DUBLIN INSTITUTE OF TECHNOLOGY

Data Mining
Assignment 1

Submitted By:

Student Name: Deepshikha Wadikar

Student ID: D17128916

Course: Data Analytics

Prog Code: DT228A (Full Time)

Academic Year: 2018-19 (1st Year,1st Sem)

Submitted to:

Prof. Brendan Tierney

1. INTRODUCTION

The main objective of this report is to provide the summary of the Data Mining Assignment related to a marketing campaign for a Portuguese banking institution. The purpose is to identify best fit model to predict which customers are most likely to subscribe to a term deposit account. All the tasks are performed using SAS Enterprise Miner. The various tasks performed are:

- Descriptive Analytics and summary
- Data Preparation
- Data Mining algorithm description
- Building different Data Models

In this bank additional data set there are 21 variables, the outcome variable is denoted by ‘y’ which is a categorical variable with the value as 0=no; 1=yes. In this report we are evaluating all the variables present in the data set.

2. DESCRIPTIVE ANALYTICS

The bank dataset is uploaded into SAS Enterprise Miner using File Import node. The ‘semi-colon’ is the delimiter used and using Import File option the file is imported.

.. Property	Value
Variables	...
Import File	C:\Users\Dell\Downlo ...
Maximum Rows to Impor	1000000
Maximum Columns to Imp	10000
Delimiter	;
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	csv

The bank additional full data contains 21 variables and 41,188 observations. The variables in the dataset are shown below with the appropriate roles and levels.

NAME	ROLE	LEVEL	DESCRIPTION
age	INPUT	INTERVAL	Age of the customer
campaign	INPUT	INTERVAL	number of contacts performed during this campaign and for this client (numeric, includes last contact)
cons_conf_idx	INPUT	INTERVAL	consumer confidence index - monthly indicator
cons_price_idx	INPUT	INTERVAL	consumer price index - monthly indicator
contact	INPUT	NOMINAL	contact communication type (categorical: 'cellular','telephone')
day_of_week	INPUT	NOMINAL	last contact day of the week
default	INPUT	NOMINAL	has credit in default? (categorical: 'no','yes','unknown')
duration	REJECTED	INTERVAL	last contact duration, in seconds (numeric).
education	INPUT	NOMINAL	education
emp_var_rate	INPUT	INTERVAL	employment variation rate - quarterly indicator
euribor3m	INPUT	INTERVAL	euribor 3 month rate - daily indicator
housing	INPUT	NOMINAL	has housing loan? (categorical: 'no','yes','unknown')
job	INPUT	NOMINAL	type of job
loan	INPUT	NOMINAL	has personal loan? (categorical: 'no','yes','unknown')
marital	INPUT	NOMINAL	marital status
month	INPUT	NOMINAL	last contact month of year
nr_employed	INPUT	INTERVAL	number of employees - quarterly indicator
pdays	INPUT	INTERVAL	number of days that passed by after the client was last contacted from a previous campaign
poutcome	INPUT	NOMINAL	outcome of the previous marketing campaign
previous	INPUT	INTERVAL	number of contacts performed before this campaign and for this client
y	TARGET	BINARY	has the client subscribed a term deposit?

Table 2.1: Portuguese Data set Variable Summary

As shown in the above table, ‘y’ is the Target variable. It is represented in a binary format (0=no; 1=yes) to decide if the client is interested in term deposit or not. Then, the duration variable displays the last call duration, in seconds. This variable is rejected because this is the call duration time and this will affect output variable only if the call is picked so to build a realistic predictive model this variable is rejected.

Descriptive Analysis is the preliminary stage of data processing to predict the desired outcome variable.
The StatExplore node is used to perform the basic descriptive analytics.The below mentioned figure shows the plots for different variables.

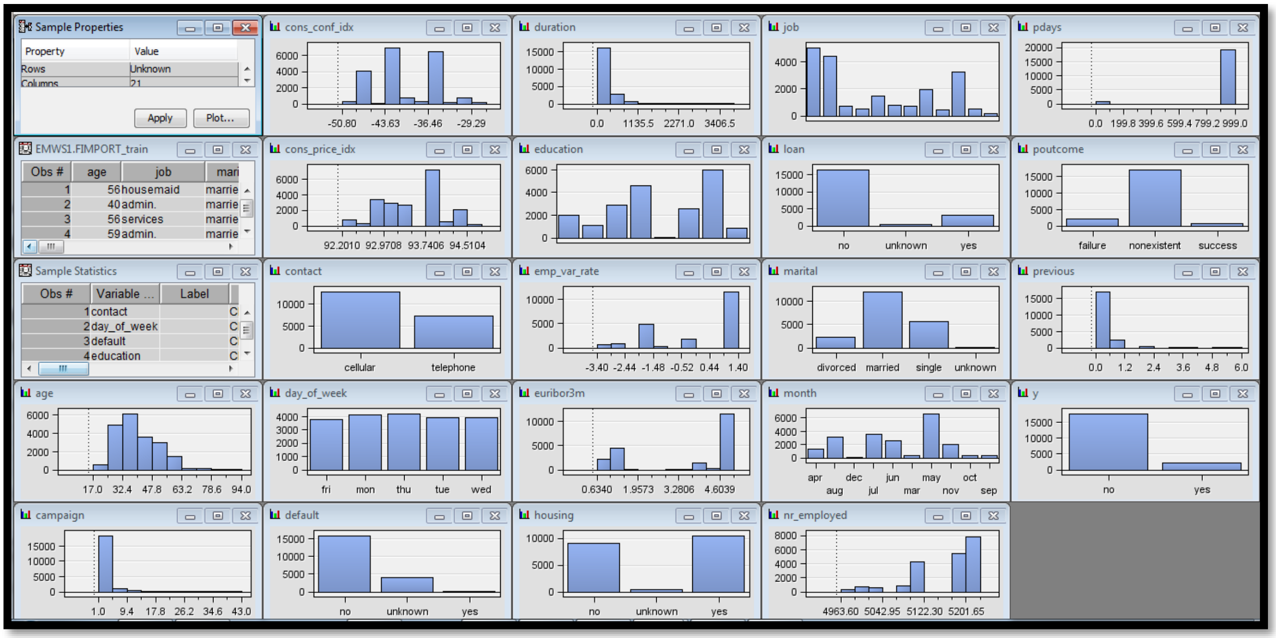


Figure 2.1: StatExplore for Variables

Chi –Square test is also performed to get the significance of each variable with the Target variable. The StatExplore node output displays the Chi-Square plot and Variable Worth plot which shows the significance of variable in deriving the output.

