INTRODUCTION

This report intends to present a prediction summary of the dataset related with direct marketing campaigns done by a Portuguese banking institution. The main aim of this study is to identify customers who are most likely to subscribe to a term deposit account based on the results of the previous marketing campaign and also to compare the decisions of the models derived from this study, with that of a previous research paper "*A Data-Driven Approach to Predict the Success of Bank Telemarketing*" done on this same dataset by researchers - S. Moro, P. Cortez and P. Rita.

Various Data Mining models and techniques were employed using SAS Enterprise Miner 14.3 on the dataset, to get the best prediction results.

1. UNDERSTANDING THE DATA

There are totally 41,188 instances, 20 input variables and 1 target variable. This dataset is a result of the previous marketing campaigns that were done by the Portuguese bank via phone calls. The clients were contacted from May 2008 to November 2010, to know whether they will subscribe to a bank term deposit or not.

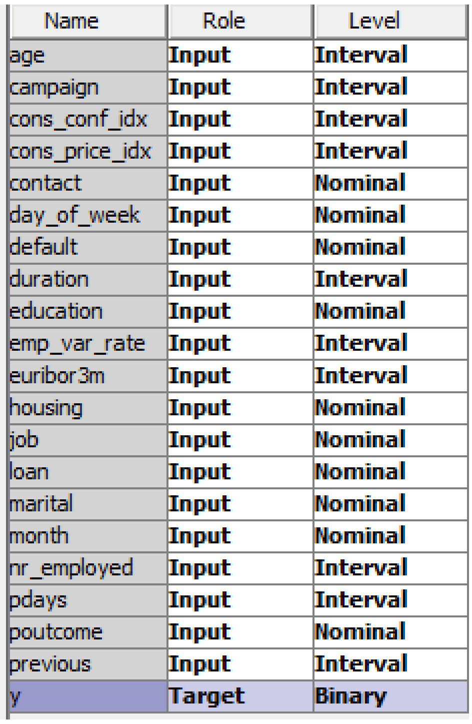
To understand what makes a client to accept a term deposit, there are a number of factors that could influence this decision by their personal indicators, such as: Their type of employment, their age, their education level, their marital status, whether they have a loan already, and by the current socio-economic status of the country and job sector they belong to, such as: employment variation rate, consumer price index, consumer confidence index, to name a few.

The input data is broadly divided into four categories:

Client personal data, Last contact of the current campaign information, Other and Socio-Economic attributes.

The target/output variable is a binary variable "Y" with the values: Yes (indicating that the customer has taken the term deposit) and No (indicating that the customer did not take the term deposit).

Table 1.1 shows the roles and levels of each of the variable:



Name	Role	Level
age	Input	Interval
campaign	Input	Interval
cons_conf_idx	Input	Interval
cons_price_idx	Input	Interval
contact	Input	Nominal
day_of_week	Input	Nominal
default	Input	Nominal
duration	Input	Interval
education	Input	Nominal
emp_var_rate	Input	Interval
euribor3m	Input	Interval
housing	Input	Nominal
job	Input	Nominal
loan	Input	Nominal
marital	Input	Nominal
month	Input	Nominal
nr_employed	Input	Interval
pdays	Input	Interval
poutcome	Input	Nominal
previous	Input	Interval
y	Target	Binary

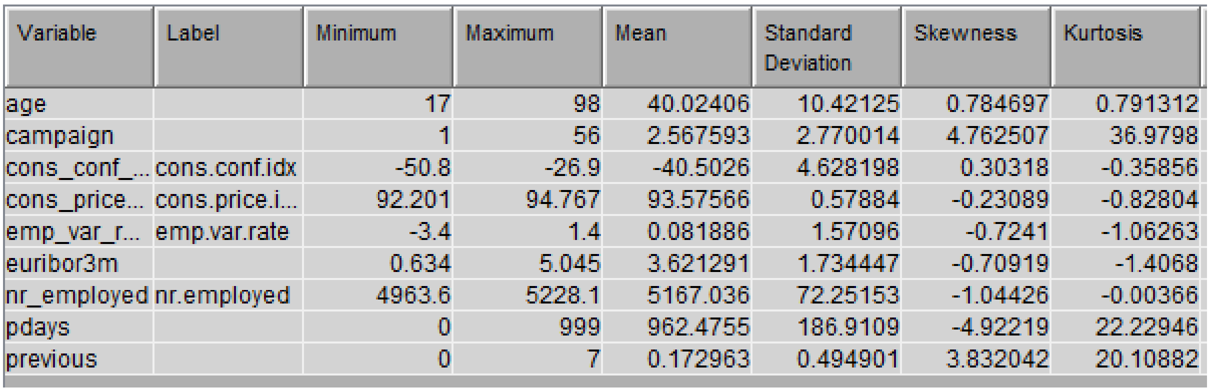
Table 1.1

1.1 DESCRIPTIVE STATISTICS

Summary statistics for interval variables:

The DMDB node from Explore tab is used to get the summary statistics of the interval variables. The following figure gives us the mean, standard deviation, skewness, kurtosis, minimum and maximum values of each of the interval variable.

From the below table, it is clear that the skewness and kurtosis of the variables campaign, pdays and previous are highly skewed and have too much kurtosis. A data transformation will be performed on these variables, which will be described in the data transformation part of the report.



Variable	Label	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
age		17	98	40.02406	10.42125	0.784697	0.791312
campaign		1	56	2.567593	2.770014	4.762507	36.9798
cons_conf_...	cons.conf.idx	-50.8	-26.9	-40.5026	4.628198	0.30318	-0.35856
cons_price...	cons.price.i...	92.201	94.767	93.57566	0.57884	-0.23089	-0.82804
emp_var_r...	emp.var.rate	-3.4	1.4	0.081886	1.57096	-0.7241	-1.06263
euribor3m		0.634	5.045	3.621291	1.734447	-0.70919	-1.4068
nr_employed	nr.employed	4963.6	5228.1	5167.036	72.25153	-1.04426	-0.00366
pdays		0	999	962.4755	186.9109	-4.92219	22.22946
previous		0	7	0.172963	0.494901	3.832042	20.10882

Table 1.2 Summary Statistics of Interval Variables

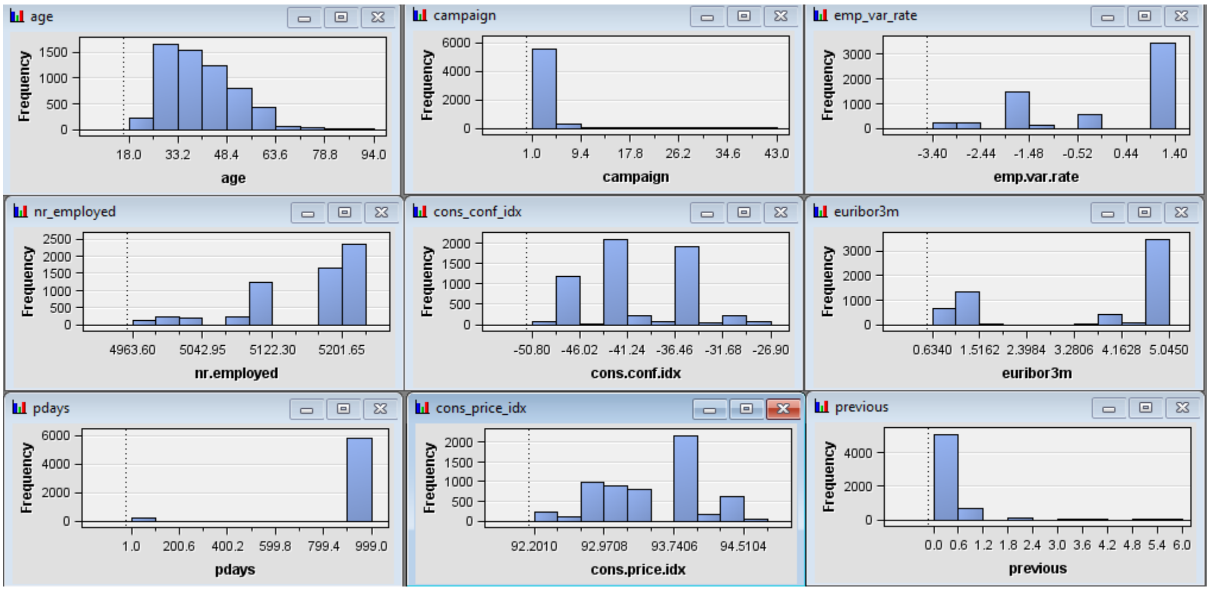


Figure 1.1: Histogram of all the interval input variables

Age:

The variable age appears to have a fairly right-skewed distribution. Age ranges from 17 to 98 of clients. The skew and kurtosis values of age are in the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further.

Campaign:

Campaign is the number of contacts performed during this campaign and for this client. It has a sharp kurtosis with poor distribution. Value of campaign ranges from 1 to 56. The skew and kurtosis values of age are not in the range of +/-2, indicating that the variable is not normal and thus needs to be transformed further.

Employment variation rate:

Employment rates are defined as a measure of the extent to which available labour resources (people available to work) are being used. They are calculated as the ratio of the employed to the working age population. Employment rates are sensitive to the economic cycle, but in the longer term they are significantly affected by governments' higher education and income support policies [1].

This is a socio-economic quarterly indicator. The distribution appears to be sparsely distributed with right skewed distribution and the range of the values are from -3.4 to 1.4. The skew and kurtosis values are in the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further.

Number of employees:

This is another socio-economic quarterly indicator. The distribution appears to be sparsely distributed and the range of the values are from 4963.6 to 5228.1. The skew and kurtosis values are in the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further. From the variable worth plot and chi-square plot, it can be inferred that this is a very important contributor to the target variable.

Consumer confidence index:

This is a socio-economic monthly indicator. The Consumer Confidence Index (CCI) is an index by The Conference Board that measures how optimistic or pessimistic consumers are with respect to the economy in the near future. The Consumer Confidence Index (CCI) is based on the concept that if consumers are optimistic, they tend to purchase more goods and services. This increase in spending inevitably stimulates the whole economy [2].

The distribution appears to be distributed across a large range of values. The skew and kurtosis values of age are in the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further. From the variable worth plot and chi-square plot, it can be inferred that this is also a very important contributor to the target variable.

Consumer price index:

The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food and medical care. It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. Changes in the CPI are used to assess price changes associated with the cost of living; the CPI is one of the most frequently used statistics for identifying periods of inflation or deflation [3].

The skew and kurtosis values are within the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further. From the variable worth plot and chi-square plot, it can be inferred that this is also a very important contributor to the target variable.

