**Prediction of success of term Deposit of Portuguese bank data**

In cutting age, competition in banking domain marketing plays an important role. Marketing is planed execution of result, which gathered from analysis of long-term customer behaviour, it is more important to know that which type of customer need to focus for successful outcome. In[2] explained various prospects of marketing strategy. As per [2] cost of marketing and retaining new customer cost five times more than existing customer.

Telemarketing is one the effective ways of marketing in banking domain and it affectively results in good customer relationship. However [2] it has been observed that around 191 million customers registered who does not want telemarketing calls in 2009.

In this research we are going to build model to predict success rate of taking term deposit on basis of 41,188 records present in dataset (<https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>). Telemarking is the mode of marketing strategy used to contact customers for this data set. There are about 21 variables, which include set of information of about 41,188 customers, which has been considered for creating models. As the research is parametric therefore model used are Decision tree, Logistic regression, Neural network and Support vector machine[1].Along with this model going to create Interactive decision tree with split rules from the list and gradient boosting approach for single predictive model. The tool used for creating model and analysis of variable is SAS Enterprise Miner.

1. INITIAL DESCRIPTIVE ANALYTICS

**1.1 Finding the variable of interests:**

To start with variable of interest for prediction using  , this will provide the variables which contributes most in predicting the outcome variable that is y. Here y is a binary variable in which if outcome is 0 then customer will more likely **not to** opt for term deposit and if output is 1 then customer will likely to opt for term deposit.

**1.2 Rejected Variables:**

With the help ofrejecting the variables at the initial stage of model creation, below are the variables, which are not under consideration for creation of model with explained reasons.

**Analysis and Findings of Rejected variable:**

1. nr\_employed (Number of employees) is contributing most for the model but after further exploration it has been observed that this variable has 5191 value with frequency 7763 and other data is missing. Consider this variable will give biased result as has only one type of data and other values as missing. On basis of existing value, we cannot predict missing data because there is no value to predict.
2. marital ('divorced','married','single','unknown') is also contributing less and R2 for

marital variable is also less therefore can be rejected.

1. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact) this is highly skewed data with high kurtosis and R2 for this variable is less and chi-square contribution is also less.
2. day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri') for this variable last contact day of week is almost has similar frequency for all the days as shown in fig 1. So it does not matter on which day customer is contacted.

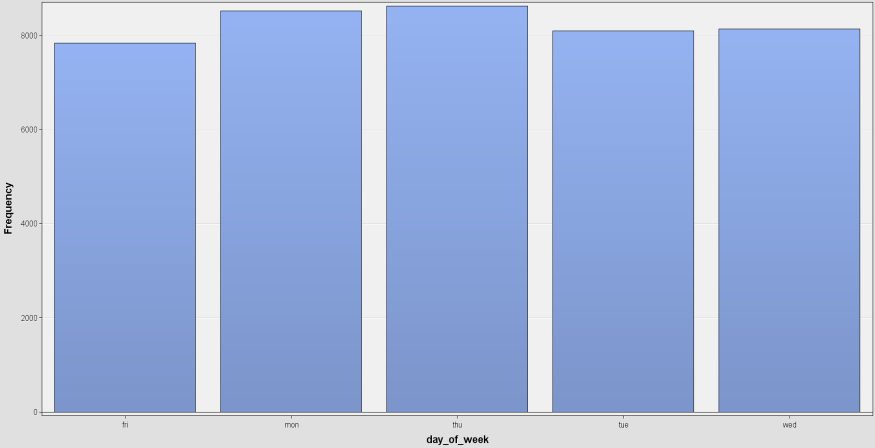


Fig1: Frequency distribution of ‘day\_of\_week’ contacted

1. housing: has housing loan? (categorical: 'no','yes','unknown') This variable also has almost similar frequency for response as ‘yes’, ’no’ and ‘unknown’ as shown in Fig 2. In other words it less likely to take term deposit for customers who are having housing loan.