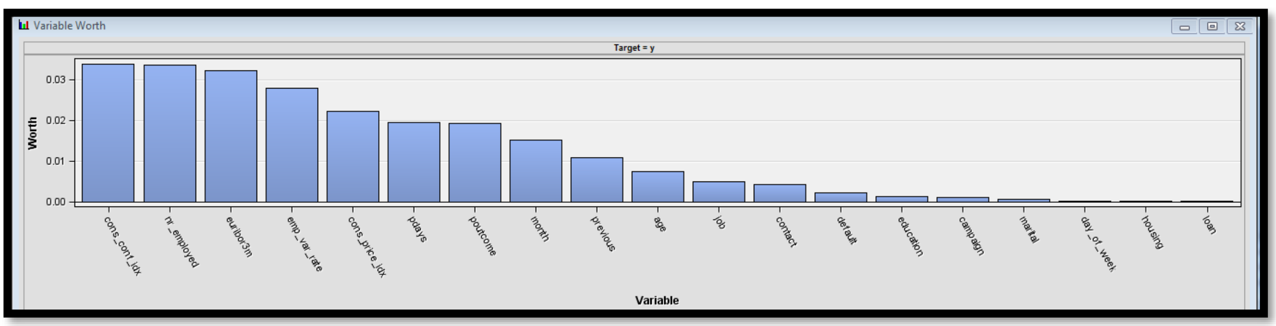


Figure 2.2: Variables Worth Plot

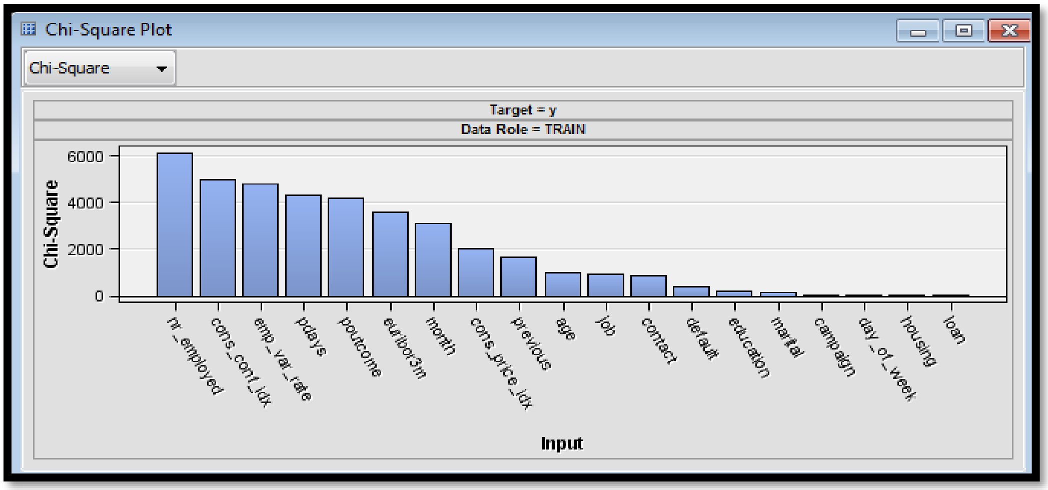


Figure 2.3 Chi-Square Plot

The below pie chart represents the output variable ‘y’. From the below pie chart we can say that 11% of people are interested to subscribe for the term deposit whereas 89% customers are not interested to subscribe for the term deposit.

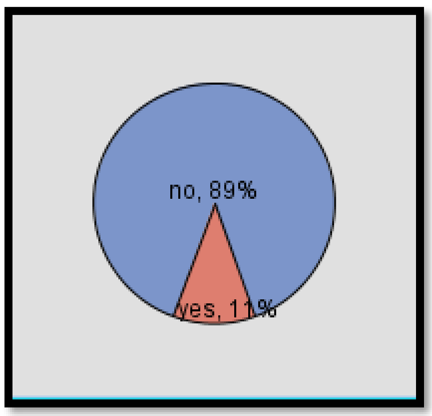


Figure 2.4: Target Variable- Term Deposit (Yes or No)

Further, few more variables contact, day_of_week and duration are analyzed with the Target variable. The below graph shows the percent of no and yes for different categories.

