range Scale[AXSPEND]
2	CC_CARD	16	Range Scale[OMONSPEND]
3	PROMOS	17	Range Scale[PREVPD]
4	DAYS	18	Range Scale[GMP]
5	FREDAYS	19	Range Scale[MARKDOWN]
6	CLASSES	20	Range Scale[RESPONSERATE]
7	COUPONS	21	Range Scale[HI]
8	STORES	22	Range Scale[PERCRET]
9	VALPHON	23	Range Scale[Amount Spent in Last 2&3 Months]
10	WEB	24	Range Scale[Amount Spend in Last 4th,5th & 6th Months]
11	CLUSTYPE	25	Range Scale[MON_Transformed]
12	Range Scale[AMSPEND]	26	Range Scale[CCSPEND_Transformed]
13	Range Scale[FREDAYS_Transformed]	27	Range Scale[GMP_Transformed]
14	Range Scale[PSSPEND]	28	Range Scale[LTFREDAY_Transformed]

Figure 4: List of Predictors

3. Comparing our model with baseline model (that is, Send mail to everyone):

   Let us assume that the cost of mailing each customer is $2, if the management decides to send out mails to all the 12,107 customers (based on testing data set), total cost spent towards mailing would be $ 24,214.00 and only 16% of them respond to it without any model. Where as with the model the team suggested, management would send mails only to 4,253 of the total customers with a response rate of 35.65% thus saving a mailing cost of $ 15,708.00. But the team also warns about losing oppurtunities with 4% of potential customers who might have responded if mailed. Marketing Management should consider this factor before making a final decision to mail promotions.
4. Further analysis of the Confusion Matrix (Contingency Table) helped in finding the customers who would respond to the direct mail marketing promotion, thus decreasing the cost of mailing for the company.

CONCLUSION, RECOMMENDATIONS, KEY TAKE AWAY

- Today marketers are trying to get most insight from their data for effective and efficient marketing campaigns with minimizing the cost factor.Our model is able to increase the response rate for direct mail marketing promotion from 16.6% to 35.5% thereby reducing overall mailing cost.
- Disclaimer here is that, since we are reduing the overall customer base of mailing the marketing promotions, there could be a loss of 4% potentail customers who may have responded to the marketing campaign. Considering the recommendation submitted to management, final call for marketing campaign implementaion have to be taken by higher management.
- While working for this project, we were very clear not only to find the quantifiable factors which lead to improvement of the customer response rate but at the same time we tried to consider factors(or variables) which could decrease the loss incurred due to mailing the non-responsive customer.
- To get high probability of the response, we tried different combinations of the variables and then, these predictors (variables) were used for building different predictive models. For almost every decision, different perspectives were discussed which assisted us in appreciating and understanding the diversity of choices and the variation of the solution accordingly.

We completely understood why the answer is mostly “It always depends!” ☺

References:

- Clothing-Store data set was obtained from book series website http://dataminingconsultant.com
- Defination of Microvision Lifestyle www.tetrad.com/pub/prices/microvision.pdf

Appendix

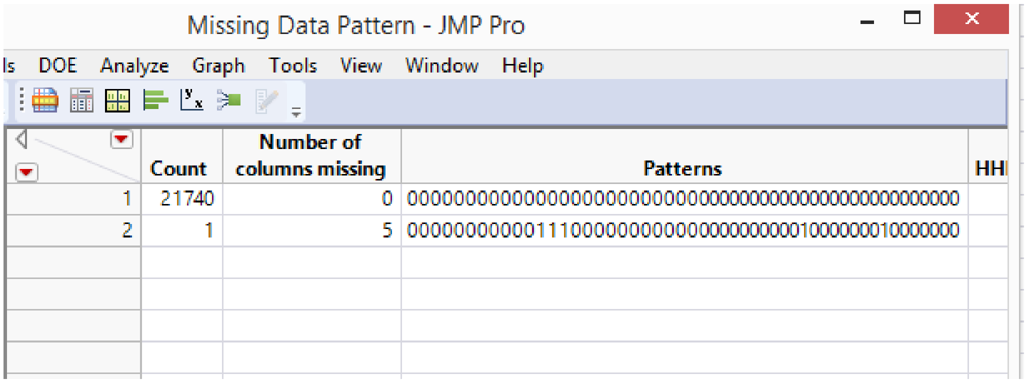
Data Pre Processing

Missing Data Description

We encountered five variables like HI, STORELOY, REC for which details are missing. That is, the data dictionary didn't have any details about these variables. So on these variables we performed correlation to the target variable and observed that only one ('HI') out of five variables has high correlation. Based on this observation, the team have chosen this one variable for modelling.

Missing Data:

The team has explored the data in all possible ways and found that only one observation has missing values in the data set. Apart from this, the data set seems pure. We did not see any redundancies and inconsistencies in data. Since the missing data is very small and negligible and have no critical functionality in the business objective, the team decided to remove this observation and go ahead further.


Missing Data Pattern - JMP Pro
ls DOE Analyze Graph Tools View Window Help

	Count	Number of columns missing	Patterns	HH
1	21740	0	000000000000000000000000000000000000000000000000000	
2	1	5	000000000001110000000000000000000001000000010000000	

Derived Variables:

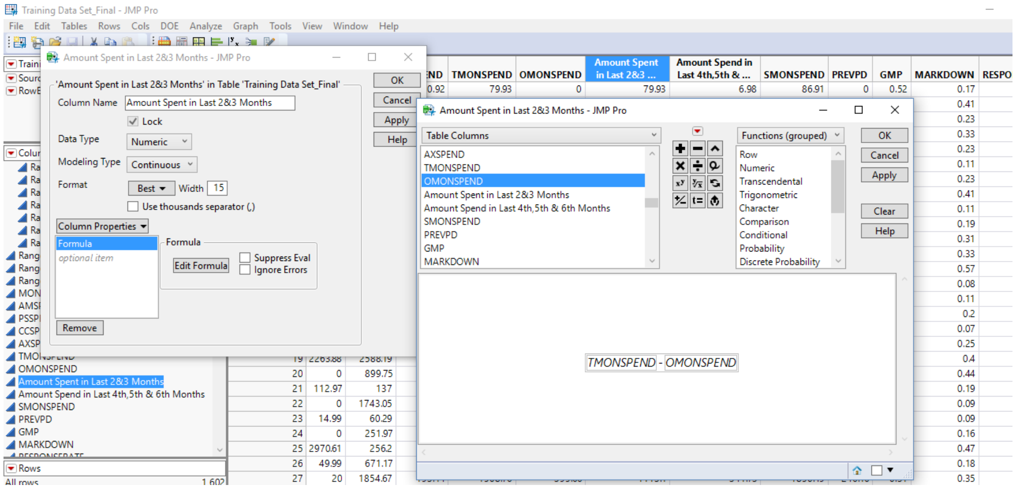
After going through the data dictionary, the team has decided to calculate Amount Spent in 2nd and 3rd month using variables OMONSPEND- 'Amount Spent in Last One Month' and TMONSPEND- 'Amount Spent in last Three Months'.