Euribor 3 month rate:

The 3 month Euribor interest rate is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months [4].

The skew and kurtosis values are within the range of +/-2, indicating that the variable is normal and that it does not need to be transformed further. From the variable worth plot and chi-square plot, it can be inferred that this is also a very important contributor to the target variable.

Pdays:

It is the number of days that passed by after the client was last contacted from a previous campaign. In this variable the value 999 indicates that the client was not previously contacted.

The skew is not within the range of +/-2 and has high kurtosis, indicating that the variable is not normal and it needs to be transformed further. From the variable worth plot and chi-square plot, it can be inferred that this is also a very important contributor to the target variable.

Previous:

This indicates the number of contacts performed before this campaign, for this client. The skew is not within the range of +/-2 and has high kurtosis, indicating that the variable is not normal and it needs to be transformed further.

Target variable Y:

This is a binary variable, indicating whether a client subscribed to a term deposit.

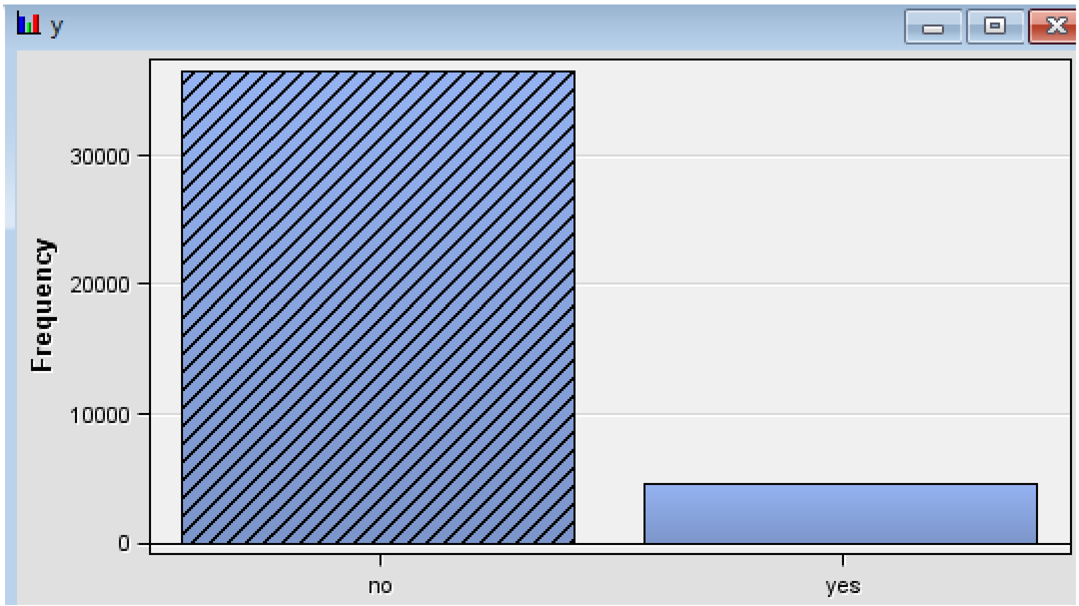


Figure 1.2: Target variable y

From the figure, variable y indicates that 88% of the clients have not taken a term deposit and only 12% of the clients have taken the term deposit.

The nodes StatExplore, Graph Explore and MultiPlot are used for descriptive statistics.

Chi-Square plot is used for both categorical and interval variables. The output shows the Chi-Square plot and Variable worth which shows the importance of variables while predicting the target variable.

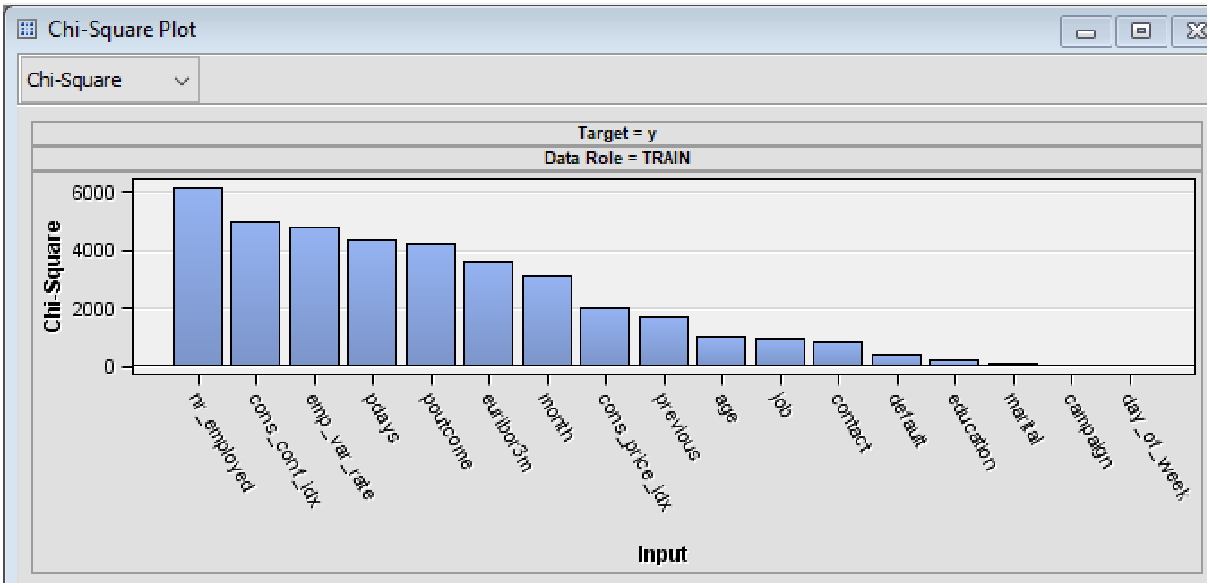


Figure 1.3: Chi-Square Plot of all variables against target variable

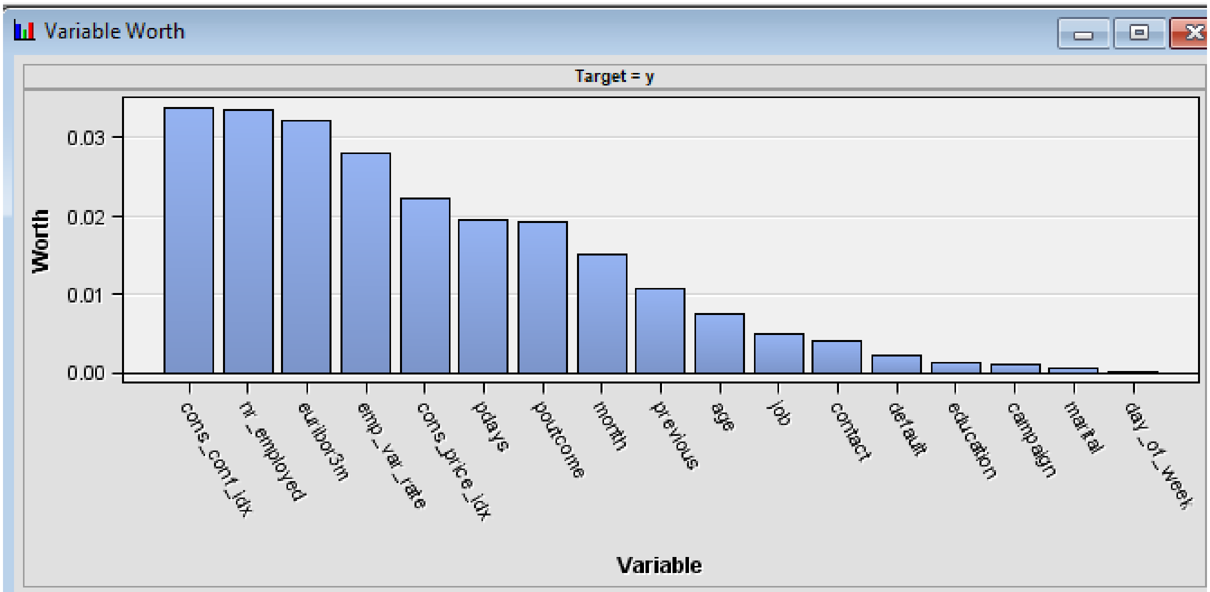
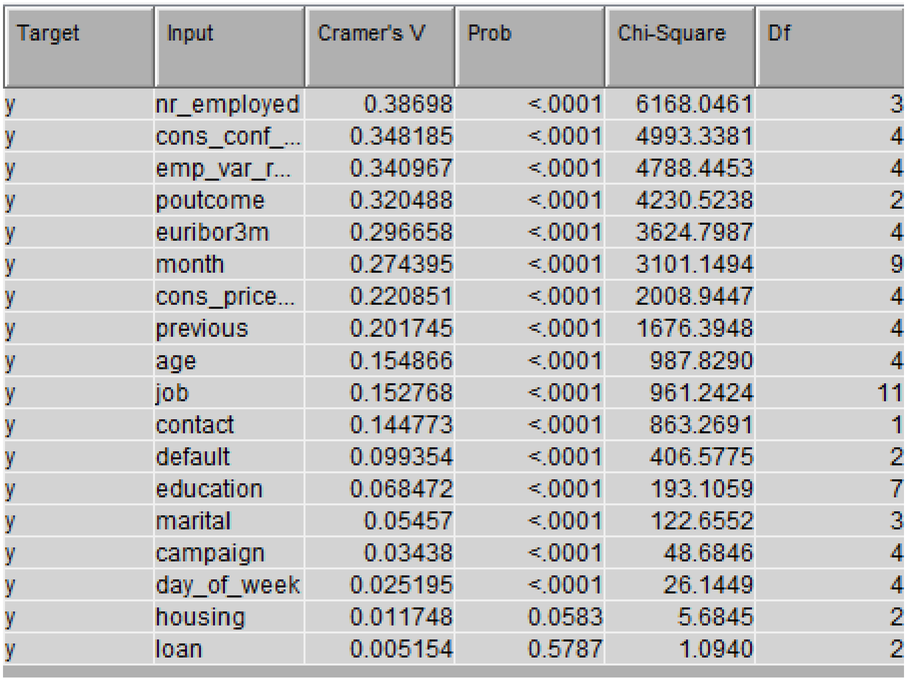


Figure 1.4: Variable worth Graph of all variables contributing to the value of target variable y

From the graph, it is clear that the variables cons.conf.idx, cons.price.idx, nr.employed, euribor3m, emp.var.rate, pdays, poutcome, month, previous contribute significantly to the target variable y.