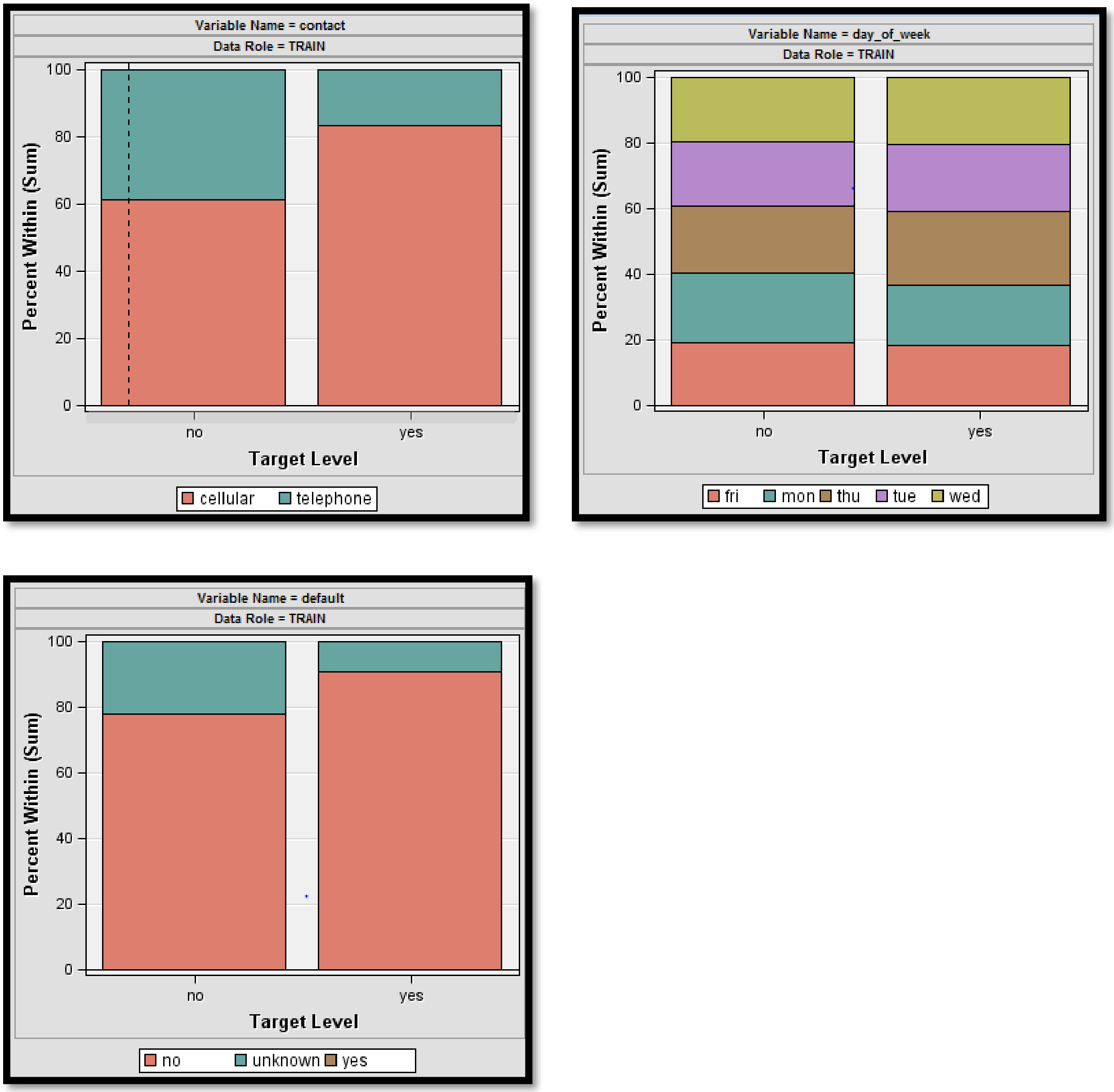


Figure 2.5: Portuguese Data set Variable Summary

The Variable Selection node is used to select the desired variables. The Target Model is selected as 'Chi-Square' in the Variable Selection node property as the output variable is binary.After running the Variable Selection node in the outcome the Variable selection result is generated with the variable and its reason for rejection. Below figures will display the configuration settings of variable selection node and output.

Property	Value
Train	
Variables	...
Max Class Level	100
Max Missing Percentage	50
Target Model	Chi-Square

Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
age	Input	Interval	Numeric		
campaign	Input	Interval	Numeric		
cons_conf_idx	Input	Interval	Numeric	cons.conf.idx	
cons_price_idx	Input	Interval	Numeric	cons.price.idx	
contact	Input	Nominal	Character		
day_of_week	Input	Nominal	Character		
default	Rejected	Nominal	Character		Varsel:Small Chi-square value
education	Input	Nominal	Character		
emp_var_rate	Rejected	Interval	Numeric	emp.var.rate	Varsel:Small Chi-square value
euribor3m	Input	Interval	Numeric		
housing	Rejected	Nominal	Character		Varsel:Small Chi-square value
job	Input	Nominal	Character		
loan	Rejected	Nominal	Character		Varsel:Small Chi-square value
marital	Rejected	Nominal	Character		Varsel:Small Chi-square value
month	Input	Nominal	Character		
nr_employed	Input	Interval	Numeric	nr.employed	
pdays	Input	Interval	Numeric		
poutcome	Input	Nominal	Character		
previous	Rejected	Interval	Numeric		Varsel:Small Chi-square value

Figure 2.6: Configuration Settings and Output for Variable Selection Node

3. HANDLING MISSING VALUES

In the given bank data set there are missing values present in some categorical variables which are all coded as an 'unknown' category. So, considering these variables missing values i.e., unknown value as category.

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	contact	INPUT	2	0	cellular	63.47	telephone	36.53
TRAIN	day_of_week	INPUT	5	0	thu	20.94	mon	20.67
TRAIN	default	INPUT	3	0	no	79.12	unknown	20.87
TRAIN	education	INPUT	8	0	university.degree	29.54	high.school	23.10
TRAIN	housing	INPUT	3	0	yes	52.38	no	45.21
TRAIN	job	INPUT	12	0	admin.	25.30	blue-collar	22.47
TRAIN	loan	INPUT	3	0	no	82.43	yes	15.17
TRAIN	marital	INPUT	4	0	married	60.52	single	28.09
TRAIN	month	INPUT	10	0	may	33.43	jul	17.42
TRAIN	poutcome	INPUT	3	0	nonexistent	86.34	failure	10.32
TRAIN	y	TARGET	2	0	no	88.73	yes	11.27

Also in other variables no missing data are found as displayed in the below figure.

Figure 3.1 Details of Missing Values

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
age	INPUT	40.02406	10.42125	41188	0	17	38	98	0.784697	0.791312
campaign	INPUT	2.567593	2.770014	41188	0	1	2	56	4.762507	36.9798
cons_conf_idx	INPUT	-40.5026	4.628198	41188	0	-50.8	-41.8	-26.9	0.30318	-0.35856
cons_price_idx	INPUT	93.57566	0.57884	41188	0	92.201	93.749	94.767	-0.23089	-0.82804
emp_var_rate	INPUT	0.081886	1.57096	41188	0	-3.4	1.1	1.4	-0.7241	-1.06263
euribor3m	INPUT	3.621291	1.734447	41188	0	0.634	4.857	5.045	-0.70919	-1.4068
nr_employed	INPUT	5167.036	72.25153	41188	0	4963.6	5191	5228.1	-1.04426	-0.00366
pdays	INPUT	962.4755	186.9109	41188	0	0	999	999	-4.92219	22.22946
previous	INPUT	0.172963	0.494901	41188	0	0	0	7	3.832042	20.10882

Figure 3.2 Summary Statistics Output of Interval variables

4. DATA PARTITION

After descriptive analysis and missing value exploration in data set the data is partitioned into Training and Validation data set. While, splitting the data set the validation data set should be larger than the Training data set to evaluate the decision tree correctly. Therefore, data is partitioned in 50-50% to avoid the erroneous results in decision tree evaluation. The partition summary is shown in below screenshot.

Data Set Allocations	
Training	50.0
Validation	50.0
Test	0.0

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Stat2_TRAIN	41188
TRAIN	EMWS1.Part_TRAIN	20593
VALIDATE	EMWS1.Part_VALIDATE	20595

Figure 4.1: Data Partition Configuration Settings and Output

5. TRANSFORMING VARIABLES

The Transform operation is performed to make the interval variables skew and kurtosis to be in the statistically significant range of +/-2. There are few variables which are considered here are not normal. The variables campaign and pdays both have the skew values outside the range of +/-2. So, by using log 10 transformation for campaign variable it is transformed statistically significant. However, the variable pdays is not getting transformed and is approaching towards normality. The below figure displays the output from Transform Variables node.

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	campaign		.	20593	0	1	43	2.555723	2.743187	4.469259	30.85667	
Input	Original	pdays		.	20593	0	0	999	963.8999	183.3585	-5.03278	23.3316	
Output	Computed	EXP_pdays	exp(pdays / 9....	.	20593	0	1	2.688E43	2.593E43	4.964E42	-5.03264	23.32969	Transformed p...
Output	Computed	LG10_campaign	log10(campaig...	.	20593	0	0.30103	1.643453	0.484566	0.21324	1.365039	2.065628	Transformed c...

Figure 5.1: Output of Transform Variable Node

6. DATA MODELLING

After, the Data preparation the next step is Data Modelling. In this step the analyzed data is used and applied into a data mining algorithm to build different models. These models help us to predict the outcome variable. Data modelling is the process by which a model is created to predict an outcome. If, the outcome is categorical it is called classification and if the outcome is numeric then it is called as regression.

Control Point node is used to simplify the matrix of connections from different modelling nodes and the model comparison node

Here, the following models are discussed:

6.1 Maximal Tree (Interactive Decision Tree)

Decision Trees are predictive modelling techniques. These models are able to handle the missing data on their own.