That is, *Amount Spent in 2nd and 3rd month* = TMONSPEND-OMONSPEND

Similarly, team has calculated amount spent in last 4th, 5th and 6th months using below formula:

Amount Spent in Last 4th, 5th & 6th month = SMONSPEND-TMONSPEND.

Team thought that having the data month wise brings more meaning compared to quarterly or half yearly. So we derived these two variables and used them for building our model. Shown below the JMP screenshot of the same.



'Amount Spent in Last 2&3 Months' in Table 'Training Data Set_Final'

Column Name: Amount Spent in Last 2&3 Months

☑ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best Width 15

□ Use thousands separator (,)

Column Properties: Formula

Formula: □ Suppress Eval □ Ignore Errors

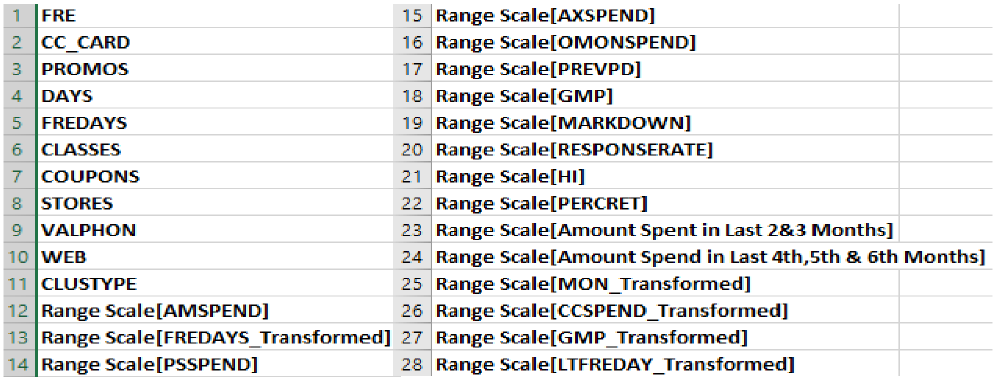
Table Columns: AXSPEND, TMONSPEND, OMONSPEND, Amount Spent in Last 2&3 Months, Amount Spend in Last 4th,5th & 6th Months, SMONSPEND, PREVPD, GMP, MARKDOWN

Formula: TMONSPEND - OMONSPEND

END	TMONSPEND	OMONSPEND	Amount Spent in Last 2&3 ...	Amount Spend in Last 4th,5th & ...	SMONSPEND	PREVPD	GMP	MARKDOWN
0.92	79.93	0	79.93	6.98	86.91	0	0.52	0.17

Selecting Variables:

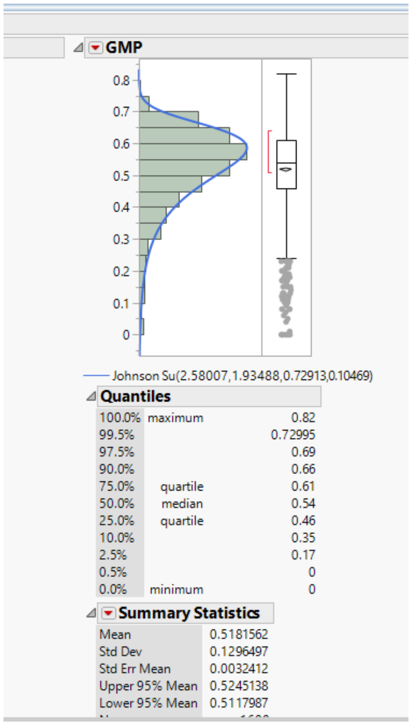
Team has identified a business problem of how to decrease the cost of mailing promotions to the unresponsive customers. For this the team has explored the data and filtered out few variables which are either encrypted or not seemed appropriate to the scope. The team has decided to go ahead with below listed 28 variables to build a model after filtering out 24 variables.



1	FRE	15	Range Scale[AXSPEND]
2	CC_CARD	16	Range Scale[OMONSPEND]
3	PROMOS	17	Range Scale[PREVPD]
4	DAYS	18	Range Scale[GMP]
5	FREDAYS	19	Range Scale[MARKDOWN]
6	CLASSES	20	Range Scale[RESPONSERATE]
7	COUPONS	21	Range Scale[HI]
8	STORES	22	Range Scale[PERCRET]
9	VALPHON	23	Range Scale[Amount Spent in Last 2&3 Months]
10	WEB	24	Range Scale[Amount Spend in Last 4th,5th & 6th Months]
11	CLUSTYPE	25	Range Scale[MON_Transformed]
12	Range Scale[AMSPEND]	26	Range Scale[CCSPEND_Transformed]
13	Range Scale[FREDAYS_Transformed]	27	Range Scale[GMP_Transformed]
14	Range Scale[PSSPEND]	28	Range Scale[LTFREDAY_Transformed]

Transformation:

Initially, the team tried transforming all the continuous variables which are not normally distributed. Then we observed that for only few transformed variables the data is uniformly distributed. The team reverted back the transformation for other variables which didn't show the normal distribution. Only 6 continuous variables are transformed successfully using Johnson Su transformation (Team used Best Fit Transformation). Rest continuous variables are used as is while modelling. Please refer to below screenshot.



Quantiles

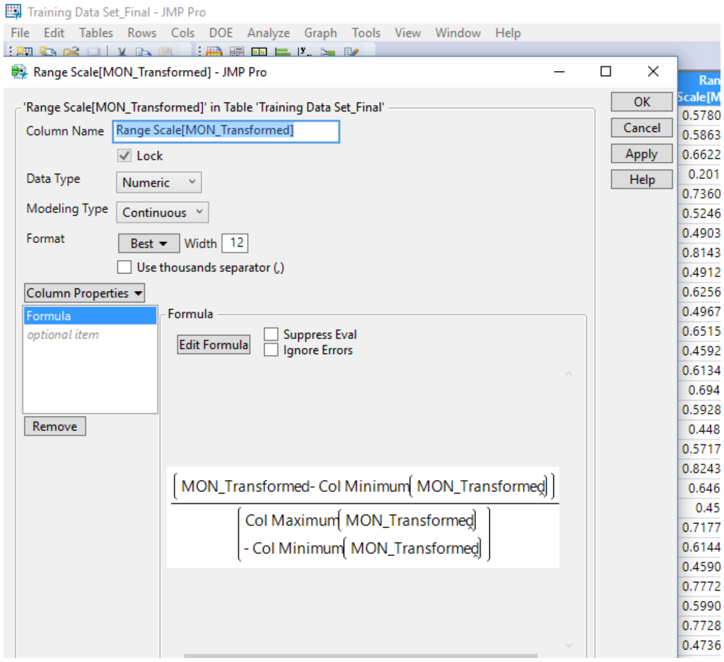
100.0%	maximum	0.82
99.5%		0.72995
97.5%		0.69
90.0%		0.66
75.0%	quartile	0.61
50.0%	median	0.54
25.0%	quartile	0.46
10.0%		0.35
2.5%		0.17
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.5181562
Std Dev	0.1296497
Std Err Mean	0.0032412
Upper 95% Mean	0.5245138
Lower 95% Mean	0.5117987

Standardization:

Standardization is done on all the continuous variables including transformed variables to bring them on a same scale of 0 to 1. Without standardization we might end up looking at erroneous values when distance is calculated and used in modelling. To avoid this and to be on a safer side, we standardized all the continuous variables. Please refer to below screenshot.