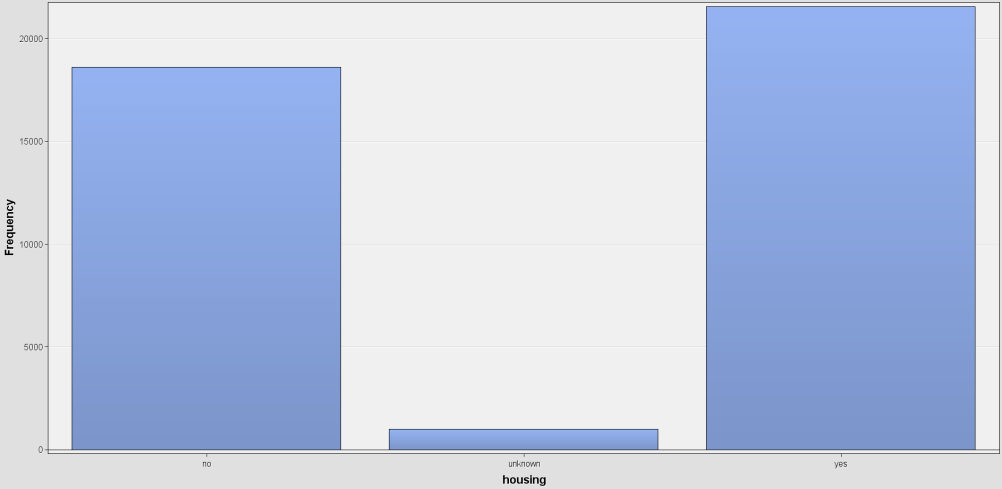
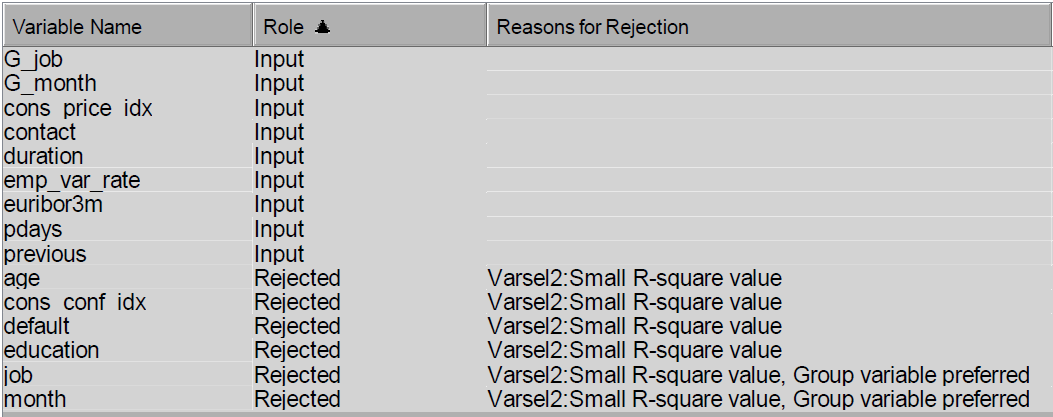


Fig2: Distribution of variable ‘Housing’

1. loan: has personal loan? (categorical: 'no','yes','unknown') Loan is biased data as 82.42% of data out of 41188 records has ‘No’ as response and 2.40 % of data is unknown which can be ignored as it is less. Moreover R2 and chi-squared contribution towards prediction is also less.
2. poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success') In this around 86.34% values are non-existent therefore this variable is also contributing very less to the prediction.

**1.3 Input variables:**

After rejection of above variables with the help of  selecting the most effective variables for the model. Variable selection will calculate the R2  for each variable and automatically reject the variable which are less likely to contribute in model. In this stage of research having 9 variables of interest which we are going to use for building model. As shown in list 1 below variables are considered for initial exploration.



List 1: List of variable on basis of R2 result

After further analysis of Input variables it has been observed that class variable contact represent the mode of contact for person is either ‘cellular’ or ‘telephone’ have 63.47% of people contacted by cellular(List2). It is a considerable percentage to taken into account. Other than this there is one more class variable job having categories like ‘admin.', 'bluecollar', 'entrepreneur' etc has 25.30% frequency for ‘admin’. The distribution of category of class variable shown in Fig3 having ‘admin.', 'bluecollar' and ‘technician’ with maximum frequency.