Maximal Decision Tree is created using Decision Tree node present in Model tab. To run the decision tree interactively we will split the node through Interactive option present in the Decision Tree Properties panel. Then simply selecting the Train Node option will create a maximal tree using the logworth value. The Misclassification Rate plot shown below for the Maximal Tree is shown below. In this the number of leaves node as 19 and Valid misclassification rate as – 0.1 and Train misclassification rate as- 0.995. Here there is not much difference in the misclassification rate. However, it is generally considered that maximal tree will not be the best model as it is over optimized so it depicts probably less accuracy rate. The smaller tree is better to get the best accurate results.

Train	
Variables	
Interactive	
Import Tree Model	No
Tree Model Data Set	
Use Frozen Tree	Yes
Use Multiple Targets	No

Figure 6.1.1 Maximal Tree – Configuration Settings

The Use Frozen Tree option is set as ‘Yes’ so that the settings will not change for the maximal tree.

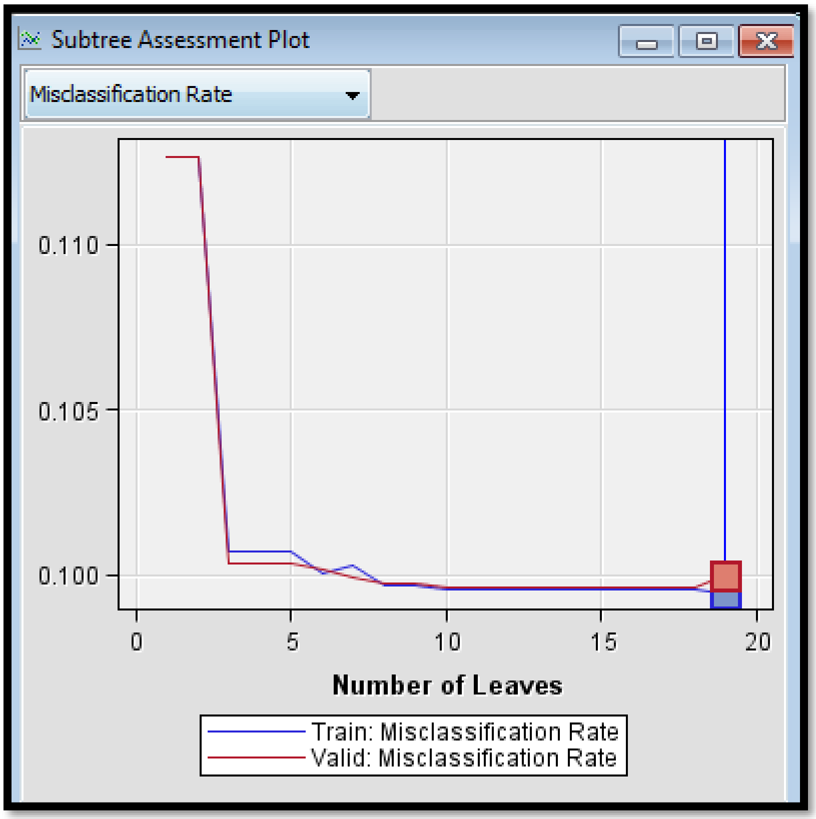


Figure 6.1.1 Maximal Tree – Misclassification Rate plot

The accuracy of the model is calculated as (True Positive + True Negative)/Total. The accuracy of the Maximal tree for Train dataset and Validate dataset is 90% and 89% respectively.

6.2 Decision Tree

The Decision Tree Model is the most widely used model and also the most powerful model to predict the result. This has the ability to predict both continuous and categorical variable for analysis. The below configurations are used to build the decision tree. The Assessment Measure is chosen as ‘Misclassification Rate’ as the output is categorical variable. By selecting the Assessment Measure as ‘Misclassification’ the tree is pruned on the basis of misclassification.

Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25

Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.

Figure 6.2.1: Decision Tree Configuration Settings

The Fit Statistics output of Decision Tree is displayed in the below figure. As by looking into the Train and validate columns we can say that there are not much difference in between Train and validate data.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y		_NOBS_	Sum of Freque...	20593	20595	.
y		_MISC_	Misclassificati...	0.097897	0.098665	.
y		_MAX_	Maximum Abs...	0.946845	0.946845	.
y		_SSE_	Sum of Square...	3232.676	3260.744	.
y		_ASE_	Average Squar...	0.07849	0.079163	.
y		_RASE_	Root Average ...	0.28016	0.28136	.
y		_DIV_	Divisor for ASE	41186	41190	.
y		_DFT_	Total Degrees ...	20593	.	.

Figure 6.2.2: Decision Tree – Fit Statistics Output

For the fit statistics the Average Squared error and Misclassification are the best statistics. The Misclassification Rate plot shows that as the depth of leaf are less the accuracy is more and as the depth of leaf node increases the accuracy decreases.

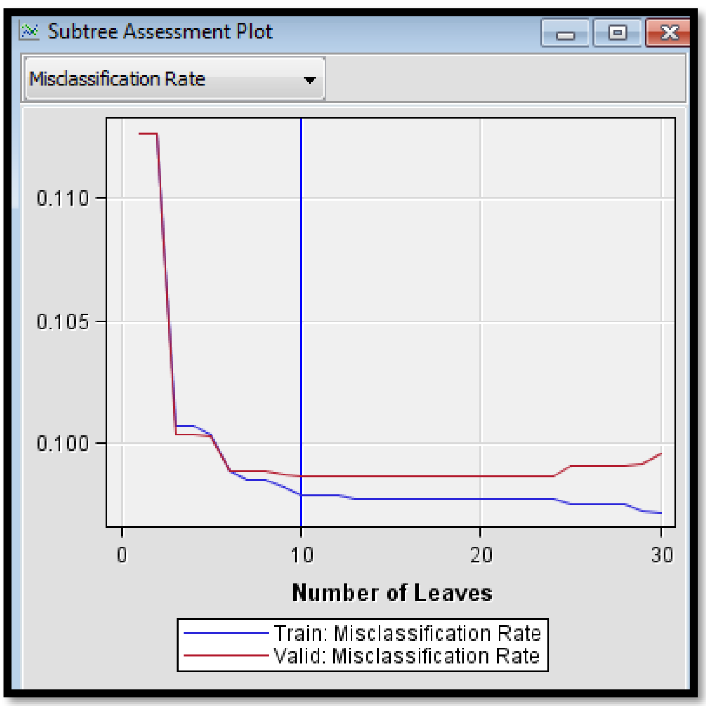


Figure 6.2.3: Decision Tree: Misclassification Rate

The accuracy of the Decision tree for both Train dataset and Validate dataset is 90%.

The below Leaf Statistics plot compares the predicted outcome percentages i.e., the Training Data output with the observed outcome percentages i.e., the validation data output. This plot shows how training data responses are getting reflected in valdation data.