Training Data Set_Final - JMP Pro
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
Range Scale[MON_Transformed] - JMP Pro
OK
Cancel
Apply
Help
Column Properties
Remove
Edit Formula
Ran
Scale[M

'Range Scale[MON_Transformed]' in Table 'Training Data Set_Final'

Column Name: Range Scale[MON_Transformed]

☑ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: Best, Width 12

☐ Use thousands separator (,)

Formula

optional item

Formula

☐ Suppress Eval

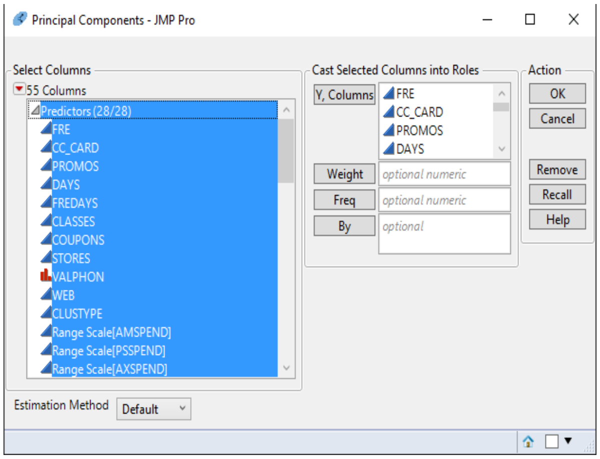
☐ Ignore Errors

$$\frac{\text{MON\_Transformed} - \text{Col Minimum}(\text{MON\_Transformed})}{\text{Col Maximum}(\text{MON\_Transformed}) - \text{Col Minimum}(\text{MON\_Transformed})}$$

0.5780
0.5863
0.6622
0.201
0.7360
0.5246
0.4903
0.8143
0.4912
0.6256
0.4967
0.6515
0.4592
0.6134
0.694
0.5928
0.448
0.5717
0.8243
0.646
0.45
0.7177
0.6144
0.4590
0.7772
0.5990
0.7728
0.4736

Principal Component Analysis:

The team performed PCA on all the 27 predictors or variables to observe the reduction in the dimensionality. We observed that 19 Principal Components could be used to retain almost 85% of variance in the original data. Since it is difficult to explain predictor contribution to the business using principal components, team decided not to use them for building a model. Principal Component Analysis is done as shown below:



Select Columns

55 Columns

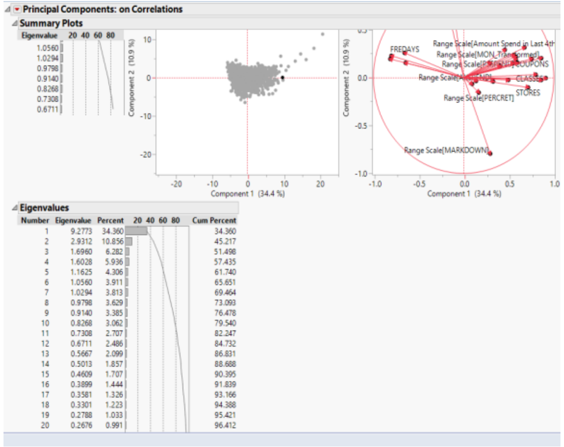
Predictors (28/28)
FRE
CC_CARD
PROMOS
DAYS
FREDAYS
CLASSES
COUPONS
STORES
VALPHON
WEB
CLUSTYPE
Range Scale[AMSPEND]
Range Scale[PSSPEND]
Range Scale[AXSPEND]

Cast Selected Columns into Roles

Y, Columns	FRE CC_CARD PROMOS DAYS
Weight	*optional numeric*
Freq	*optional numeric*
By	*optional*

Action

Estimation Method: Default



Eigenvalues

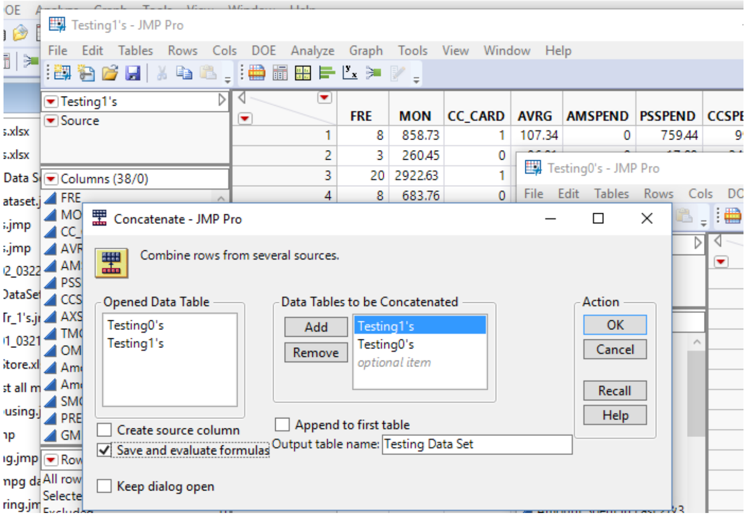
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	9.2773	34.360		34.360
2	2.9312	10.856		45.217
3	1.6960	6.282		51.498
4	1.6028	5.936		57.435
5	1.1625	4.306		61.740
6	1.0560	3.911		65.651
7	1.0294	3.813		69.464
8	0.9798	3.629		73.093
9	0.9140	3.385		76.478
10	0.8268	3.062		79.540
11	0.7308	2.707		82.247
12	0.6711	2.486		84.732
13	0.5667	2.099		86.831
14	0.5013	1.857		88.688
15	0.4609	1.707		90.395
16	0.3899	1.444		91.839
17	0.3581	1.326		93.166
18	0.3301	1.223		94.388
19	0.2788	1.033		95.421
20	0.2676	0.991		96.412

Balancing Data Set:

We separated our data with respect to responses that is, 0’s and 1’s and made two subset tables respectively. Then team has randomly selected a sample of 800 observations of 0’s each for testing and validation and then concatenated them to similarly selected 800 observations of 1’s each for testing and validation using MS Excel.

Unbalanced Data Set for Testing the model:

Now to run our trained and validated model, the team has selected the remaining data after balancing. To uphold the proportion of yes to no responses in the original data set (that is 1:5), the team selected observations maintaining the ratio of the original data.