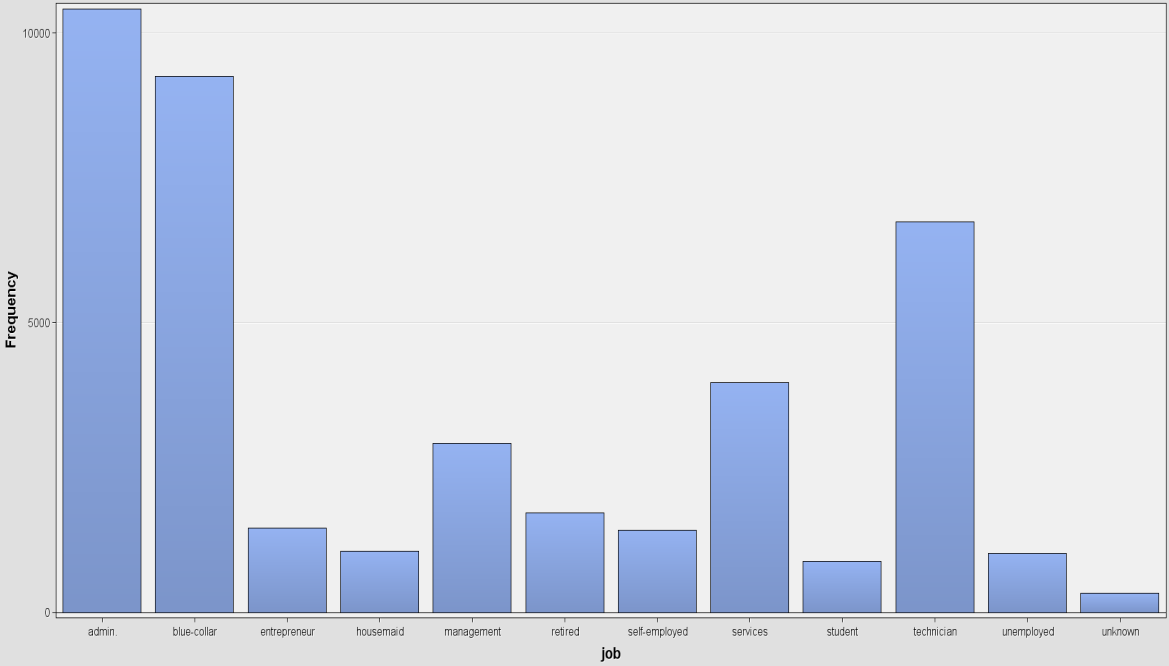
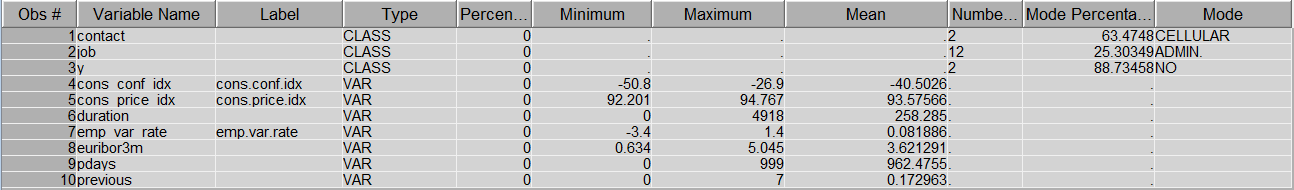


Fig3: Frequency distribution of variable ’job’

Other than class variable ‘cons.conf.idx’, ‘cons.price.idx’, ’emp.var.rate’, ‘duration’ and ‘euribor3m’ have normal distribution. Little skew and kurtosis observed for these variables but it is within exactable range of +2 and -2, therefore no need of transformation. The variable explained monthly, quarterly and daily indicator for employment variation, consumer price and confidence index.



List2: Description of Input variables

1. **Additional Data Preparation:**

Data preparation has been done with the help of  , there are two variables which has skew over the range of +2 and -2. These variables are ‘pdays’ and ‘duration’ which explains number of days passed by after the customer is contacted and last contact duration.

1. ‘pdays’: Although this variable is continuous but has only two values as response. Either number of days passed after the customer is contacted is 99 or 999. There are 1515 customers has been contacted 99 days before and 39673 customers has been contacted 999 days before. This is shown in fig4.

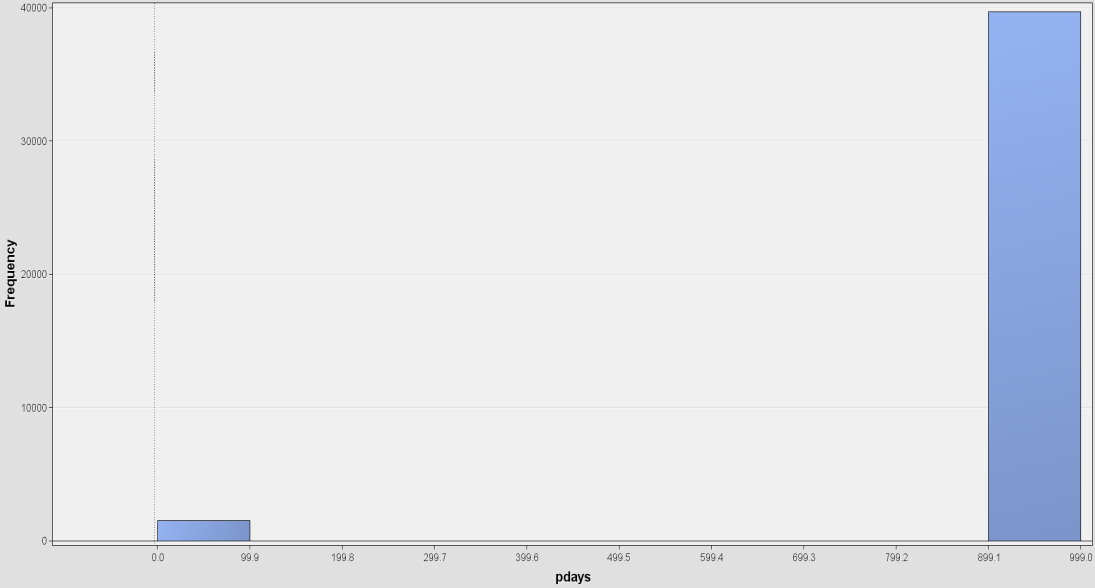


Fig4: Distribution of ‘Pdays’

To transform this variable using bucket as transformation function which will divide the data in two category , one with value 99 and another with value 999. This will result in formation of new variable ‘BIN Pdays’.