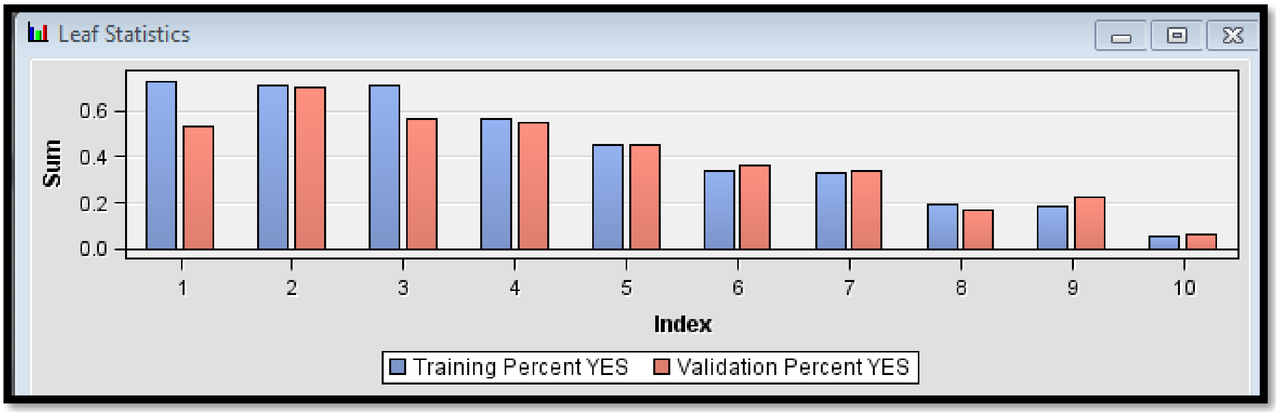
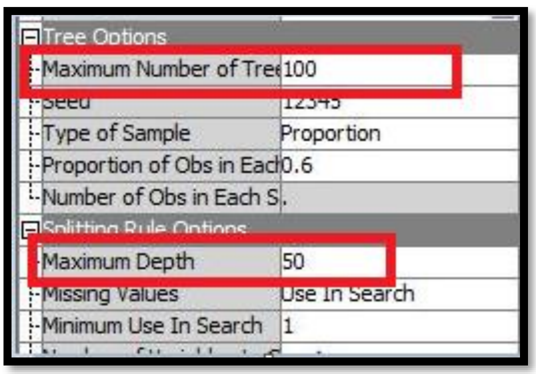


Figure 6.2.4: Decision Tree Leaf Statistics

6.3 HP Forest Model

The Forest Tree is a group of many trees. In this algorithm all the trees are running simultaneously and each tree is built at the bagged data. In this data mining algorithm many trees are trained using the subset of data and averaged together to predict the final predictability probability. It have high predictive power.

The accuracy of the HP Forest Model for both Train dataset and Validate dataset is 90%.



Tree Options	
Maximum Number of Tre	100
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Eac	0.6
Number of Obs in Each S	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1

Figure 6.3.1: HP Forest Model – Configuration Settings

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y		_ASE_	Average Sq...	0.075653	0.077105	.
y		_DIV_	Divisor for A...	41186	41190	.
y		_MAX_	Maximum A...	0.96927	0.96927	.
y		_NOBS_	Sum of Fre...	20593	20595	.
y		_RASE_	Root Avera...	0.275051	0.277678	.
y		_SSE_	Sum of Squ...	3115.844	3175.961	.
y		_DISF_	Frequency ...	20593	20595	.
y		_MISC_	Misclassific...	0.098869	0.099539	.
y		_WRONG_	Number of ...	2036	2050	.

Figure 6.3.2: HP Forest Model – Fit Statistics Output

6.4 Support Vector Machine

The Support Vector Machine data modelling is a binary classification model which construct a line that maximizes margin between two classes. The points which lies on the separator line are called as Support Vectors. HP SVM node is used in SAS enterprise miner to create the SVM model.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y		_ASE_	Average Squar...	0.101362	0.101499	.
y		_DIV_	Divisor for ASE	41186	41190	.
y		_MAX_	Maximum Abs...	0.999457	0.999403	.
y		_NOBS_	Sum of Freque...	20593	20595	.
y		_RASE_	Root Average ...	0.318374	0.318589	.
y		_SSE_	Sum of Square...	4174.687	4180.754	.
y		_DISF_	Frequency of C...	20593	20595	.
y		_MISC_	Misclassificati...	0.102365	0.102598	.
y		_WRONG_	Number of Wro...	2108	2113	.

Figure 6.4.1: SVM Model – Fit Statistics Output

The accuracy of the SVM for both Train dataset and Validate dataset is 89.7%.

6.5 Neural Network

Neural Network model is used for classification, feature mining, prediction analysis. It is a non linear statistical data modelling tool.

The result is displayed in the below graphs. In the Fit Statistics the mean squared error is displayed for Train and validation data and the values are quite close.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y		_DFT_	Total Degrees ...	20593	.	.
y		_DFE_	Degrees of Fr...	20463	.	.
y		_DFM_	Model Degree...	130	.	.
y		_NW_	Number of Est...	130	.	.
y		_AIC_	Akaike's Infor...	11366.13	.	.
y		_SBC_	Schwarz's Bay...	12397.39	.	.
y		_ASE_	Average Squar...	0.076423	0.078388	.
y		_MAX_	Maximum Abs...	0.981983	0.982453	.
y		_DIV_	Divisor for ASE	41186	41190	.
y		_NOBS_	Sum of Freque...	20593	20595	.
y		_RASE_	Root Average ...	0.276448	0.279978	.
y		_SSE_	Sum of Squar...	3147.575	3228.793	.
y		_SUMW_	Sum of Case ...	41186	41190	.
y		_FPE_	Final Predictio...	0.077394	.	.
y		_MSE_	Mean Squared...	0.076909	0.078388	.
y		_RFPE_	Root Final Pre...	0.278199	.	.
y		_RMSE_	Root Mean Sq...	0.277325	0.279978	