	FRE	MON	CC_CARD	AVRG	AMSPEND	PSSPEND	CCSPE
1	8	858.73	1	107.34	0	759.44	9
2	3	260.45	0				
3	20	2922.63	1				
4	8	683.76	0				

Combine rows from several sources.

Opened Data Table: Testing0's, Testing1's

Data Tables to be Concatenated: Testing1's, Testing0's, *optional item*

□ Create source column

☑ Save and evaluate formulas

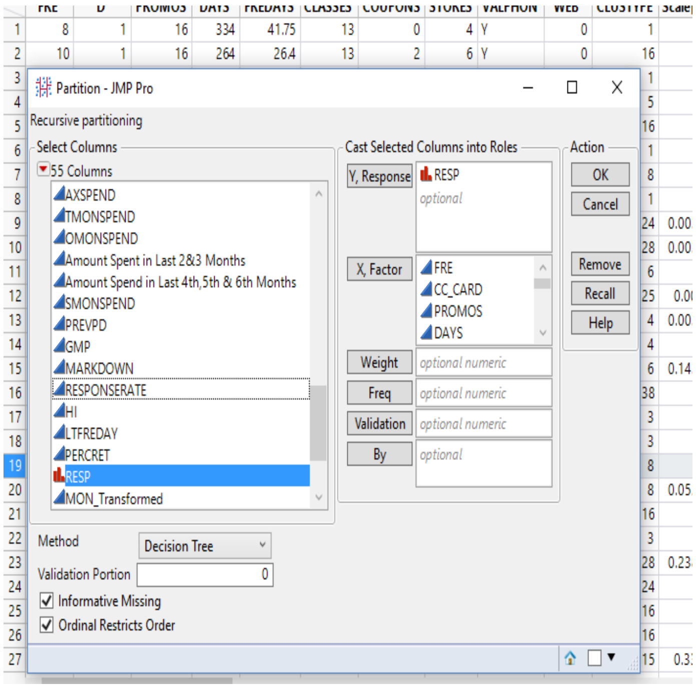
□ Keep dialog open

□ Append to first table

Output table name: Testing Data Set

Data Modeling

Training model using Decision Tree on Balanced data set as follows:



Y, Response: RESP

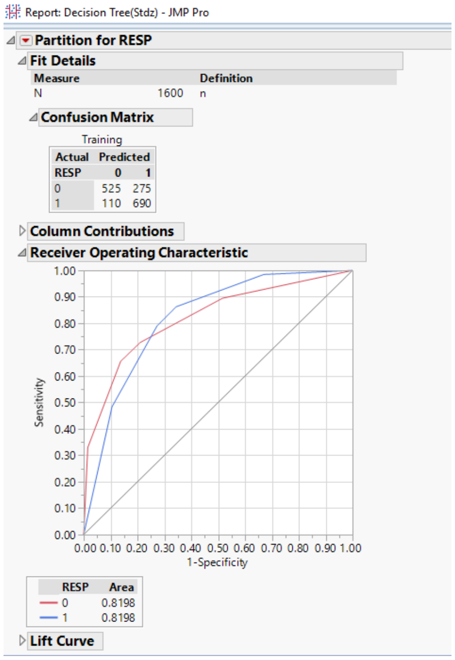
X, Factor: FRE, CC_CARD, PROMOS, DAYS

Method: Decision Tree

Validation Portion: 0

☑ Informative Missing

☑ Ordinal Restricts Order



Partition for RESP

Fit Details

Measure		Definition
N	1600	n

Confusion Matrix

Training

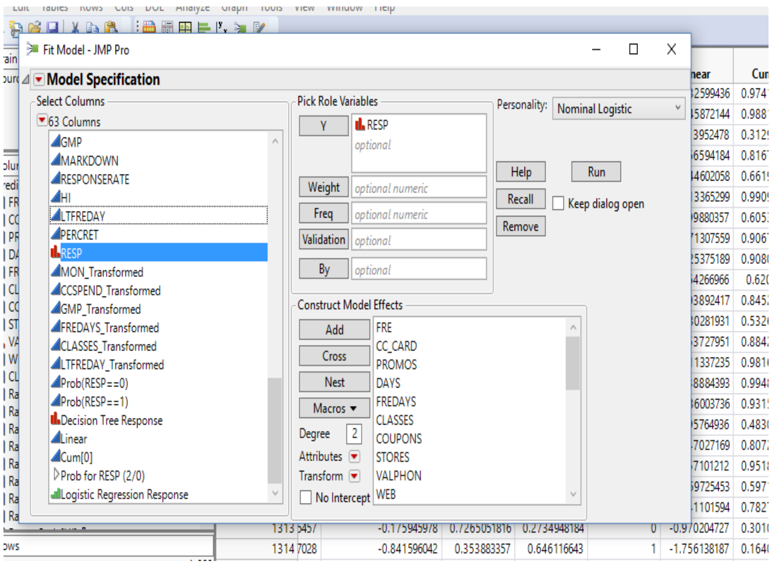
Actual RESP	Predicted 0	1
0	525	275
1	110	690

Receiver Operating Characteristic

RESP	Area
0	0.8198
1	0.8198

Please refer Jmp report file ‘Decision Tree report-Training.jrp’ for detailed results.

Training model using Logistic Regression method using balanced data set is shown below:



Y: RESP

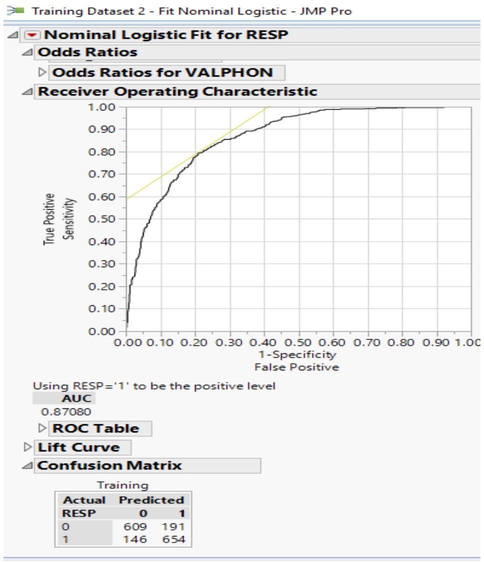
Personality: Nominal Logistic

□ Keep dialog open

Construct Model Effects: FRE, CC_CARD, PROMOS, DAYS, FREDAYS, CLASSES, COUPONS, STORES, VALPHON, WEB