Target	Input	Cramer's V	Prob	Chi-Square	Df
y	nr_employed	0.38698	<.0001	6168.0461	3
y	cons_conf_...	0.348185	<.0001	4993.3381	4
y	emp_var_r...	0.340967	<.0001	4788.4453	4
y	poutcome	0.320488	<.0001	4230.5238	2
y	euribor3m	0.296658	<.0001	3624.7987	4
y	month	0.274395	<.0001	3101.1494	9
y	cons_price...	0.220851	<.0001	2008.9447	4
y	previous	0.201745	<.0001	1676.3948	4
y	age	0.154866	<.0001	987.8290	4
y	job	0.152768	<.0001	961.2424	11
y	contact	0.144773	<.0001	863.2691	1
y	default	0.099354	<.0001	406.5775	2
y	education	0.068472	<.0001	193.1059	7
y	marital	0.05457	<.0001	122.6552	3
y	campaign	0.03438	<.0001	48.6846	4
y	day_of_week	0.025195	<.0001	26.1449	4
y	housing	0.011748	0.0583	5.6845	2
y	loan	0.005154	0.5787	1.0940	2

Table 1.3: Chi-Square Plot

Examining the table values of Chi-Square Plot from table 1.3, all the variables have statistically significant correlation with the target variable (p-value < 0.05) apart from the variables loan and housing. Although, the variable housing is approaching normality with a value of 0.0583, it is still a source of concern for including this variable in a predictive model.

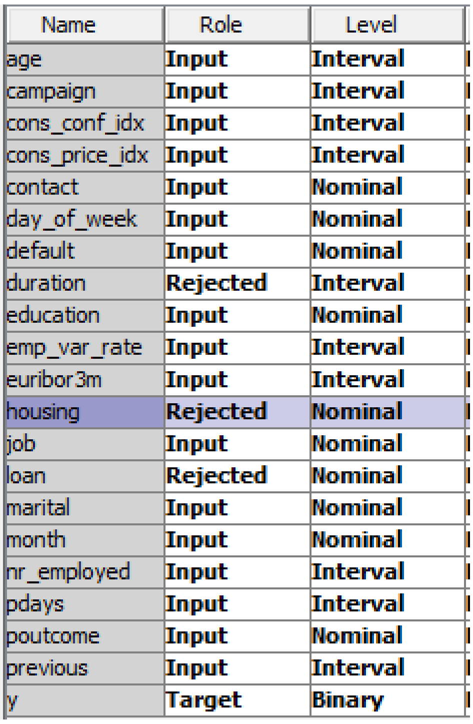
1.2 DATA PREPARATION

This is the stage where data cleaning, imputing, transforming, variable selection, variable reduction are applied on the dataset, for it to be used in analysis. It is essential that we ensure data is clean before beginning any analysis.

1.2.1 FEATURE REDUCTION

The variable duration is the total number of seconds spoken between the customer and agent in the last contact. This variable highly affects the output target and hence for avoiding bias in our dataset, it is recommended to reject it, to get a realistic prediction model.

The variables loan and housing are rejected based on table 1.3's p-values. They indicate that there is no statistically significant correlation between the target variable. Hence below table is the final list of variables that will be used in this analysis:



Name	Role	Level
age	Input	Interval
campaign	Input	Interval
cons_conf_idx	Input	Interval
cons_price_idx	Input	Interval
contact	Input	Nominal
day_of_week	Input	Nominal
default	Input	Nominal
duration	Rejected	Interval
education	Input	Nominal
emp_var_rate	Input	Interval
euribor3m	Input	Interval
housing	Rejected	Nominal
job	Input	Nominal
loan	Rejected	Nominal
marital	Input	Nominal
month	Input	Nominal
nr_employed	Input	Interval
pdays	Input	Interval
poutcome	Input	Nominal
previous	Input	Interval
y	Target	Binary

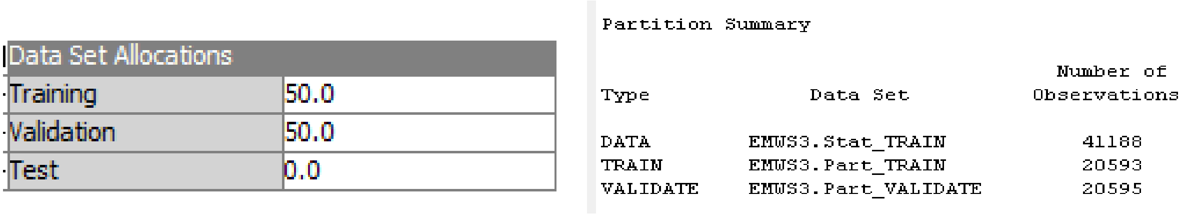
Table 1.4

1.2.2 HANDLING MISSING VALUES

There are no NA values in this dataset. Only unknown values are present in categorical variables. For the purpose of this study, I am going to treat unknown as it is. Thus, unknown values will be considered as a separate category in the dataset.

1.2.3 DATA PARTITION

In this section we distribute the data set into Training and Validation. Dataset is partitioned into train and validation is because it helps in managing the quality of model. This is done because, when a model is run on a dataset, the same result cannot be expected when another researcher is trying to fit in the same model in the same dataset. To ensure this doesn't happen, the same model is run on a randomly split train and validation dataset While splitting dataset we have to ensure that validation dataset is almost equal as training data set or else, it might lead to erroneous results while evaluating the models. Therefore, data has been divided as 50% for training and 50% for validation. Partition summary is as shown below:



Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0

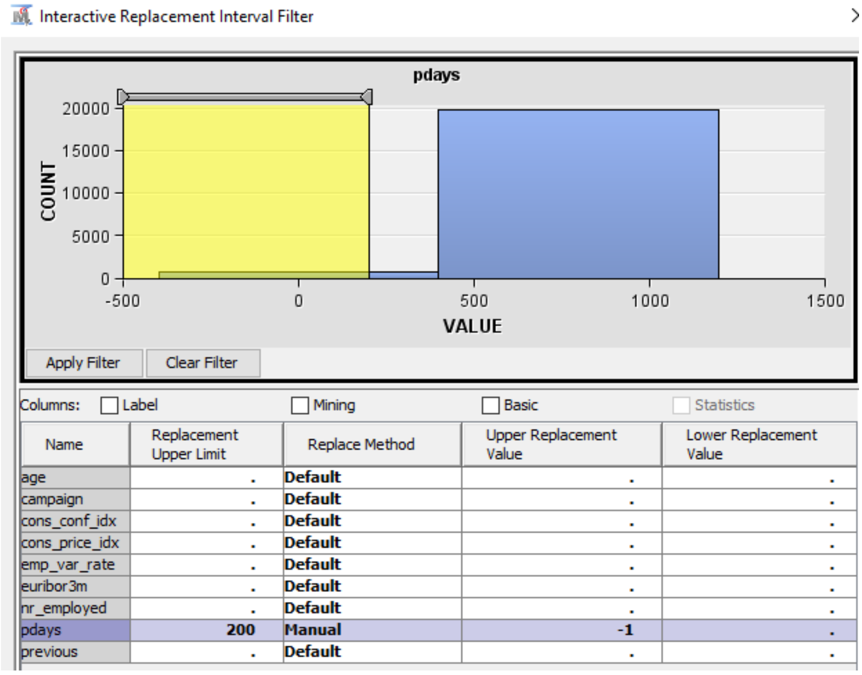
```
Partition Summary

                                          Number of
Type           Data Set                  Observations

DATA        EMWS3.Stat_TRAIN                41188
TRAIN       EMWS3.Part_TRAIN                20593
VALIDATE    EMWS3.Part_VALIDATE             20595
```

Figure 1.5 Data partition

1.2.4 DATA REPLACEMENT

The variable pdays has a vlue “999” which indicates that a client was never previously contacted. This is an unusually large value for a variable and this value is present in 96.3% of the pdays observation, which will highly bias the variable’s mean and standard deviation, due to its large value.

To avoid that, I have replaced the value “999” with “-1” to indicate the customer was not previously contacted, using the Replacement node.

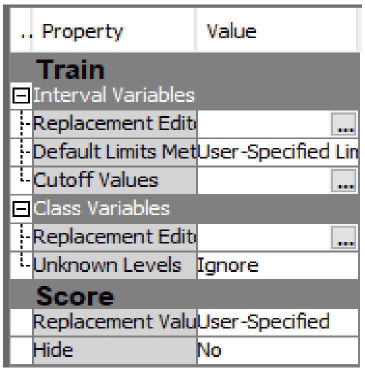


Columns: □ Label □ Mining □ Basic □ Statistics

Name	Replacement Upper Limit	Replace Method	Upper Replacement Value	Lower Replacement Value
age	.	Default	.	.
campaign	.	Default	.	.
cons_conf_idx	.	Default	.	.
cons_price_idx	.	Default	.	.
emp_var_rate	.	Default	.	.
euribor3m	.	Default	.	.
nr_employed	.	Default	.	.
pdays	200	Manual	-1	.
previous	.	Default	.	.

Figure 1.6 Data Replacement

The upper replacement limit is set to 200 and all the values of “999” will be replaced with -1, as set using the Upper Replacement Value.



Property	Value
Train	
Interval Variables	
Replacement Edit	...
Default Limits Met	User-Specified Lin
Cutoff Values	...
Class Variables	
Replacement Edit	...
Unknown Levels	Ignore
Score	
Replacement Valu	User-Specified
Hide	No

Figure 1.7

The Default Limits Method and Replacement Value properties are set to User-Specified.

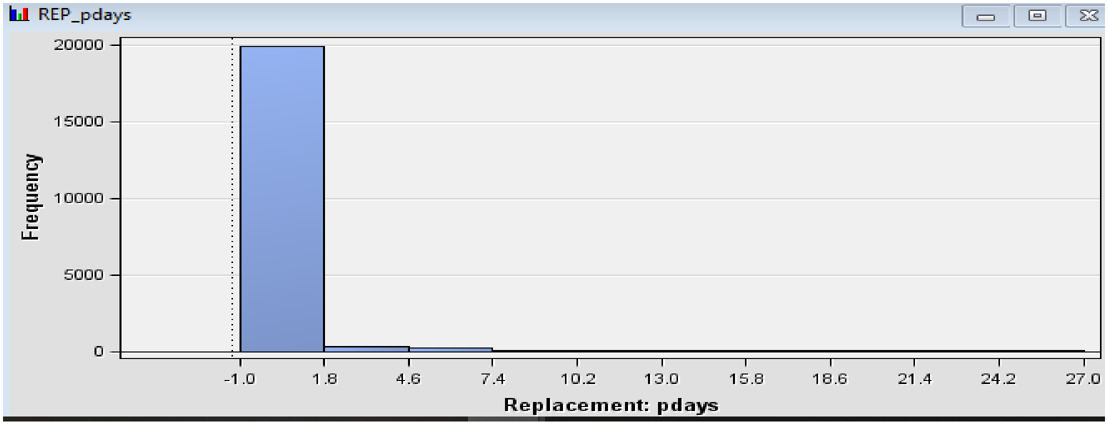
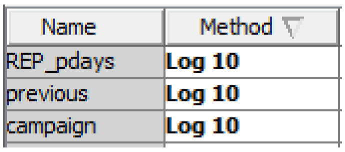


Figure 1.8 Histogram of transformed pdays

This is the resulting distribution of the replaced pdays variable. The minimum value is -1 and maximum value is 27.0. Totally 39673 observations were replaced with the -1 value. The replaced pdays variable is still highly skewed and have high kurtosis and hence it will be transformed further.