Figure 6.5.1: Neural Network – Fit Statistics Output

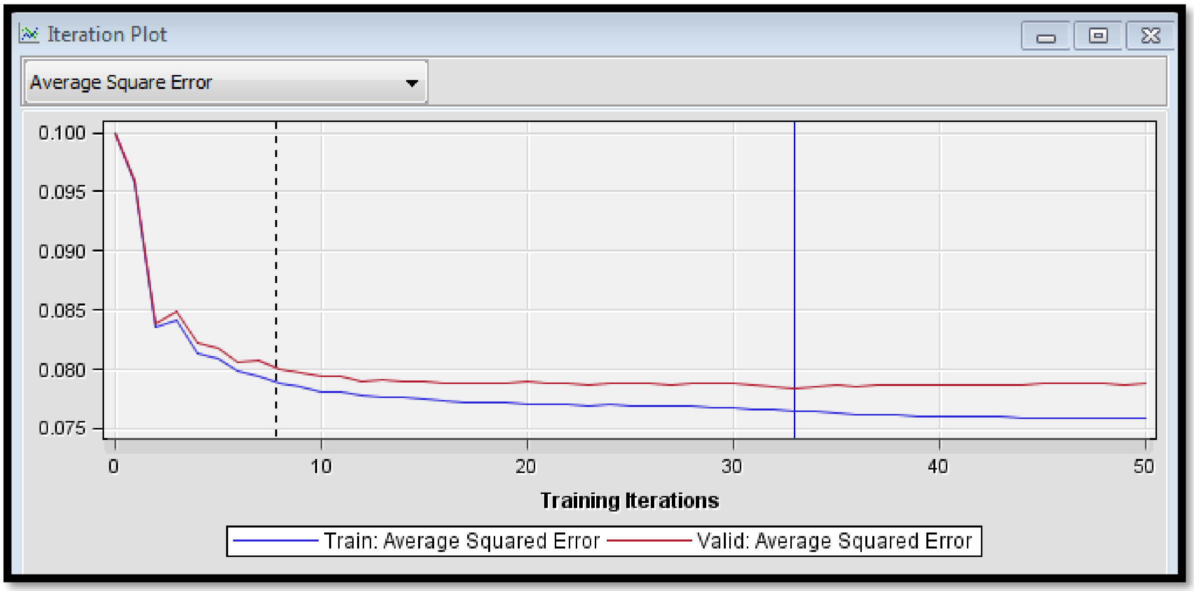


Figure 6.5.2: Neural Network – Fit Statistics Output

The accuracy of the Neural Network for both Train dataset and Validate dataset is 90%.

6.6 Regression

Regression is a data mining tool used in predictive analysis. Regression is of two types – linear regression and logistic regression. When the output variable is continuous linear regression is used and when the output is categorical then logistic regression is used. Here as the output is binary so logistic regression is used.
The logistic regression is a parametric model. The variable with value p<.0001 are statistically significant.

Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Suppress Intercept	No
Input Coding	Deviation

Figure 6.6.1: Regression Model – Configuration Settings

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
EXP_pdays	1	18.0830	<.0001
LG10_campaign	1	6.4246	0.0113
cons_conf_idx	1	18.1641	<.0001
contact	1	61.0925	<.0001
day_of_week	4	23.9913	<.0001
month	9	268.3591	<.0001
nr_employed	1	772.2305	<.0001
poutcome	2	77.9488	<.0001

Figure 6.6.2: Regression Model – Output

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
y		_AIC_	Akaike's Infor...	11448.6	.	.
y		_ASE_	Average Squar...	0.078521	0.079158	.
y		_AVERR_	Average Error ...	0.276953	0.280622	.
y		_DFE_	Degrees of Fr...	20572	.	.
y		_DFM_	Model Degree...	21	.	.
y		_DFT_	Total Degrees ...	20593	.	.
y		_DIV_	Divisor for ASE	41186	41190	.
y		_ERR_	Error Function	11406.6	11558.81	.
y		_FPE_	Final Predictio...	0.078681	.	.
y		_MAX_	Maximum Abs...	0.980823	0.97927	.
y		_MSE_	Mean Square ...	0.078601	0.079158	.
y		_NOBS_	Sum of Freque...	20593	20595	.
y		_NW_	Number of Est...	21	.	.
y		_RASE_	Root Average ...	0.280215	0.28135	.
y		_RFPE_	Root Final Pre...	0.280501	.	.
y		_RMSE_	Root Mean Sq...	0.280358	0.28135	.
y		_SBC_	Schwarz's Bay...	11615.18		

Figure 6.6.3: Regression Model – Fit Statistics Output

The accuracy of the Regression model for both Train dataset and Validate dataset is 90%.

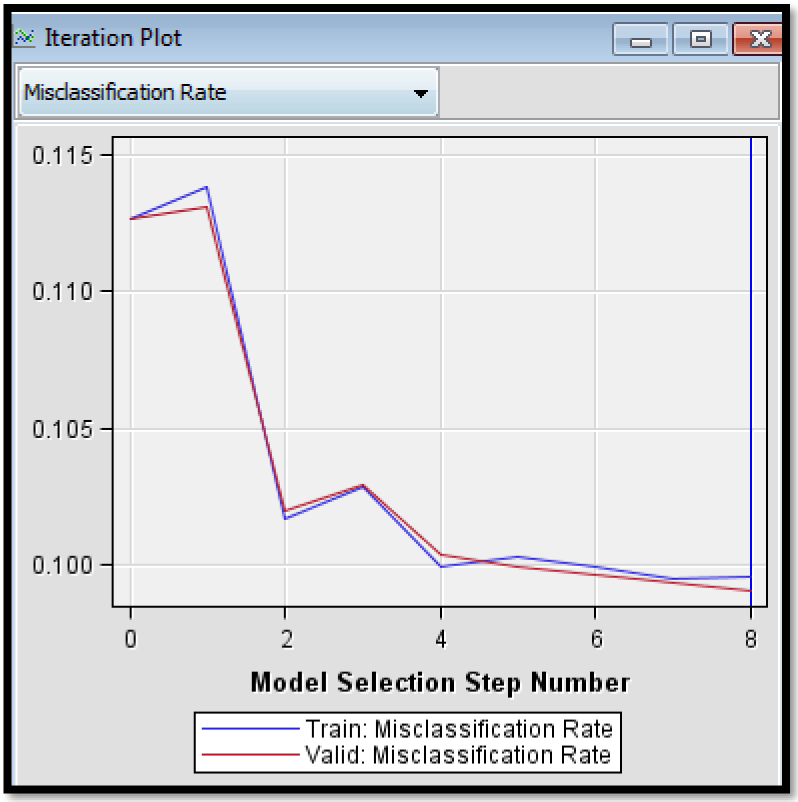


Figure 6.6.4: Regression Model – Misclassification Rate Plot

7. MODEL COMPARISON

Model Evaluation is done by using Model Comparison node. A Control Point is used in between to control the points from all the models and then control point is connected to Model Comparison node.

For this particular dataset Decision Tree is the best fit model. The below figures depict the output from model comparison node.