Degree: 2

□ No Intercept



Nominal Logistic Fit for RESP

Receiver Operating Characteristic

Using RESP='1' to be the positive level

AUC

0.87080

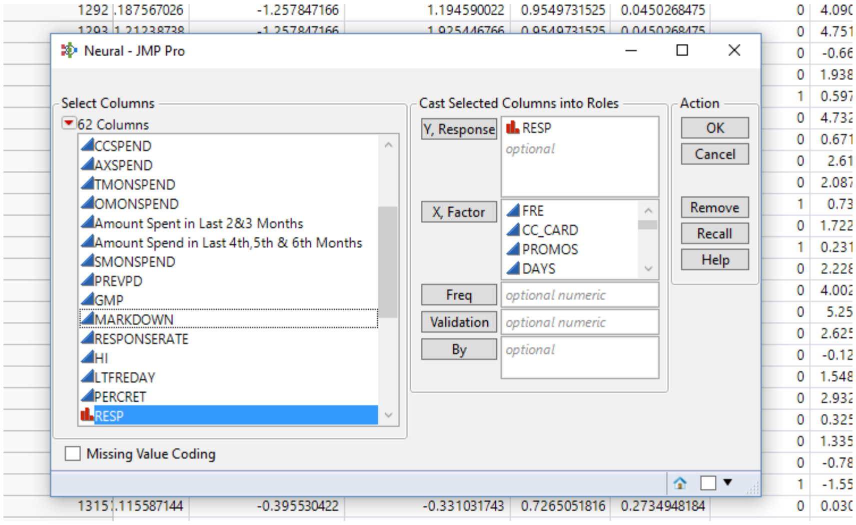
Confusion Matrix

Training

Actual RESP	Predicted 0	1
0	609	191
1	146	654

Please refer Jmp report file ‘Logistic Regression report-Training.jrp’ for detailed results.

Training model using Neural Networks method using balanced data set is shown below:



```
1292 .187567026   -1.257847166   1.194590022  0.9549731525  0.0450268475   0  4.09(
1293 1.21238738   -1.257847166   1.925446766  0.9549731525  0.0450268475   0  4.751
1315 .115587144   -0.395530422  -0.331031743  0.7265051816  0.2734948184   0  0.03(
```

Neural - JMP Pro

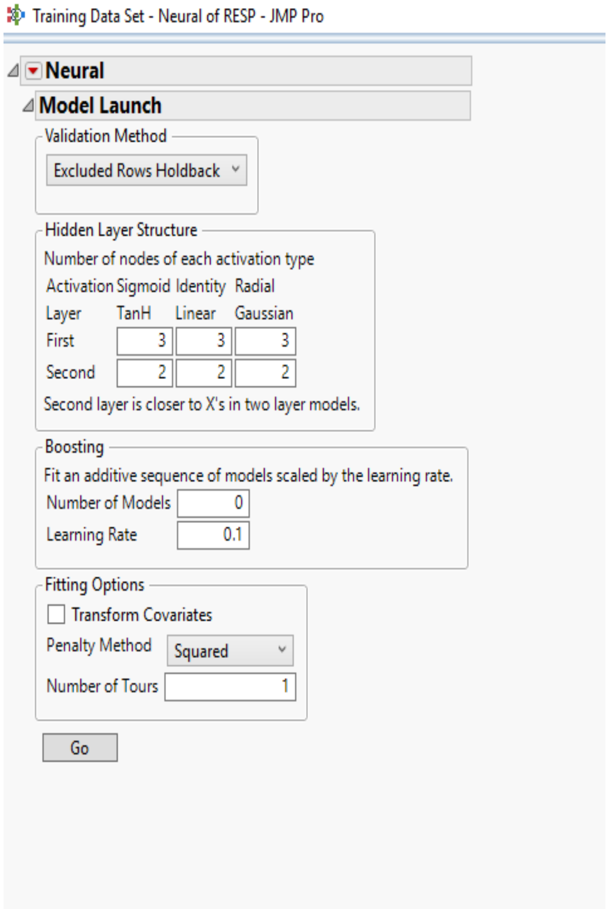
Select Columns

62 Columns

- CCSPEND
- AXSPEND
- TMONSPEND
- OMONSPEND
- Amount Spent in Last 2&3 Months
- Amount Spend in Last 4th,5th & 6th Months
- SMONSPEND
- PREVPD
- GMP
- MARKDOWN
- RESPONSERATE
- HI
- LTFREDAY
- PERCRET
- RESP

Cast Selected Columns into Roles

Role	Columns
Y, Response	RESP optional
X, Factor	FRE CC_CARD PROMOS DAYS
Freq	optional numeric
Validation	optional numeric
By	optional



Training Data Set - Neural of RESP - JMP Pro

Neural

Model Launch

Validation Method: Excluded Rows Holdback

Hidden Layer Structure

Number of nodes of each activation type

Activation	Sigmoid	Identity	Radial
Layer	TanH	Linear	Gaussian
First	3	3	3
Second	2	2	2

Second layer is closer to X's in two layer models.

Boosting

Fit an additive sequence of models scaled by the learning rate.

Number of Models: 0

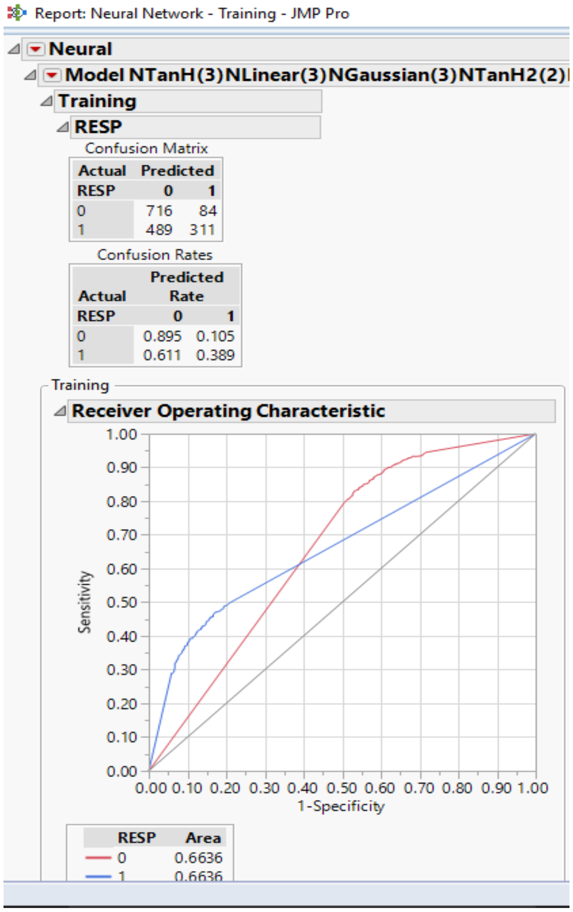
Learning Rate: 0.1

Fitting Options

□ Transform Covariates

Penalty Method: Squared

Number of Tours: 1



Report: Neural Network - Training - JMP Pro

Neural

Model NTanH(3)NLinear(3)NGaussian(3)NTanH2(2)