1. ‘duration’ : This is also continuous variable which has minimum value as 0 and maximum as 4918. As explained in data set description this variable is highly affects target. This is not predefined variable, after end of the call the duration of customer contact can be defined in seconds. Frequency distribution of this variable is shown in Fig5.

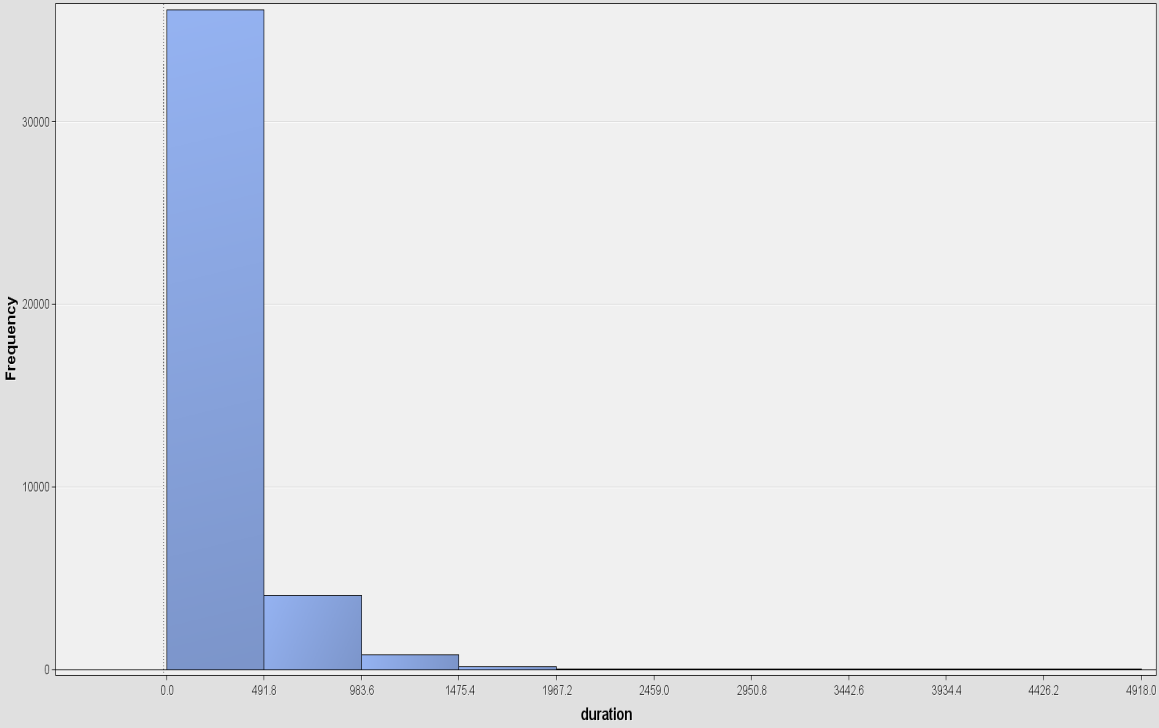


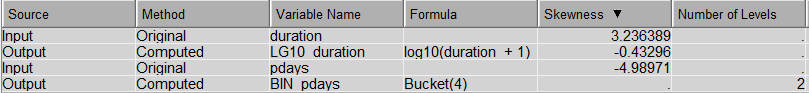
Fig5**:** Distribution of variable ‘duration’

To overcome this skewness in data, transforming it with log10 transformation. This will result in creation of new variable ‘LG10 duration’.

**2.1 Summary of transformation:**

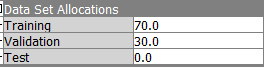
After transformation of variables, the skewness for variable ‘duration’ is reduced and came down to “-0.43296” which is within the acceptable range of +2 and -2. Before transformation, the skewness was “3.236389”.

Moreover, ‘pdays’ also changed to two levels after the transformation which is 99 and 999. After this transformation, data is ready for model creation.



1. **Data Mining Algorithms**

There are six data mining algorithms used for creating the model that are NN(Neural Network), Decision tree(DT), Logistic Regression(LR), Support Vector Machin(SVM), Interactive Decision Tree(IDT) and Gradient Boosting(GB). Data partitioning is done in the ratio 70 to 30 where 70 % is training data and 30 % is validation data. All the models are created with help of this data partitioning.