1.2.5 DATA TRANSFORMATION

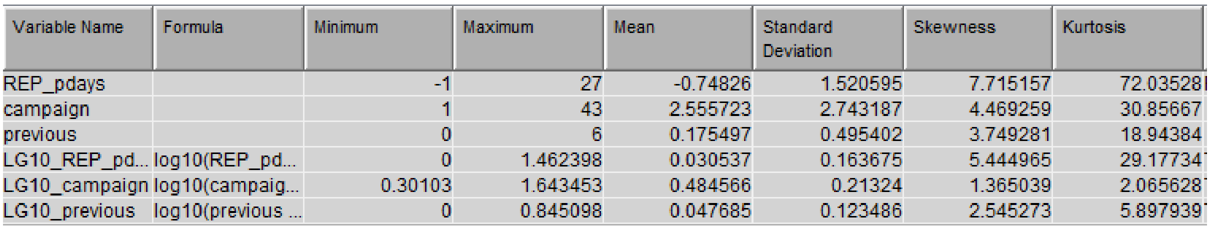
Transforming the data can improve model response as it stabilizes variance, and removes non-linearity and non-normality. This leads to better model fit.
The variables REP_pdays, previous and campaign are transformed using Log 10 method, as these three variables have high skew and kurtosis.



Name	Method
REP_pdays	**Log 10**
previous	**Log 10**
campaign	**Log 10**

Table 1.5

As seen in Table 1.6, the skew and kurtosis values of campaign variable are now normal. The skew of previous variable is normal and the kurtosis is approaching normality. The skew and kurtosis of pdays variable is still not normal, but have considerably reduced after the LOG 10 transformation.



Variable Name	Formula	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
REP_pdays		-1	27	-0.74826	1.520595	7.715157	72.03528
campaign		1	43	2.555723	2.743187	4.469259	30.85667
previous		0	6	0.175497	0.495402	3.749281	18.94384
LG10_REP_pd...	log10(REP_pd...	0	1.462398	0.030537	0.163675	5.444965	29.17734
LG10_campaign	log10(campaig...	0.30103	1.643453	0.484566	0.21324	1.365039	2.065628
LG10_previous	log10(previous ...	0	0.845098	0.047685	0.123486	2.545273	5.897939

Table 1.6

Thus, this ends the data preparation and data cleaning part of the analysis. This is the summary of what all were done:

- The variables loan and housing were removed on the basis that they were not statistically significant with the target variable y, based on the Chi-Square Plot table.
- Dataset was partitioned into training and validation in the ratio 50:50.
- The large value of 999 of variable pdays was replaced with the value -1.
- The variables pdays, campaign and previous were transformed using LOG 10 transformation method, as the skew and kurtosis of these variables were not normal.

2. MODELING

Modeling is the process of using data mining and probability techniques to predict outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. A data mining model is an empty object until it is processed. When a mode is processed, the data that is cached by the structure is passed through a filter, if a filter is defined in the model, and is analyzed by the algorithm. The algorithm computes a set of summary statistics that describes the data, identifies the rules and patterns within the data, and then uses these rules and patterns to populate the model. After it has been processed, the mining model contains a plethora of information about the data and the patterns found through analysis, including statistics, rules, and regression formulas.

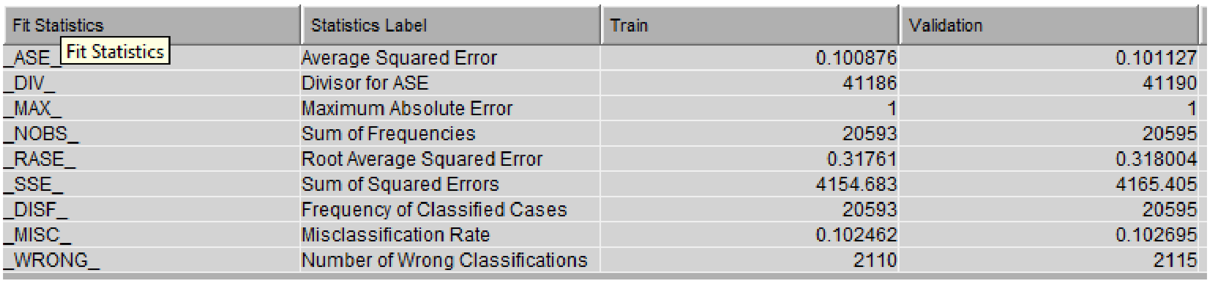
For the purpose of this study Gradient Boosting, HP SVM, Logistic Regression, Decision Tree, Auto Neural and Neural Network models are used.

It is advisable to use a control point node before and after using the data mining models. It simplifies the flow of diagram by reducing the number of connections that would otherwise be used in absence of a control point node. The Control Point node performs the same data flow distribution to the models, with lesser number of connections which makes the models look clean and less messy.

2.1 HP SVM MODEL

A support vector machine (SVM) is a supervised machine-learning method that is used to perform classification and regression analysis. The standard SVM model solves binary classification problems that produce non-probability output (only sign +1/-1) by constructing a set of hyperplanes that maximize the margin between two classes. The HP SVM Node supports only binary classification problems, thus our target variable is converted to type binary. Input variables with missing values are ignored during model training.

RESULT



Fit Statistics

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.100876	0.101127
DIV	Divisor for ASE	41186	41190
MAX	Maximum Absolute Error	1	1
NOBS	Sum of Frequencies	20593	20595
RASE	Root Average Squared Error	0.31761	0.318004
SSE	Sum of Squared Errors	4154.683	4165.405
DISF	Frequency of Classified Cases	20593	20595
MISC	Misclassification Rate	0.102462	0.102695
WRONG	Number of Wrong Classifications	2110	2115

Table 2.1.1

It is evident that the SVM model has a large misclassification rate of 0.1.

2.2 DECISION TREE

An empirical tree represents a segmentation of the data that is created by applying a series of simple rules. Each rule assigns an observation to a segment based on the value of one input. One rule is applied after another, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node. A node with all its successors forms a branch of the node that created it. The final nodes are called leaves. For each leaf, a decision is made and applied to all observations in the leaf. In predictive modeling, the decision is the predicted value. Decision trees produce a set of rules that can be used to generate predictions for a new data set. This information can then be used to drive business decisions.

An advantage of the Decision Tree node over other modeling nodes, such as the Neural Network node, is that it produces output that describes the scoring model with interpretable Node Rules. Another advantage of the Decision Tree node is the treatment of missing data. The search for a splitting rule uses the missing values of an input. Surrogate rules are available as backup when missing data prohibits the application of a splitting rule.

The parameter setting for splitting rule, leaf node and subtree details are changed for the purpose of this analysis. The value of maximum depth of tree should be changed to 10. This allows SAS to train tree up to 10 generations of splits to build more complex tree. Leaf size is changed to 8 which means it requires minimum of eight training observations for performing the modeling. Finally, the number of surrogate rules is set to 4. This feature helps when variable liable for splitting rule is having missing value to use four surrogate rules to determine the splits in non-leaf nodes.

RESULT:

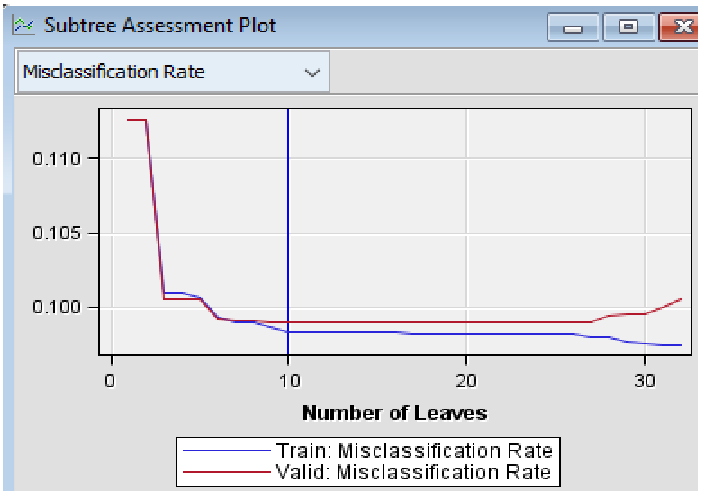


Figure 2.2.1

The model with the lowest misclassification rate comes from Step 10.