Training

RESP

Confusion Matrix

Actual	Predicted	
RESP	0	1
0	716	84
1	489	311

Confusion Rates

Actual	Predicted Rate	
RESP	0	1
0	0.895	0.105
1	0.611	0.389

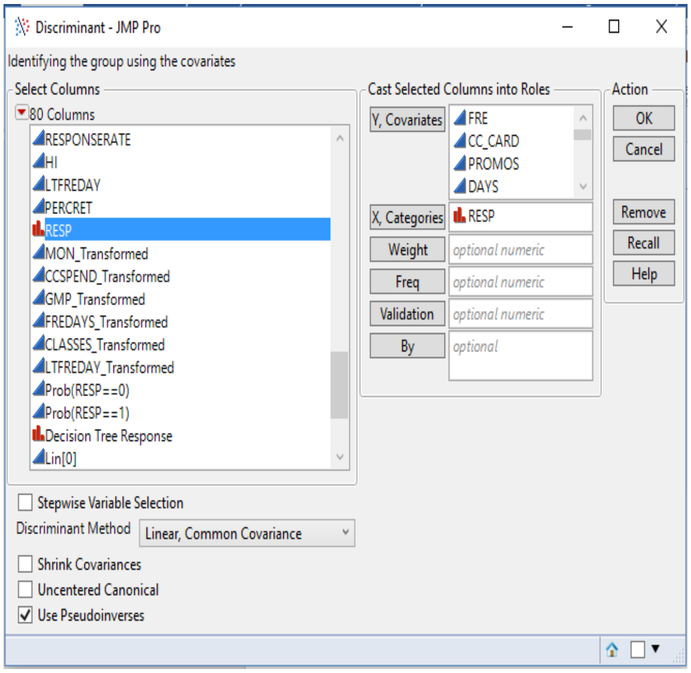
Training

Receiver Operating Characteristic

RESP	Area
0	0.6636
1	0.6636

Please refer Jmp report file ‘Neural Network - Training.jrp’ for detailed results.

Training model using Discriminative Analysis method using balanced data set is shown below:



Discriminant - JMP Pro

Identifying the group using the covariates

Select Columns

80 Columns

- RESPONSERATE
- HI
- LTFREDAY
- PERCRET
- RESP
- MON_Transformed
- CCSPEND_Transformed
- GMP_Transformed
- FREDAYS_Transformed
- CLASSES_Transformed
- LTFREDAY_Transformed
- Prob(RESP==0)
- Prob(RESP==1)
- Decision Tree Response
- Lin[0]

Cast Selected Columns into Roles

Y, Covariates	FRE CC_CARD PROMOS DAYS
X, Categories	RESP
Weight	optional numeric
Freq	optional numeric
Validation	optional numeric
By	optional

Action

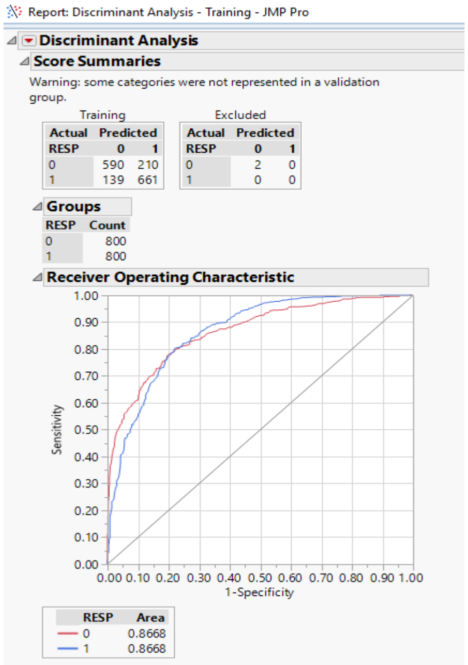
□ Stepwise Variable Selection

Discriminant Method: Linear, Common Covariance

□ Shrink Covariances

□ Uncentered Canonical

☑ Use Pseudoinverses



Report: Discriminant Analysis - Training - JMP Pro

Discriminant Analysis

Score Summaries

Warning: some categories were not represented in a validation group.

Training

Actual	Predicted	
RESP	0	1
0	590	210
1	139	661

Excluded

Actual	Predicted	
RESP	0	1
0	2	0
1	0	0

Groups

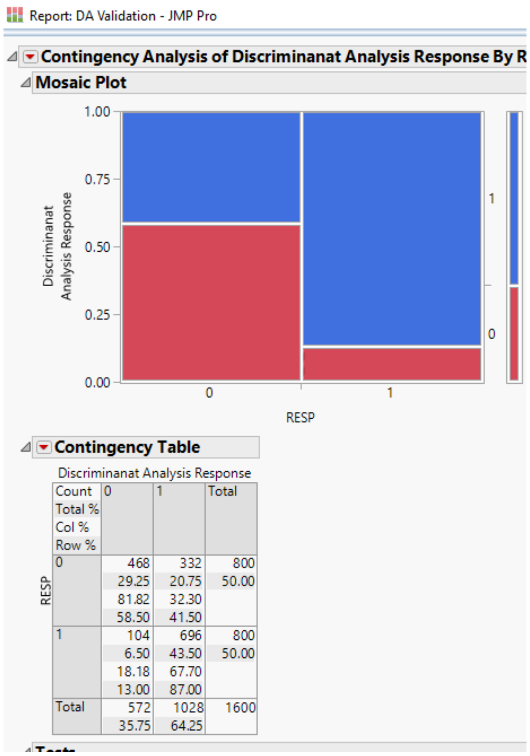
RESP	Count
0	800
1	800

Receiver Operating Characteristic

RESP	Area
🟥 0	0.8668
🟦 1	0.8668

Please refer Jmp report file ‘Descriptive Analysis - Training.jrp’ for detailed results.

Discriminant Analysis and Neural Network contingency tables results on validation data:



Report: DA Validation - JMP Pro

Contingency Analysis of Discriminanat Analysis Response By R

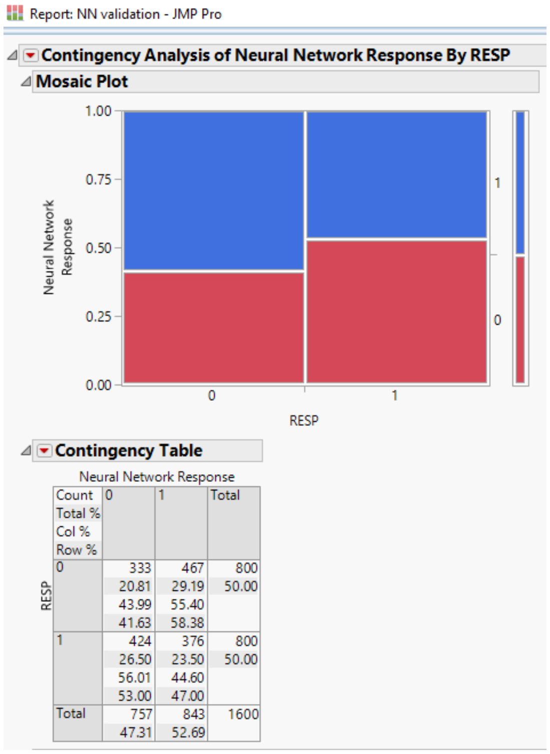
Mosaic Plot

Contingency Table

Discriminanat Analysis Response

RESP	Count Total % Col % Row %	0	1	Total
	0	468 29.25 81.82 58.50	332 20.75 32.30 41.50	800 50.00
	1	104 6.50 18.18 13.00	696 43.50 67.70 87.00	800 50.00
	Total	572 35.75	1028 64.25	1600