**3.1 Logistic Regression:**

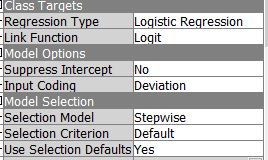
The regression model is built by  node in the SAS Enterprise Miner. The Logistic regression function is used when target variable is binary [4]and for this research our target variable Y is binary. Logistic regression work on the concept of probability, its prediction of the occurrence of binary value of target based on several independent variables.

Logistic regression works on K model equation (1.1).

 …1.1

Here Pr{y = k+1} is the category of dependent variable; x1, x2, ……. , xp are the independent variables; a0i, a1i, ... , api are model parameters.

*I* = 1, 2, ……., k represent the possible category of prediction.



The configuration used for creating model is logit function, which provides result of cumulative logistic distribution of the function. Selection model used is Stepwise.

**3.2 Neural Network:**

For creating Neural Network  node used in SAS Enterprise miner. Neural network is complex form of regression the concept. Neural network works on the nonlinear relationship between input and target variable which is shown in Fig:

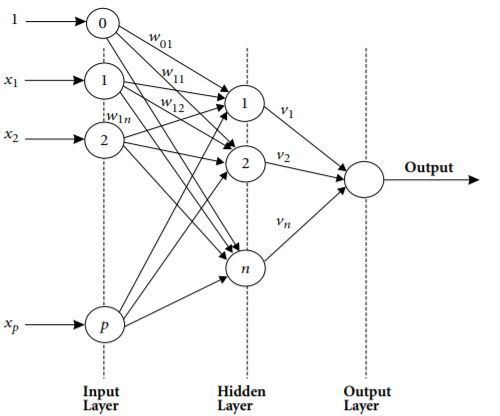


Fig6: Neural network internal architecture

The setting used in SAS for Neural network is **direct connection** as **Yes** which connects input and output units directly. We will also use number of **hidden units** as **5** which will consider 5 units of hidden layer.

**3.3 Decision Tree:**

Decision tree works on splitting the independent variables into several sub groups and this process continuous till not further splits can be done. Decision tree explains the mostly the behavior of target variable with respect to input variables. In SAS Enterprise Miner select the **maximum depth** as **10** to consider 10 split for the generation and **leaf size** to **8** for 8 training observation. As we don’t have missing values in data set so no need to set Number of surrogate rules which will determine the split in non-leaf node. **Assessment Measure** changes to **Decision**.

**3.4 Gradient Boosting:**

Gradient Boosting approach do partitioning for single target variable and it resample the analysis of data several times. This model used to prevent from overfitting of model. We will use  for creating gradient boosting model in SAS Enterprise Miner. Setting the **maximum depth** to **10** will consider for 10 split of generation and **Number of Surrogate rule** to **2** for missing values.

**3.5 Support Vector Model:**

SVM is machine learning algorithm which is similar to neural network. This algorithm is suitable linear data set. We have used default setting in SAS Miner for optimized result.

The node used for SVM is .

1. **Evaluation of Data Mining Models**

As we have created 6 Models , To decide which model is best fit for the prediction we will use ROC chart. First lets look at the Neural Network which is at the top with **ROC** **index value** as **0.94** but ther then ROC value we have to take squared error into the consideration. Fig7 Shows that the area under the curve for NN is maximum as compared to other models and misclassification rate for the model. For Neural network **missclassification rate** is **0.086348**.