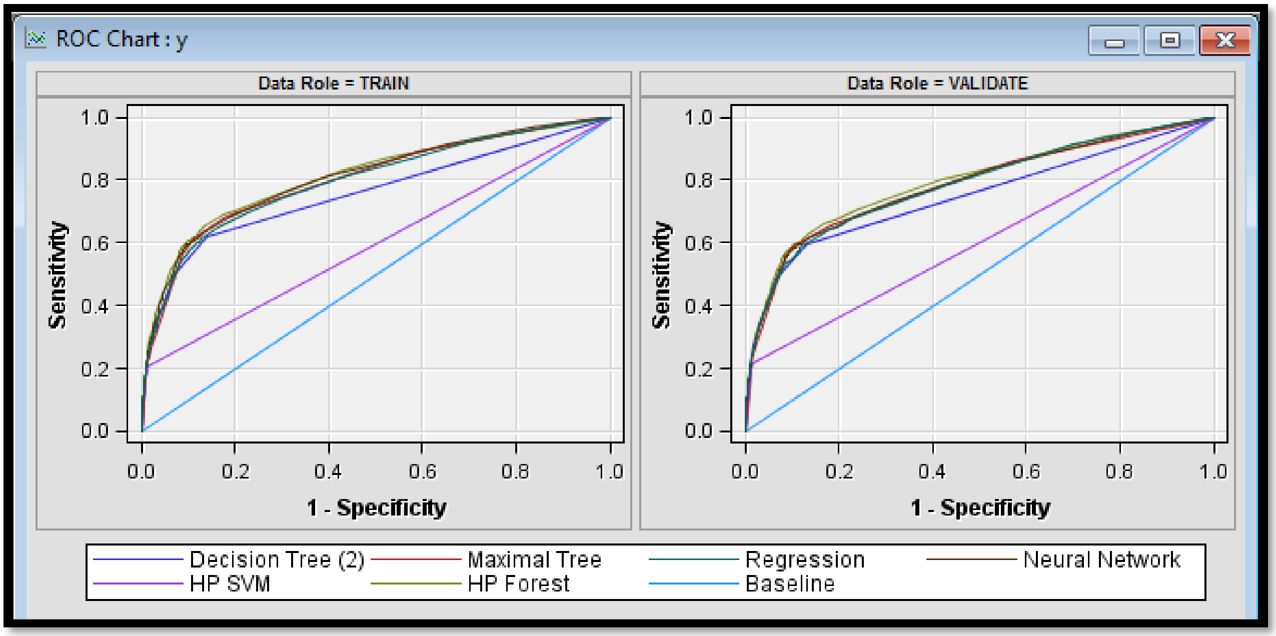


Figure 7.1: ROC Curve

The more wider the curve is better is the model.

The Cumulative Lift curve is almost close for all models.

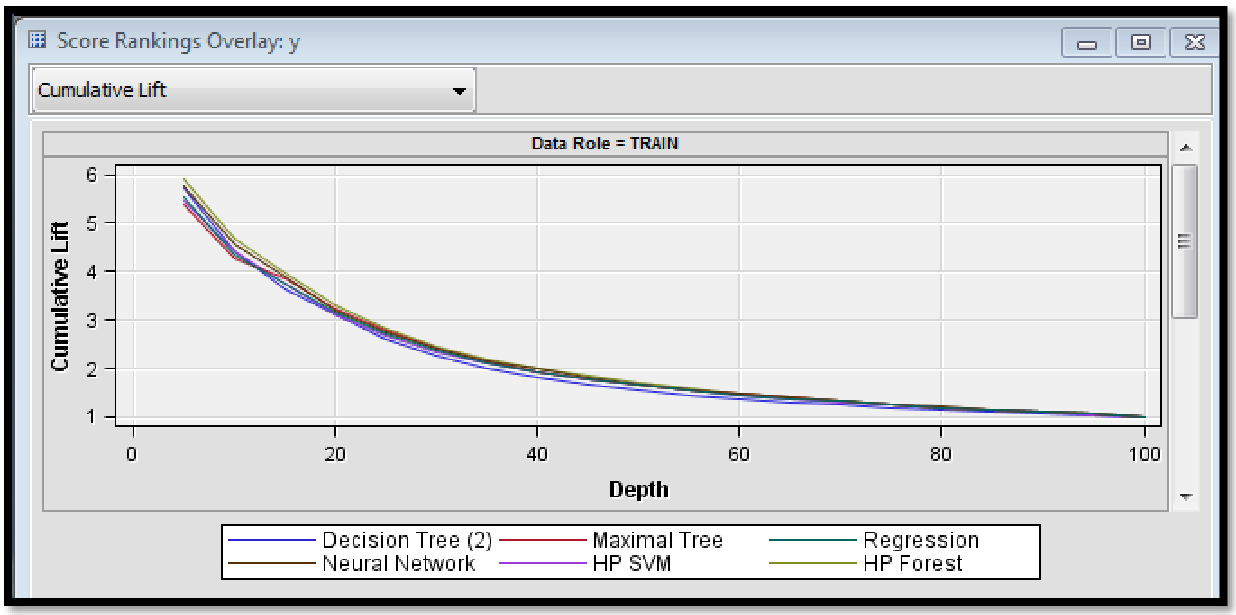


Figure 7.2: Cumulative Lift Curve

The below Fit Statistics curve shows that the Decision Tree model is the best fit model for this dataset to predict the outcome. The selection criteria is Misclassification Rate the lower the misclassification rate of the model the better the model is to predict the outcome

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Sum of Frequencies	Train: Misclassification Rate	Train: Maximum Absolute Error	Train: Sum of Squared Errors	Train: Average Squared Error	Train: Root Average Squared Error
Y	Tree2	Tree2	Decision Tr...	y		0.098665	20593	0.097897	0.946845	3232.676	0.07849	0.28016
	Neural	Neural	Neural Net...	y		0.099053	20593	0.097897	0.981983	3147.575	0.076423	0.276448
	Reg	Reg	Regression	y		0.099102	20593	0.099597	0.980823	3233.946	0.078521	0.280215
	HPDMForest	HPDMForest	HP Forest	y		0.099539	20593	0.098869	0.96927	3115.844	0.075653	0.275051
	Tree	Tree	Maximal Tree	y		0.100024	20593	0.0995	0.973202	3194.256	0.077557	0.27849
	HPSVM	HPSVM	HP SVM	y		0.102598	20593	0.102365	0.999457	4174.687	0.101362	0.318374

Figure 7.3: Model Comparison– Fit Statistics Output

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

Model Node	Model Description	Data Role	Target	Target Label	False Negative	True Negative	False Positive	True Positive
Tree2	Decision Tree (2)	TRAIN	y		1687	17944	329	633
Tree2	Decision Tree (2)	VALIDATE	y		1697	17940	335	623
HPDMForest	HP Forest	TRAIN	y		1852	18089	184	468
HPDMForest	HP Forest	VALIDATE	y		1842	18067	208	478
HPSVM	HP SVM	TRAIN	y		1850	18015	258	470
HPSVM	HP SVM	VALIDATE	y		1823	17985	290	497
Reg	Regression	TRAIN	y		1790	18012	261	530
Reg	Regression	VALIDATE	y		1769	18003	272	551
Neural	Neural Network	TRAIN	y		1720	17977	296	600
Neural	Neural Network	VALIDATE	y		1722	17957	318	598
Tree	Maximal Tree	TRAIN	y		1837	18061	212	483
Tree	Maximal Tree	VALIDATE	y		1816	18031	244	504

Figure 7.4: Model Comparison– Classification Table Output

The model workflow diagram created in SAS Enterprise Miner is shown in the below figure 7.5