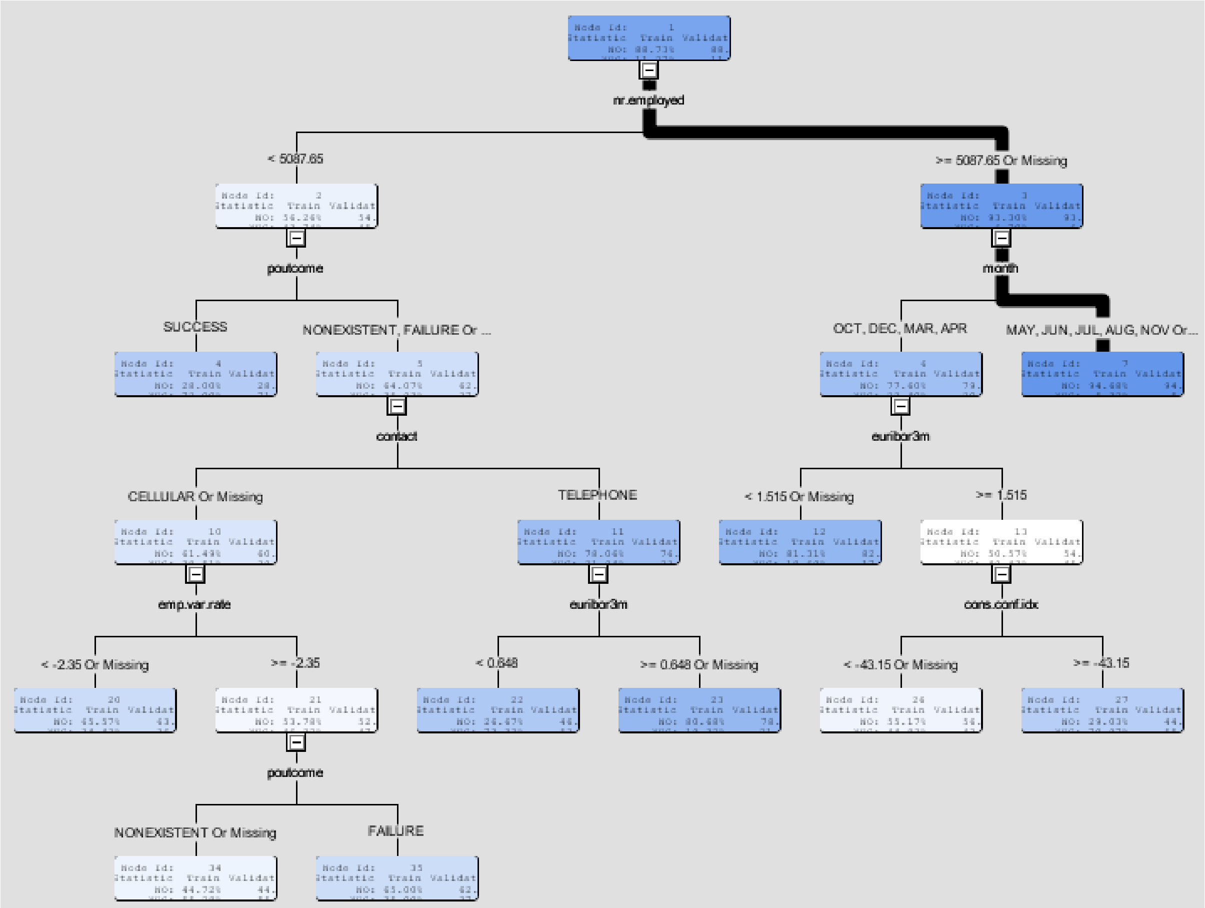
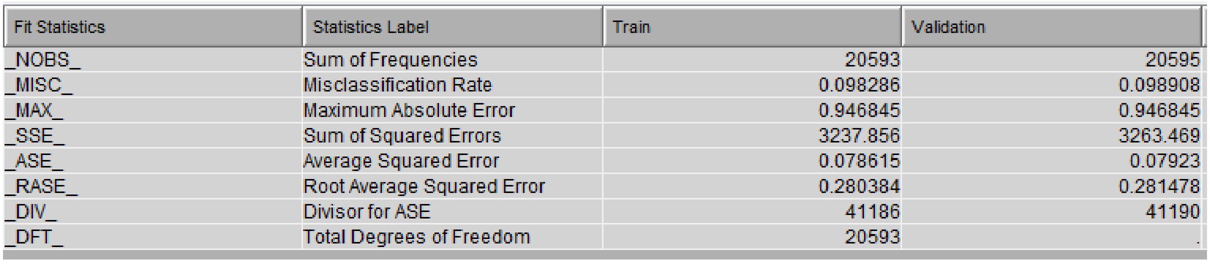


Figure 2.2.2 Decision Tree



Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20593	20595
MISC	Misclassification Rate	0.098286	0.098908
MAX	Maximum Absolute Error	0.946845	0.946845
SSE	Sum of Squared Errors	3237.856	3263.469
ASE	Average Squared Error	0.078615	0.07923
RASE	Root Average Squared Error	0.280384	0.281478
DIV	Divisor for ASE	41186	41190
DFT	Total Degrees of Freedom	20593	.

Table 2.2.1 Fit statistics

The Misclassification Rate for decision tree 0.098, which is almost closer to 1.

2.3 LOGISTIC REGRESSION

Logistic regression attempts to predict the probability that a binary or ordinal target will acquire the event of interest as a function of one or more independent inputs. The Regression node uses either a logit, complementary log-log, or probit link function and a binomial distribution error function for a logistic regression analysis.

RESULT

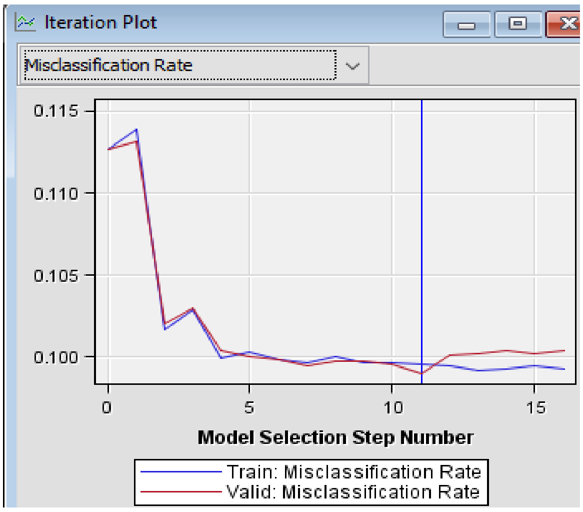
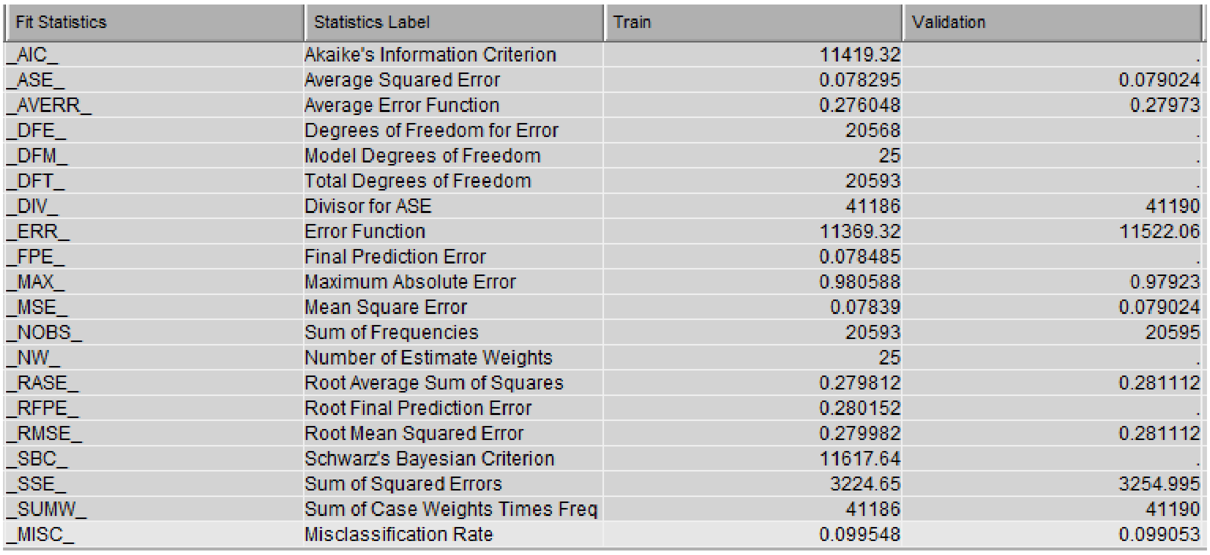


Figure 2.3.1 Iteration Plot

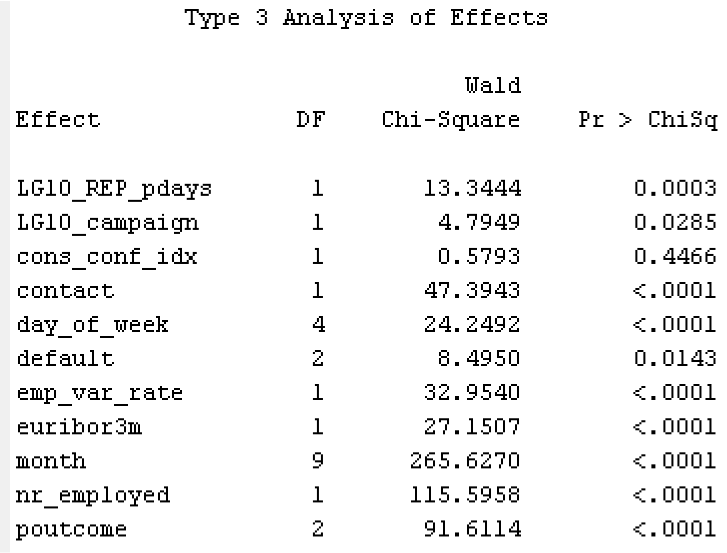
The model with the lowest misclassification rate comes from Step 11.



Fit Statistics	Statistics Label	Train	Validation
AIC	Akaike's Information Criterion	11419.32	.
ASE	Average Squared Error	0.078295	0.079024
AVERR	Average Error Function	0.276048	0.27973
DFE	Degrees of Freedom for Error	20568	.
DFM	Model Degrees of Freedom	25	.
DFT	Total Degrees of Freedom	20593	.
DIV	Divisor for ASE	41186	41190
ERR	Error Function	11369.32	11522.06
FPE	Final Prediction Error	0.078485	.
MAX	Maximum Absolute Error	0.980588	0.97923
MSE	Mean Square Error	0.07839	0.079024
NOBS	Sum of Frequencies	20593	20595
_NW_	Number of Estimate Weights	25	.
RASE	Root Average Sum of Squares	0.279812	0.281112
RFPE	Root Final Prediction Error	0.280152	.
RMSE	Root Mean Squared Error	0.279982	0.281112
SBC	Schwarz's Bayesian Criterion	11617.64	.
SSE	Sum of Squared Errors	3224.65	3254.995
SUMW	Sum of Case Weights Times Freq	41186	41190
MISC	Misclassification Rate	0.099548	0.099053

Table 2.3.1 Fit Statistics

Misclassification rate for Logistic regression is 0.099



```
             Type 3 Analysis of Effects

                                    Wald
Effect              DF    Chi-Square    Pr > ChiSq

LG10_REP_pdays       1       13.3444        0.0003
LG10_campaign        1        4.7949        0.0285
cons_conf_idx        1        0.5793        0.4466
contact              1       47.3943        <.0001
day_of_week          4       24.2492        <.0001
default              2        8.4950        0.0143
emp_var_rate         1       32.9540        <.0001
euribor3m            1       27.1507        <.0001
month                9      265.6270        <.0001
nr_employed          1      115.5958        <.0001
poutcome             2       91.6114        <.0001
```