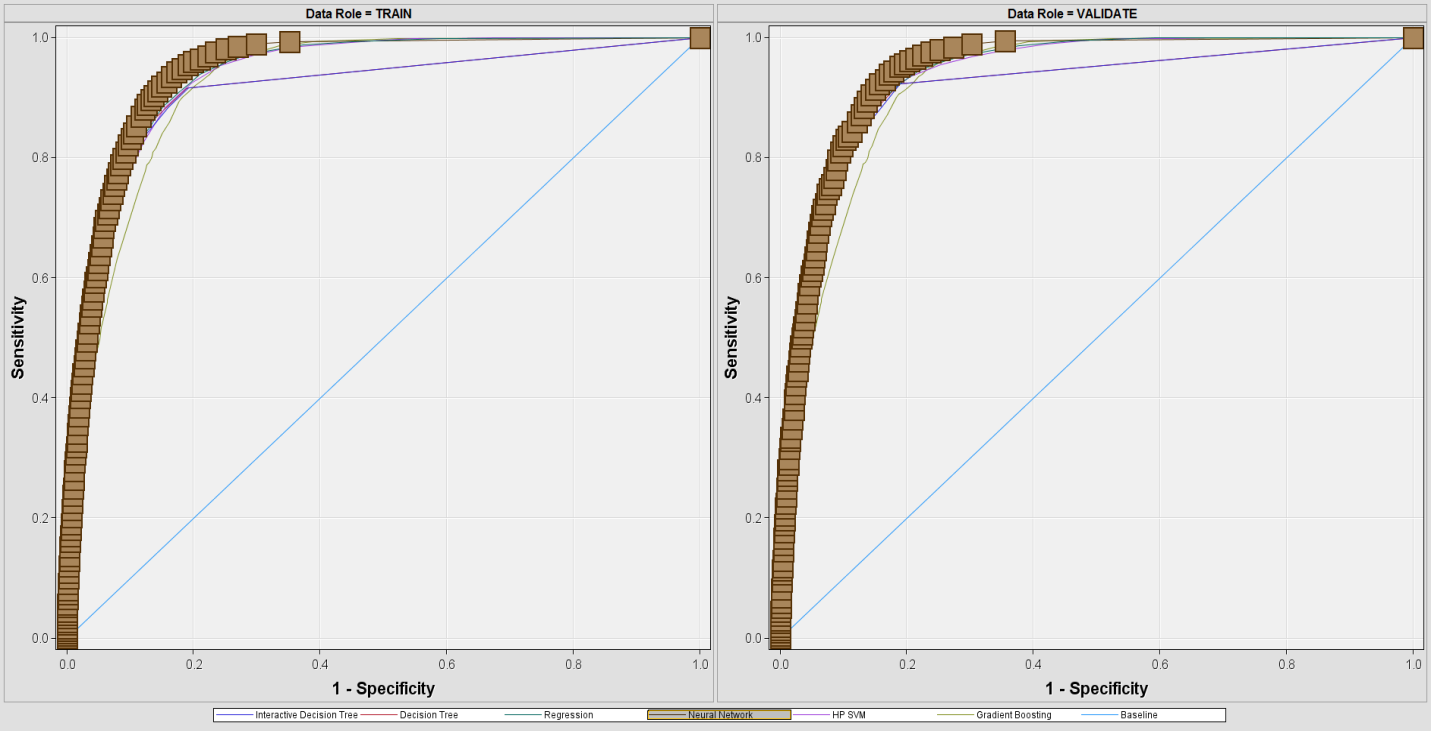


Fig7: Nueral Network ROC chart

Second Model which is in the top of the list is Decision Tree index value as 0.91 but other than ROC value we have to take squared error into the consideration. Fig 8 Shows that the area under the curve for DT is less then NN. For DT missclassification rate is 0.086752 which is more than NN.