Tests



Report: NN validation - JMP Pro

Contingency Analysis of Neural Network Response By RESP

Mosaic Plot

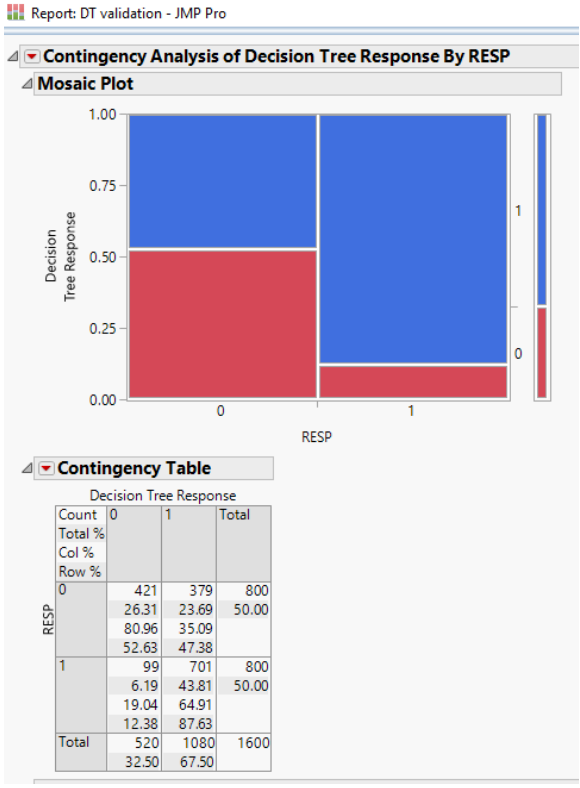
Contingency Table

Neural Network Response

RESP	Count Total % Col % Row %	0	1	Total
	0	333 20.81 43.99 41.63	467 29.19 55.40 58.38	800 50.00
	1	424 26.50 56.01 53.00	376 23.50 44.60 47.00	800 50.00
	Total	757 47.31	843 52.69	1600

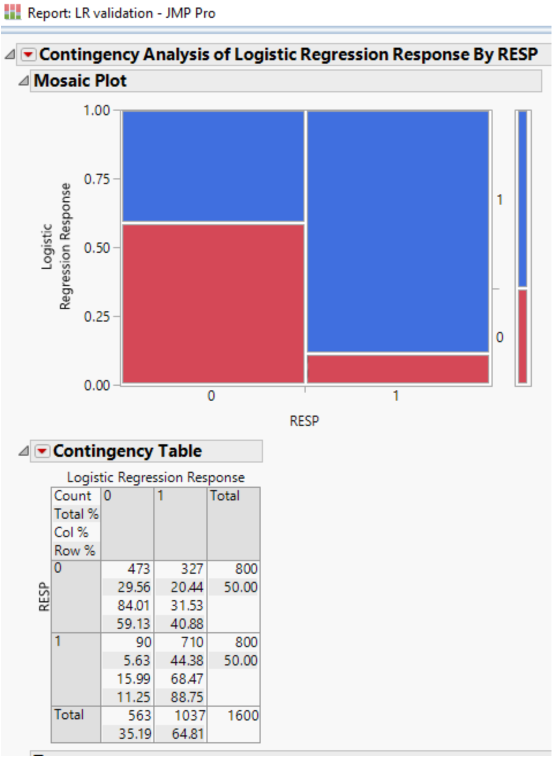
Please refer Jmp report files ‘DA Validation.jrp’ & ‘NN Validation.jrp’ for detailed results.

Decision Tree and Logistic Regression Contingency plots results on validation data set:


Report: DT validation - JMP Pro
Contingency Analysis of Decision Tree Response By RESP
Mosaic Plot
Decision Tree Response
1.00
0.75
0.50
0.25
0.00
0
1
RESP
1
0
Contingency Table

Decision Tree Response

	Count Total % Col % Row %	0	1	Total
RESP	0	421 26.31 80.96 52.63	379 23.69 35.09 47.38	800 50.00
	1	99 6.19 19.04 12.38	701 43.81 64.91 87.63	800 50.00
	Total	520 32.50	1080 67.50	1600

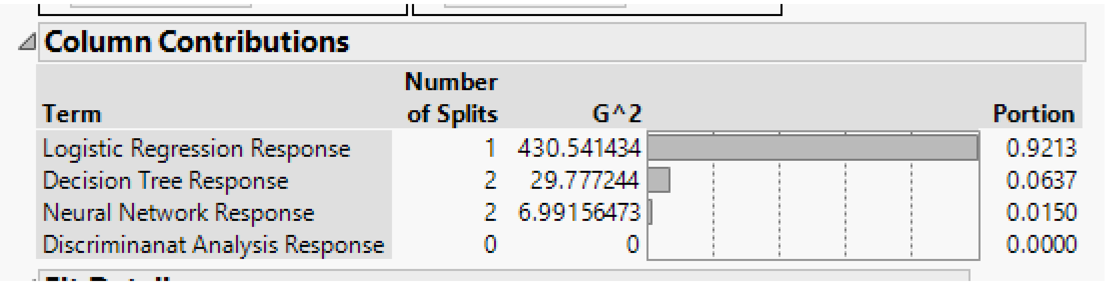

Report: LR validation - JMP Pro
Contingency Analysis of Logistic Regression Response By RESP
Mosaic Plot
Logistic Regression Response
1.00
0.75
0.50
0.25
0.00
0
1
RESP
1
0
Contingency Table

Logistic Regression Response

	Count Total % Col % Row %	0	1	Total
RESP	0	473 29.56 84.01 59.13	327 20.44 31.53 40.88	800 50.00
	1	90 5.63 15.99 11.25	710 44.38 68.47 88.75	800 50.00
	Total	563 35.19	1037 64.81	1600

Please refer Jmp report files ‘LR Validation.jrp’ & ‘DT Validation.jrp’ for detailed results.