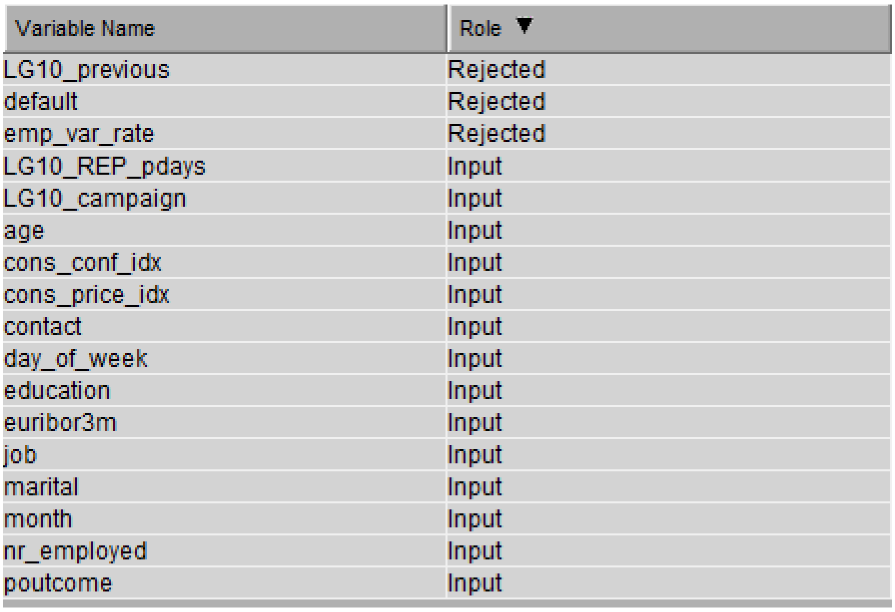
Figure 2.3.2 Analysis of Effects

All the variables except cons_conf_idx (p-value = 0.45) are statistically significant with p-value < 0.05.

2.4 NEURAL NETWORK

Neural networks are especially useful for prediction problems where: no mathematical formula is known that relates inputs to outputs, prediction is more important than explanation and there is a lot of training data. Common applications of neural networks include credit risk assessment, direct marketing, and sales prediction.

The Neural Network model ignores missing values, therefore imputing should be done before proceeding with this model. Variable selection is done for Neural Network model alone, so that it gives better results. In variable selection, input variables with low Chi-square values are rejected.



Variable Name	Role ▼
LG10_previous	Rejected
default	Rejected
emp_var_rate	Rejected
LG10_REP_pdays	Input
LG10_campaign	Input
age	Input
cons_conf_idx	Input
cons_price_idx	Input
contact	Input
day_of_week	Input
education	Input
euribor3m	Input
job	Input
marital	Input
month	Input
nr_employed	Input
poutcome	Input

Table 2.4.1

The variables previous, default and emp_var_rate are rejected because of low chi-square values.

RESULT

In the parameter setting for Neural Network, Multilayer Perceptron is most widely used architecture for Neural Network Model. Number of hidden layers are set to 5.

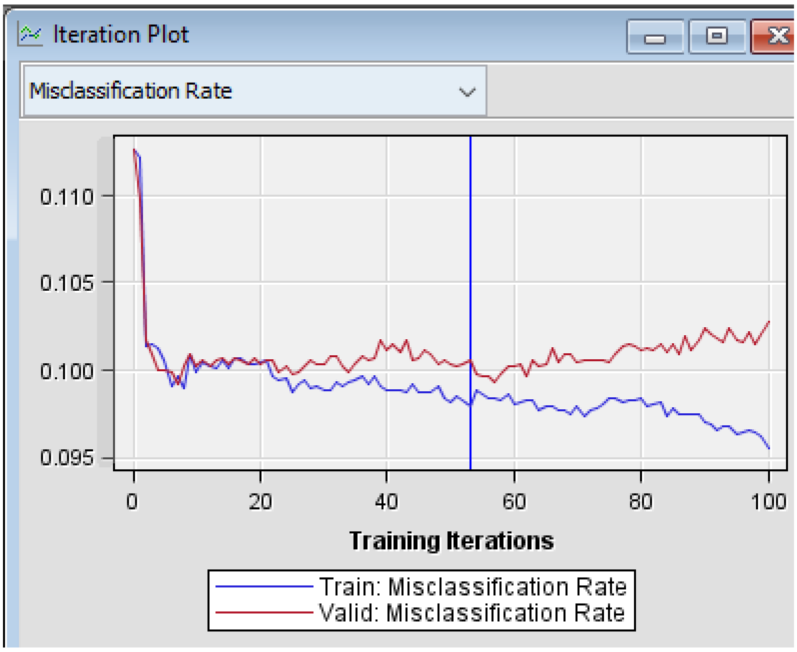
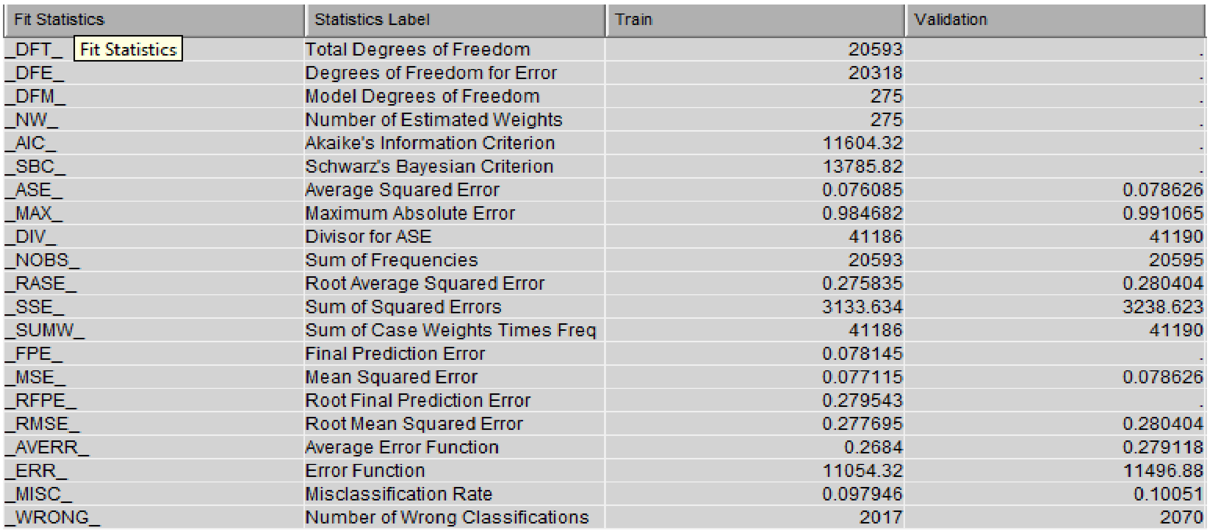


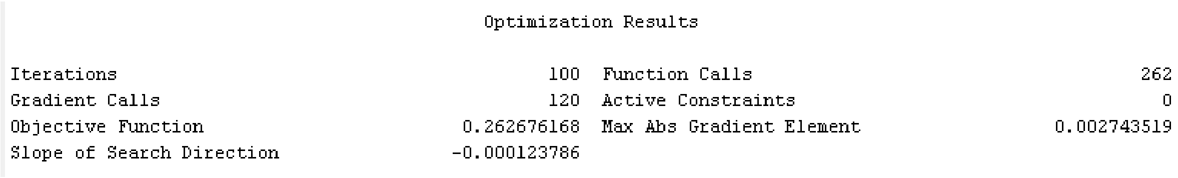
Figure 2.4.1

The model with the lowest misclassification rate comes from Step 53.



Fit Statistics	Statistics Label	Train	Validation
DFT Fit Statistics	Total Degrees of Freedom	20593	.
DFE	Degrees of Freedom for Error	20318	.
DFM	Model Degrees of Freedom	275	.
_NW_	Number of Estimated Weights	275	.
AIC	Akaike's Information Criterion	11604.32	.
SBC	Schwarz's Bayesian Criterion	13785.82	.
ASE	Average Squared Error	0.076085	0.078626
MAX	Maximum Absolute Error	0.984682	0.991065
DIV	Divisor for ASE	41186	41190
NOBS	Sum of Frequencies	20593	20595
RASE	Root Average Squared Error	0.275835	0.280404
SSE	Sum of Squared Errors	3133.634	3238.623
SUMW	Sum of Case Weights Times Freq	41186	41190
FPE	Final Prediction Error	0.078145	.
MSE	Mean Squared Error	0.077115	0.078626
RFPE	Root Final Prediction Error	0.279543	.
RMSE	Root Mean Squared Error	0.277695	0.280404
AVERR	Average Error Function	0.2684	0.279118
ERR	Error Function	11054.32	11496.88
MISC	Misclassification Rate	0.097946	0.10051
WRONG	Number of Wrong Classifications	2017	2070

Table 2.4.1



```
                                          Optimization Results

Iterations                                  100   Function Calls                                  262
Gradient Calls                              120   Active Constraints                                0
Objective Function                  0.262676168   Max Abs Gradient Element                0.002743519
Slope of Search Direction          -0.000123786
```

Figure 2.4.2

2.5 AUTO NEURAL NETWORK

The AutoNeural node conducts limited searches in order to find better network configurations. In this, Hidden nodes are added one at a time. A node may contain one or more neurons.

The following properties are changed: Train Action is set to Search. This configures the AutoNeural node to sequentially increase the network complexity. Set the Number of Hidden Units to 1, With this option, each iteration adds one hidden unit. Set Tolerance to Low, This prevents preliminary training from occurring. Select Direct to No, This deactivates direct connections between the inputs and the target. Select Normal to No, This deactivates the normal distribution activation function.