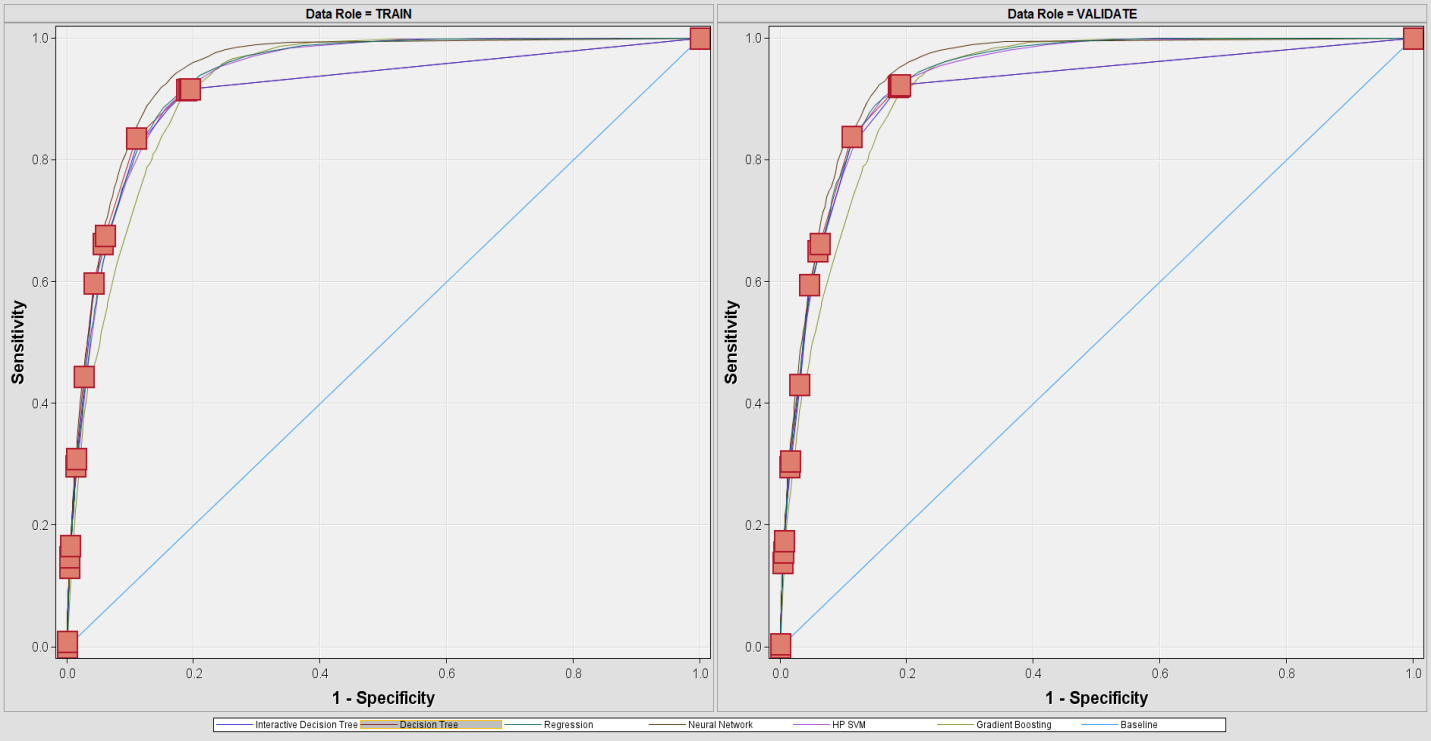


Fig8: Decision tree ROC chart

Third Model which is in the top of the list is Interactive Decision Tree with **index value** as **0.91** but other than ROC value we have to take squared error into the consideration. Fig 9 Shows that the area under the curve for IDT is less then NN. For DT **missclassification rate** is **0.087724** which is more than NN and DT. Now here ROC value for DT and IDT is same so we need to consider sum of squared error which is 1509.72 for IDT and comparitively more them DT and NN.

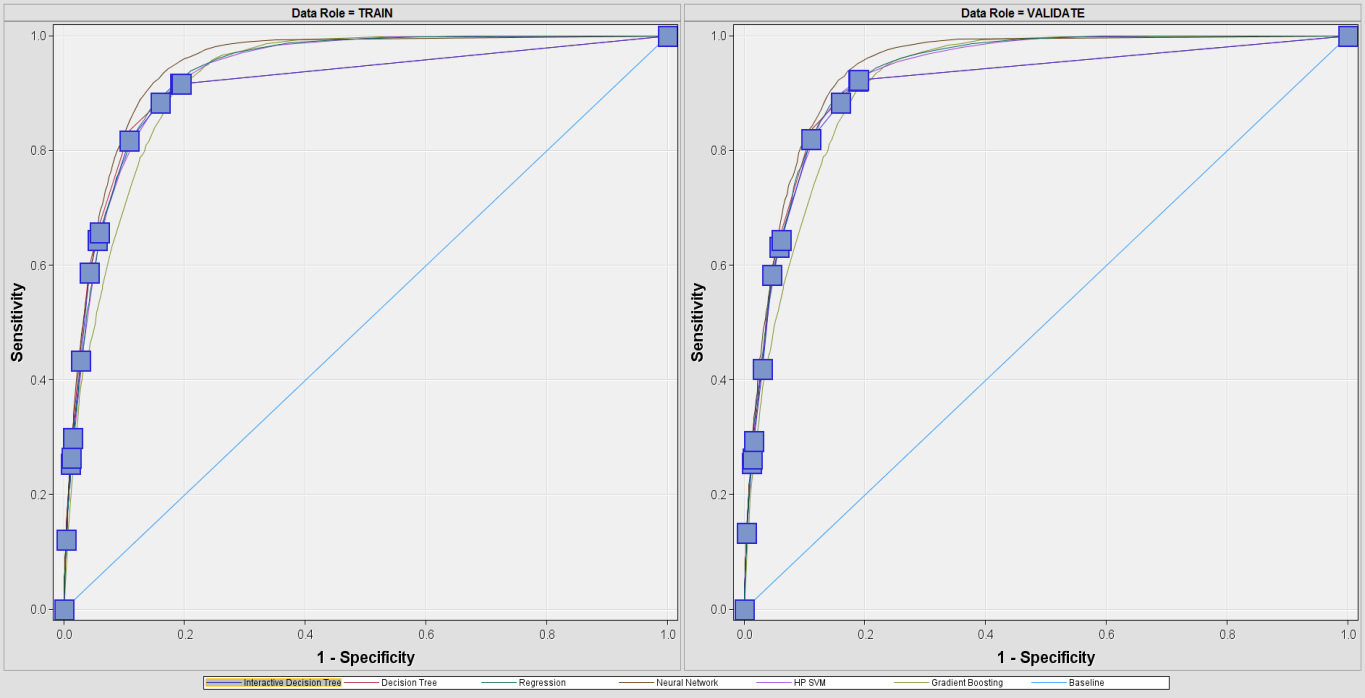
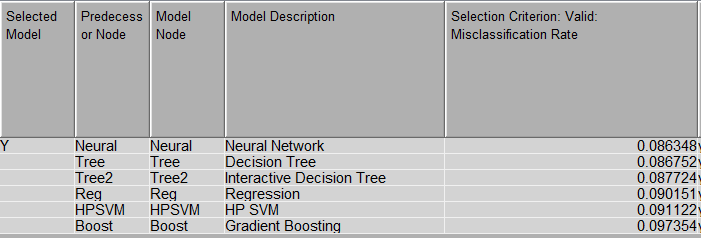