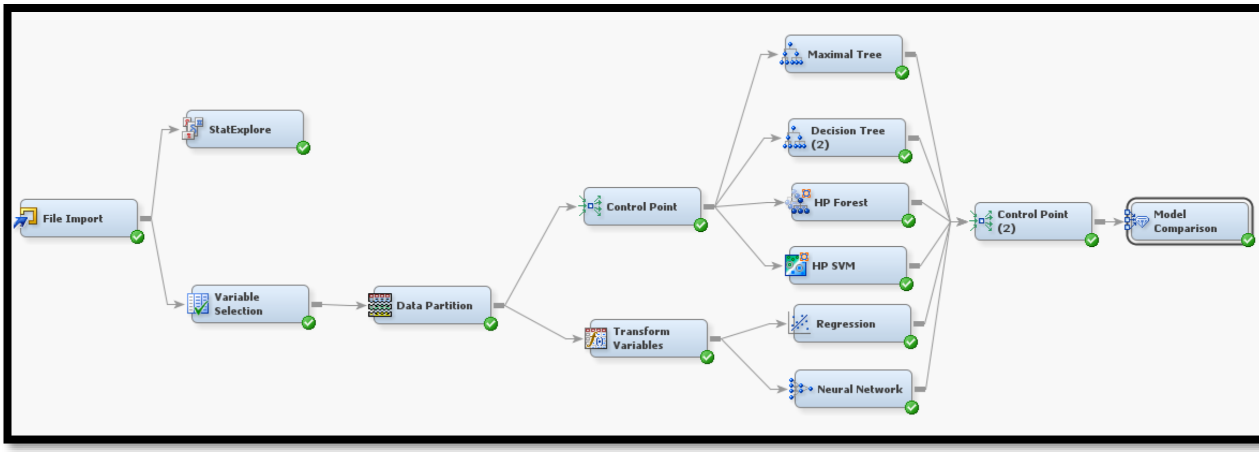


Figure 7.5: Model Workflow Diagram

8. RESULT COMPARISON

As per the fit statistics Misclassification Rate output the Decision Tree is the best fit model for this data set followed by Neural Network, Regression model and so on. The model fitness can be determined by different measures but mainly Misclassification Rate is considered to check the performance of the model. The Decision Tree with the Assessment Measure as 'Misclassification' is considered as the best model for both the bank additional full dataset and small dataset. The nr_employed, pdays and month variables were the most influential variables to predict the output whether the customer will subscribe for a Term deposit or not.

However, few differences were noticed while working with the small and big dataset. It is observed that Decision Tree model is good for type of data set whether it is small or big but SVM performs well with small dataset as compared to big dataset. Also the Neural Network model was found as a best fit for large dataset as compared to small dataset when compared using the Misclassification Rate of different models.

In the given research paper, four Data Mining models Decision Tree, Neural Network, Logistic Regression and Support Vector Machines were compared and the Neural Network is the best model amongst all. These models were compared using two metrics, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT). The AUC in the original research paper is 0.80 and ALIFT is 0.67.

In the current study, I am considering the Misclassification Rate as the selection criteria of the best model and it is found that for Decision Tree it is best. Then also the ROC index for the Decision Tree is 0.75 which depicts that the performance of the model is good.

The cumulative curve for decision tree is depicted in the below diagram. As there is not much difference in the Cumulative Lift curve of all the models but still we can say that the Cumulative Lift curve is slightly higher for the Decision Tree model.

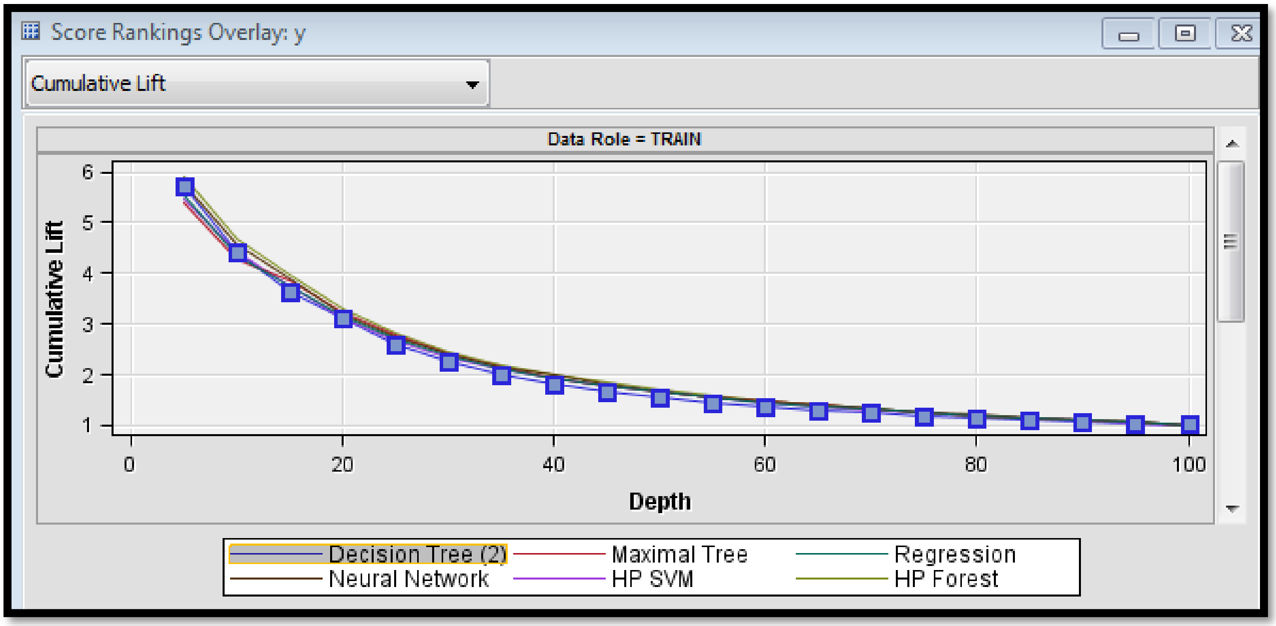


Figure 8.1: Model Comparison – Cumulative Lift Plot

9. CONCLUSION

The Portuguese bank dataset is studied and based on the experiments performed, it is concluded that Decision Tree is the best fit model followed by Neural Network based on the Misclassification Rate. There is a very slight difference between the misclassification rates of the two models. By considering the above comparisons we can say like the Decision Tree algorithm which is a nonlinear and non parametric model is the best and well known model which handle the missing values without the need for imputation. Also it is easy to understand conceptually.

10. REFERENCES

Moro,S.,Cortez, P. and Paulo,R.(2014) A Data-Driven Approach to Predict the Success of Bank Telemarketing. Retrieved from: https://repositorio.iscteiul.pt/bitstream/10071/9499/1/post_print_dss_v3.pdf

Han,J., Kamber,M. and Pei,J (2017). Data Mining Concepts and Techniques.

Fawzy, D., Moussa, S., & Badr, N. (2016). The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics. Asian Journal of Applied Sciences Retrieved from https://www.researchgate.net/publication/305768258_The_Evolution_of_Data_Mining_Techniques_to_Big_Data_Analytics_An_Extensive_Study_with_Application_to_Renewable_Energy_Data_Analytics