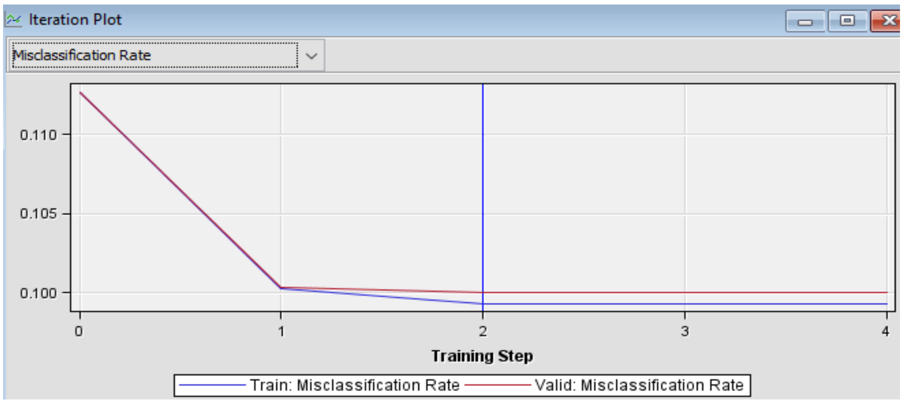
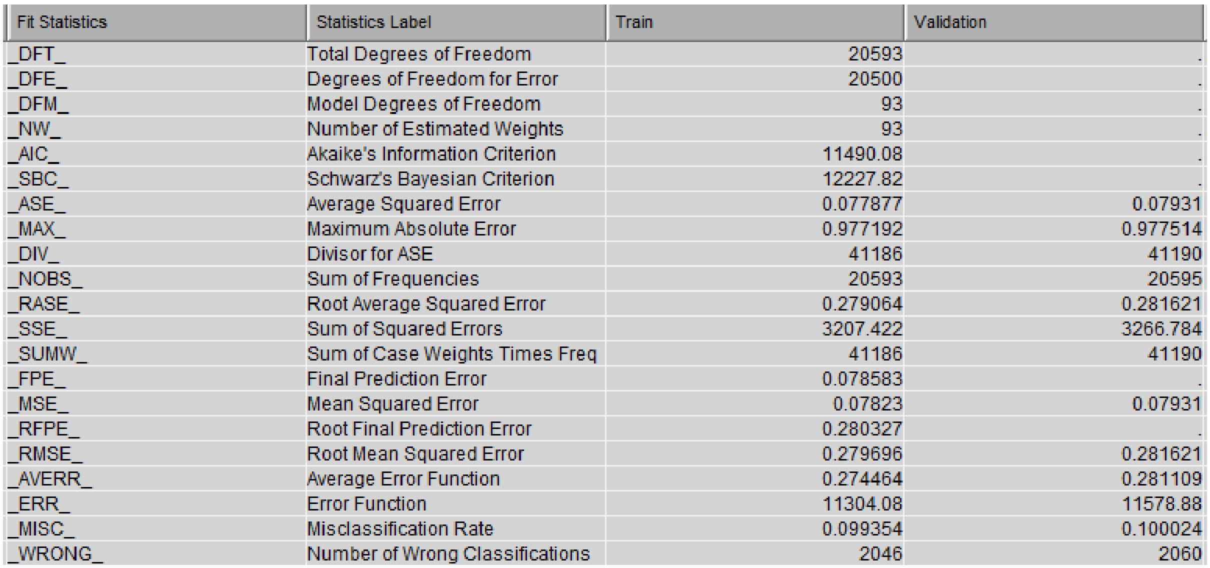


FIGURE 2.5.1 Iteration Plot

The model with the lowest misclassification rate comes from Step 2 and remains stable after that.



Fit Statistics	Statistics Label	Train	Validation
DFT	Total Degrees of Freedom	20593	.
DFE	Degrees of Freedom for Error	20500	.
DFM	Model Degrees of Freedom	93	.
_NW_	Number of Estimated Weights	93	.
AIC	Akaike's Information Criterion	11490.08	.
SBC	Schwarz's Bayesian Criterion	12227.82	.
ASE	Average Squared Error	0.077877	0.07931
MAX	Maximum Absolute Error	0.977192	0.977514
DIV	Divisor for ASE	41186	41190
NOBS	Sum of Frequencies	20593	20595
RASE	Root Average Squared Error	0.279064	0.281621
SSE	Sum of Squared Errors	3207.422	3266.784
SUMW	Sum of Case Weights Times Freq	41186	41190
FPE	Final Prediction Error	0.078583	.
MSE	Mean Squared Error	0.07823	0.07931
RFPE	Root Final Prediction Error	0.280327	.
RMSE	Root Mean Squared Error	0.279696	0.281621
AVERR	Average Error Function	0.274464	0.281109
ERR	Error Function	11304.08	11578.88
MISC	Misclassification Rate	0.099354	0.100024
WRONG	Number of Wrong Classifications	2046	2060

TABLE 2.5.1 Fit Statistics

Misclassification rate is 0.099. Almost same as other models.

2.6 GRADIENT BOOSTING

Gradient boosting is a boosting approach that resamples the analysis data set several times to generate results that form a weighted average of the re-sampled data set. Tree boosting creates a series of decision trees which together form a single predictive model. A tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function.

Interval targets define the residual using the squared error loss function. To compute the residual for an interval target using squared error loss you simply subtract the predicted value from the target value. Binary targets define the residual using the negative binomial log-likelihood loss function. The negative binomial log-likelihood loss function is also called logistic loss. Each time the data is used to grow a tree and the accuracy of the tree is computed. The successive samples are adjusted to accommodate previously computed inaccuracies.

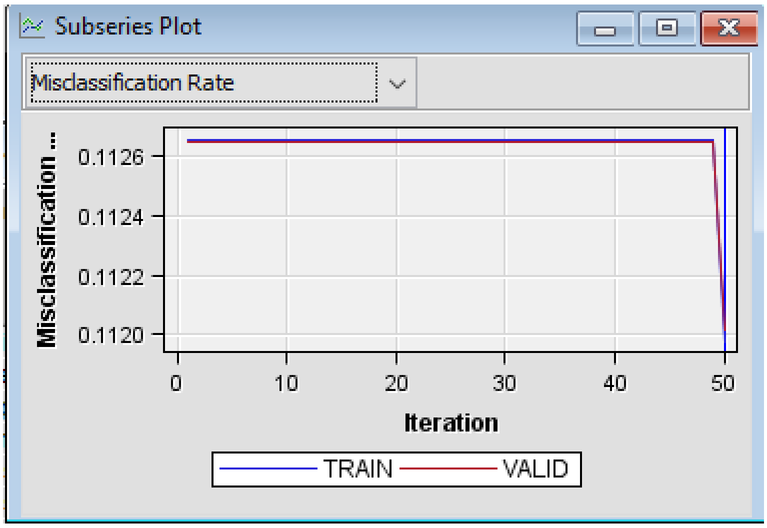
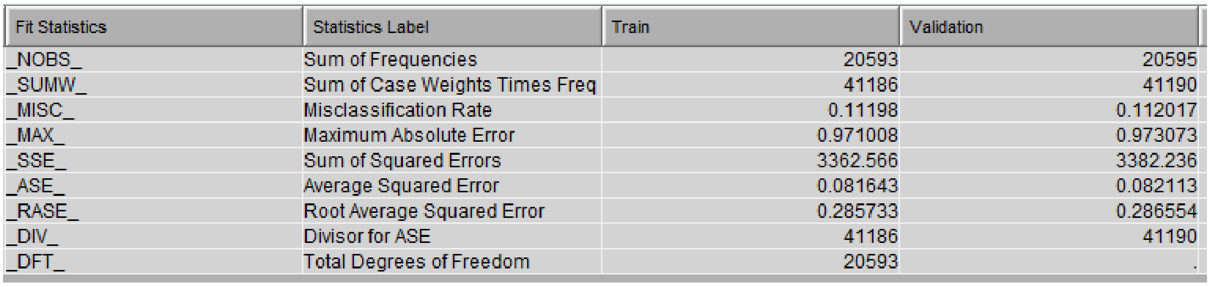


Figure 2.5.1



Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20593	20595
SUMW	Sum of Case Weights Times Freq	41186	41190
MISC	Misclassification Rate	0.11198	0.112017
MAX	Maximum Absolute Error	0.971008	0.973073
SSE	Sum of Squared Errors	3362.566	3382.236
ASE	Average Squared Error	0.081643	0.082113
RASE	Root Average Squared Error	0.285733	0.286554
DIV	Divisor for ASE	41186	41190
DFT	Total Degrees of Freedom	20593	.

Table 2.5.1

This model has the largest misclassification so far, with 0.11 value.

3. MODEL COMPARISON

After running all the models, the next step is to evaluate and compare which model produces the best result. Model Evaluation is done with the help of Model Comparison Node. We again use the Control Point node before comparison, by connecting all the outputs of the models to Control Point node and then connect it with the model comparison node.

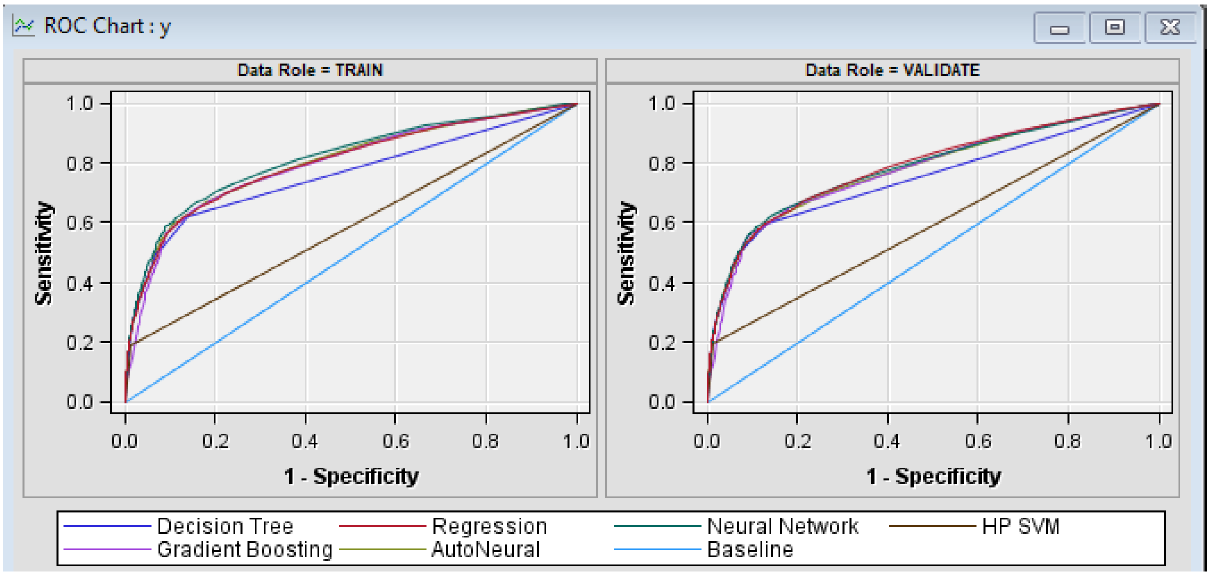
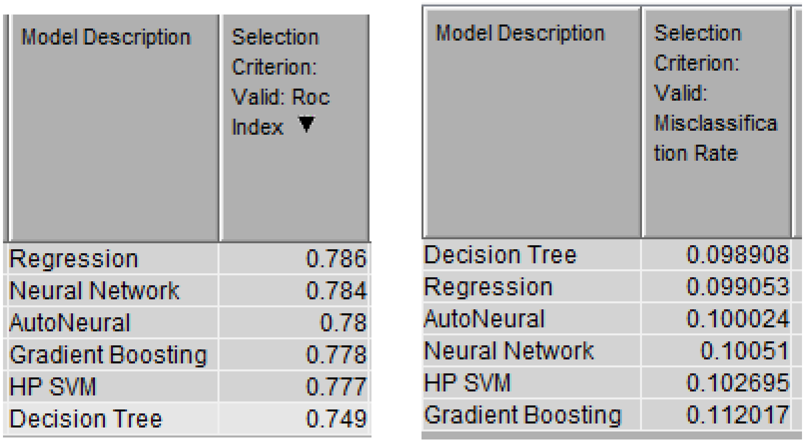


Figure 3.1: ROC Curve

Minute changes can be observed in curves for Train and Validate data sets which means models are not over fitted.