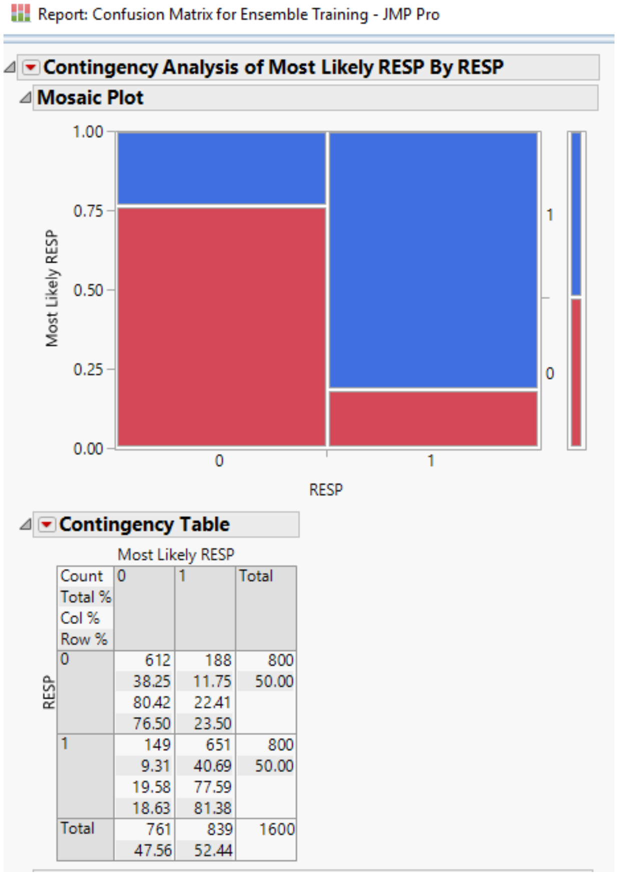
Contributions of different models to actual responses is shown below:



Column Contributions

Term	Number of Splits	G^2	Portion
Logistic Regression Response	1	430.541434	0.9213
Decision Tree Response	2	29.777244	0.0637
Neural Network Response	2	6.99156473	0.0150
Discriminanat Analysis Response	0	0	0.0000

Contingency analysis for Ensemble model on training and validation data set as follows:


Report: Confusion Matrix for Ensemble Training - JMP Pro

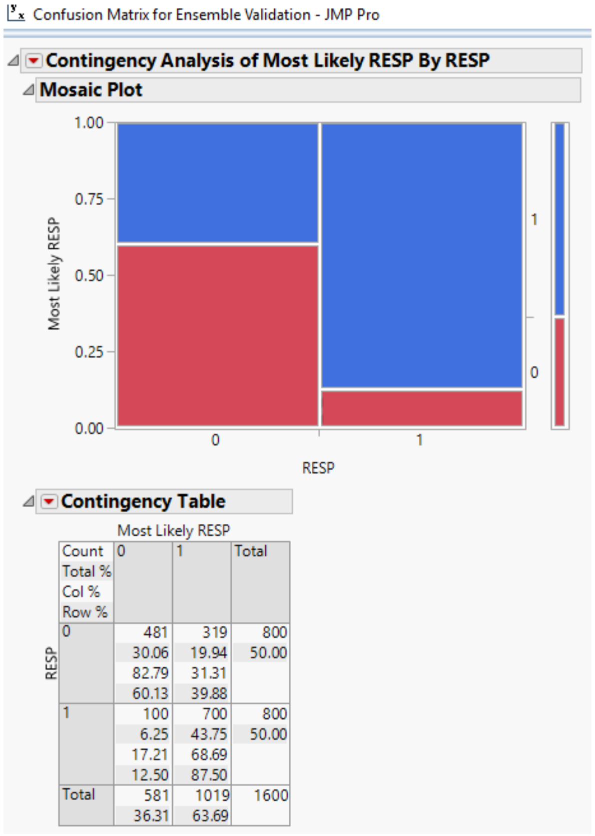
Contingency Analysis of Most Likely RESP By RESP

Mosaic Plot

Contingency Table

Most Likely RESP

RESP	Count Total % Col % Row %	0	1	Total
	0	612 38.25 80.42 76.50	188 11.75 22.41 23.50	800 50.00
	1	149 9.31 19.58 18.63	651 40.69 77.59 81.38	800 50.00
	Total	761 47.56	839 52.44	1600


Confusion Matrix for Ensemble Validation - JMP Pro

Contingency Analysis of Most Likely RESP By RESP

Mosaic Plot

Contingency Table

Most Likely RESP

RESP	Count Total % Col % Row %	0	1	Total
	0	481 30.06 82.79 60.13	319 19.94 31.31 39.88	800 50.00
	1	100 6.25 17.21 12.50	700 43.75 68.69 87.50	800 50.00
	Total	581 36.31	1019 63.69	1600

Please refer to ‘Confusion Matrix for Ensemble Training.jrp’ and Confusion Matrix for Ensemble Validation.jrp’ for detailed results.

Confusion Matrix for all the models on traning and validation data sets is listed as below:

True Negative	False Negative
DT Model = 525	DT Model = 275
LR Model = 609	LR Model = 191
NN Model = 445	NN Model = 355
DA Model = 590	DA Model = 210
False Positive	True Positive
DT Model = 110	DT Model = 690
LR Model = 146	LR Model = 654
NN Model = 457	NN Model = 343
DA Model = 139	DA Model = 661

Training Data Confusion Matrix

True Negative	False Negative
DT Model = 421	DT Model = 379
LR Model = 473	LR Model = 327
NN Model = 333	NN Model = 467
DA Model = 468	DA Model = 332
False Positive	True Positive
DT Model = 99	DT Model = 701
LR Model = 90	LR Model = 710
NN Model = 424	NN Model = 376
DA Model = 104	DA Model = 696

Validation Data Confusion Matrix

Confusion Matrix and Measures:

Shown below is the general structure of confusion matrix where outcome is the actual results from the data set and test indicator is the most likely predicted values using the model.

- True Negative is that the model predicted a No and the actual result is also a No.
- False Negative is that the model predicted a No while the actual result is a Yes.
- False Positive is that the model predicted Yes while the actual result is a No.
- True Positive is that the model predicted No and the actual result is also a No.

		Test Indicator	
		No	Yes
Outcome	No	a True Negative	b False Positive
	Yes	c False Negative	d True Positive

True Negatives and True Positives determine the accuracy of the model, while False positives and False Negatives should be as low as possible in a good model. There should be a trade-off on these two decision outcomes based on the business setting.

Accuracy of Model:

Accuracy is the proportion of individuals who were correctly classified– that is, the proportions of True Positives and True Negatives.

Percentage of Model Accuracy is:

$$\frac{(\boldsymbol{a}+\boldsymbol{d}) * \mathbf{100}}{(\boldsymbol{a}+\boldsymbol{b}+\boldsymbol{c}+\boldsymbol{d})}$$

Misclassification in Model:

This is negation to the accuracy of the model. It is calculated as:

100 – % of Model Accuracy

Specificity in Model:

Specificity is the proportion of observed negatives that were predicted to be negatives. That is,

$$\frac{\mathbf{100} * \boldsymbol{a}}{(\boldsymbol{a}+\boldsymbol{b})}$$

Sensitivity in Model:

Sensitivity is the proportion of observed positives that were predicted to be positive. That is,

$$\frac{\mathbf{100} * \boldsymbol{d}}{(\boldsymbol{c}+\boldsymbol{d})}$$

Specificity and Sensitivity are inversely proportional to each other. That is, increasing sensitivity leads to decrease in specificity. This should be balanced accordingly depending on the business setting where the models are used.