Model Description	Selection Criterion: Valid: Roc Index ▼
Regression	0.786
Neural Network	0.784
AutoNeural	0.78
Gradient Boosting	0.778
HP SVM	0.777
Decision Tree	0.749

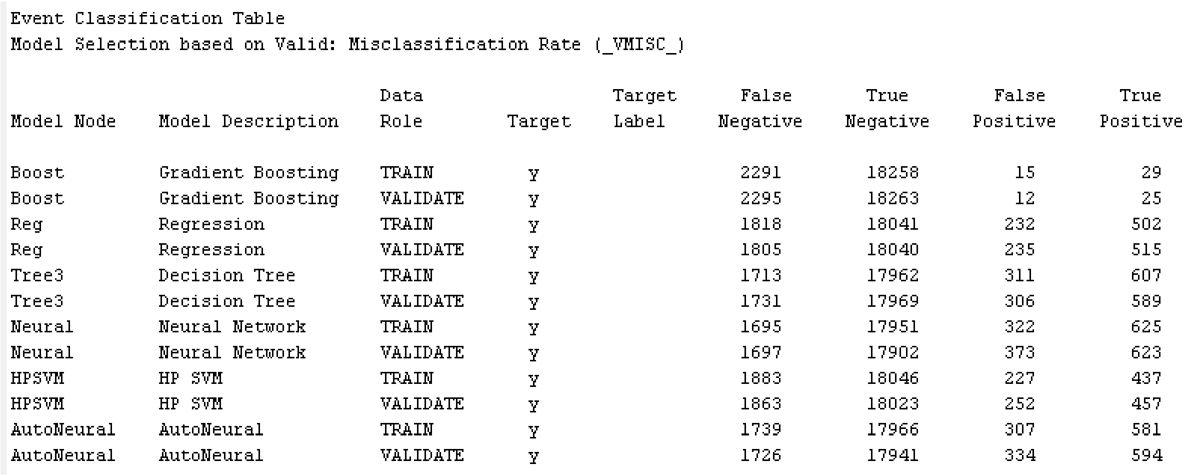
Model Description	Selection Criterion: Valid: Misclassification Rate
Decision Tree	0.098908
Regression	0.099053
AutoNeural	0.100024
Neural Network	0.10051
HP SVM	0.102695
Gradient Boosting	0.112017

Table 3.1 ROC Index and Misclassification Rate

In Fit Statistics we consider the Misclassification Rate to check the performance of the model. The misclassification value should be on lower side to illustrate that the model is performing better. Misclassification Rate is the number of times a decision was wrongly made. From Table 3.1, we can see that Decision Tree Model has the lowest Misclassification Rate (0.098908) for Validation data set, this proves that Decision Tree is the best model among the 6. While, Gradient Boosting has the highest Misclassification Rate (0.112017).

Considering upon another parameter - ROC Index, Regression Model performs better with a value of 0.786.

Thus, in accordance with the obtained misclassification rate, we can conclude positively that the decision tree performs the best for this dataset and for this study.



```
Event Classification Table
Model Selection based on Valid: Misclassification Rate (_VMISC_)

                                  Data                 Target     False      True      False      True
Model Node    Model Description   Role        Target   Label    Negative   Negative   Positive   Positive

Boost         Gradient Boosting   TRAIN         y                  2291      18258         15         29
Boost         Gradient Boosting   VALIDATE      y                  2295      18263         12         25
Reg           Regression          TRAIN         y                  1818      18041        232        502
Reg           Regression          VALIDATE      y                  1805      18040        235        515
Tree3         Decision Tree       TRAIN         y                  1713      17962        311        607
Tree3         Decision Tree       VALIDATE      y                  1731      17969        306        589
Neural        Neural Network      TRAIN         y                  1695      17951        322        625
Neural        Neural Network      VALIDATE      y                  1697      17902        373        623
HPSVM         HP SVM              TRAIN         y                  1883      18046        227        437
HPSVM         HP SVM              VALIDATE      y                  1863      18023        252        457
AutoNeural    AutoNeural          TRAIN         y                  1739      17966        307        581
AutoNeural    AutoNeural          VALIDATE      y                  1726      17941        334        594
```

Figure 3.2 Confusion Matrix

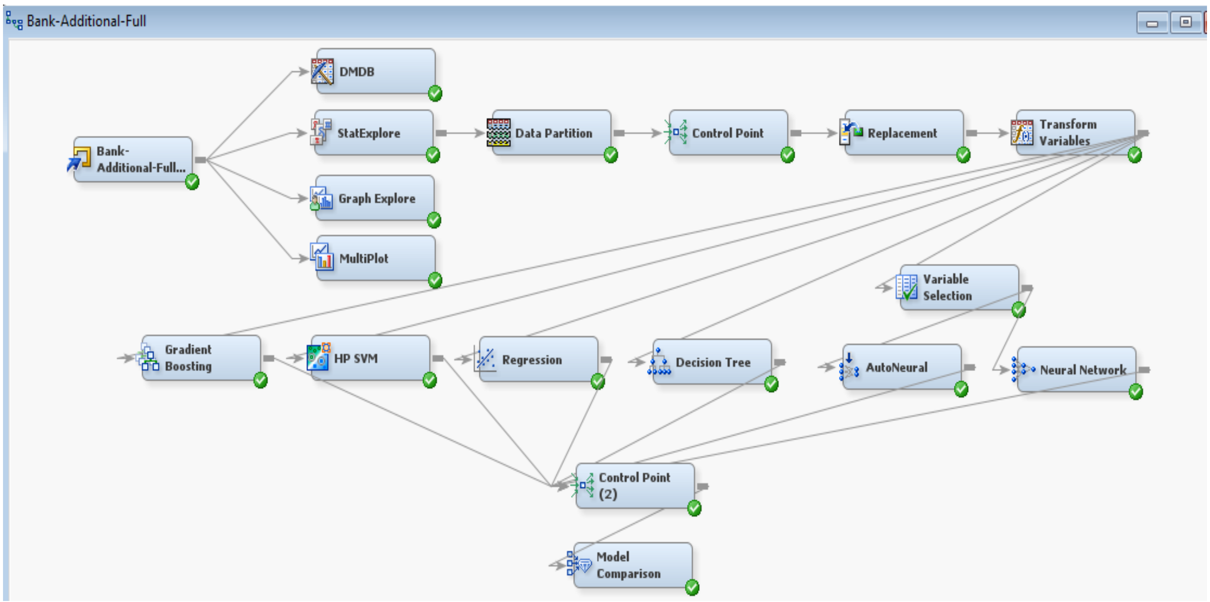


Figure 3.3 Model diagram

3.1 MODEL COMPARISON WITH ORIGINAL RESEARCH:

The original research *"A Data-Driven Approach to Predict the Success of Bank Telemarketing"* was performed by S. Moro, P. Cortez and P. Rita on this same dataset. Four data mining models were performed: Decision tree, Logistic regression, support vector machine and neural network and were compared using the following two metrics: Area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT).

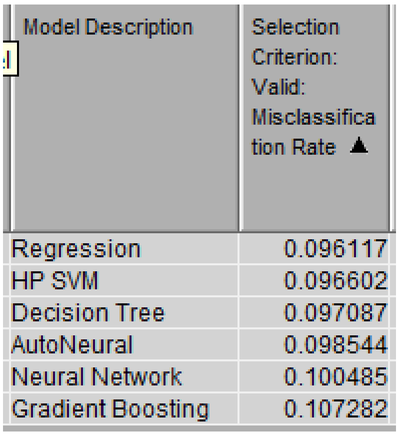
In both the metrics, Neural Network performed the best results for their research with an AUC of 0.80 and ALIFT of 0.67. They were able to get a good prediction rate that 79% of the successful selling can be achieved when contacting only half of the clients.

Based on the result obtained from this study on the misclassification rate metric, Decision Tree performs the best. Thus, the results of this study are not consistent with the original research, which concludes that Neural Network performs the best.

3.2 MODEL COMPARISON WITH SAMPLE DATASET:

The sample dataset contains 10% sample (4,119 records) of the entire original data set. The smallest dataset is provided to test computationally demanding machine learning algorithms.

For the purpose of this study, we are just going to compare the model results obtained using the sample dataset and the full dataset. The same data preparation, data partition, data cleaning rules that were applied to the original dataset were applied to the sample dataset, so that there won't be any difference.



Model Description	Selection Criterion: Valid: Misclassification Rate ▲
Regression	0.096117
HP SVM	0.096602
Decision Tree	0.097087
AutoNeural	0.098544
Neural Network	0.100485
Gradient Boosting	0.107282

Table 3.2: Misclassification rate of sample validation dataset

In the sample dataset, Regression model performs well with the smallest Misclassification rate of 0.096. This rate is even smaller than the misclassification rate found using the original dataset (DECISION TREE --> 0.098908). Thus, Regression model found using the sample dataset is the best model compared to the Decision tree model found using the original full dataset.

CONCLUSION:

Thus from this research we can conclude that Decision Tree performs the best among all the other models.

REFERENCES:

[1] https://data.oecd.org/emp/employment-rate.htm

[2] https://www.investopedia.com/terms/c/cci.asp

[3] https://www.investopedia.com/terms/c/consumerpriceindex.asp

[4] https://www.global-rates.com/interest-rates/euribor/euribor-interest-3-months.aspx

[5] Moro, S., Cortez, P. and Rita, P. (2018). Decision Support Systems. [online] Journals.elsevier.com. Available at: https://www.journals.elsevier.com/decision-support-systems [Accessed 12 Dec. 2018].

STUDENT NAME: Sangita Sriram

STUDENT NUMBER: D17129392

COURSE: DT228A Full Time

CLASS: MSc Computing Data Analytics

YEAR: 2018-2019

ASSIGNMENT: Data Mining

LECTURER NAME: Brendan Tierney