Fig9: Interactive Decision Tree ROC curve

Regression have **ROC value(0.93)** more than DT and IDT. But sum of squared error is more for regression model. The same kind of scenarios observed with SVM and Boost model where sum of sqaured error is more. For SVM model average suqared error is **3644.24** where as for Boost model average squared error is **1660.07**.

The below list shows that which model is best fit for the prediction:



Below is the lis of ROC index and sum of squared error for the model:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Statistics** | **Neural** | **Tree** | **Tree2** | **Reg** | **HPSVM** | **Boost** |
| Valid: Roc Index | 0.94 | 0.91 | 0.91 | 0.93 | 0.93 | 0.92 |
| Valid: Sum of Squared Errors | 1445.76 | 1489.56 | 1509.72 | 1532.74 | 3644.24 | 1660.07 |

1. **Comparision**

In the original research paper they have created 4 model and for this data mining using 6 models, we can compare the both the research on the basis of common models used.

For existing research they used reduction of variable and considered test data set for analysis but for this analysis the Data partition used is 70% and 30% of validation and traning data set and test dataset 0 for the research. This is differece in data set consideration.

Now if look in the tool used for research the original research paper that R studio and rminder package but for this assignment we used SAS Enterprise Miner.

**Model comparision:**

In original researach paper top three model fit for the prediction are NN>SVM>DT but in this research the more precision we are getting for NN>DT>IDT>REG>SVM>BOOST.

In research paper the ROC curve has significant difference between regression model and other but for the analysis done with SAS all look similar to each other.

For Nural network original research paper has ROC index of 0.91 but here we got ROC index for NN as 0.94. Similarly for DT research paper has ROC index as 0.83 but here we got ROC index as 0.91.

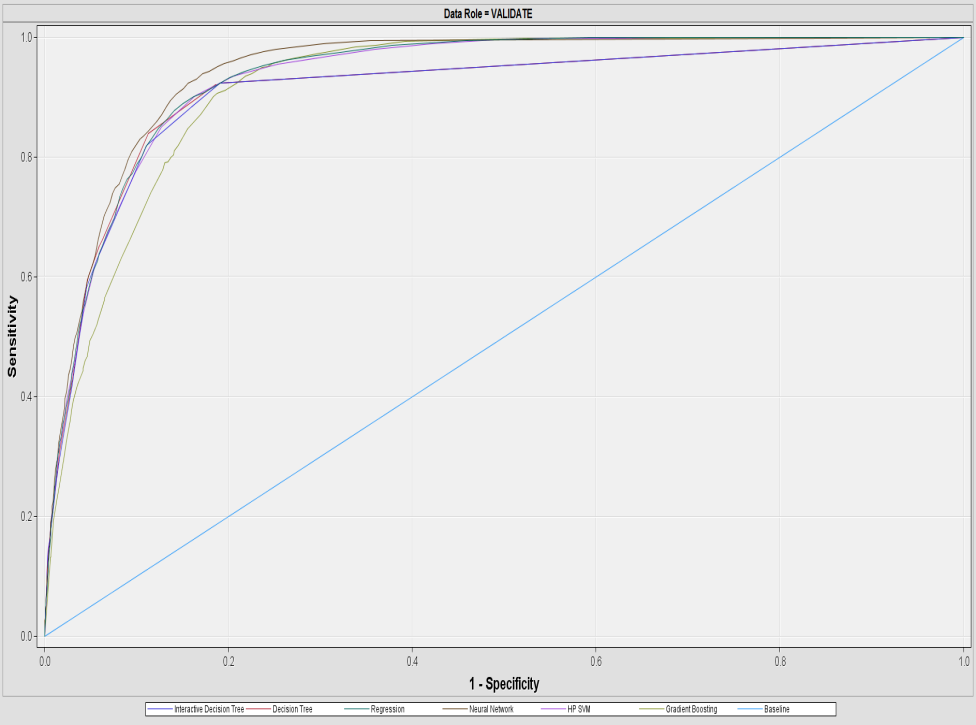


Fig: ROC of this analysis

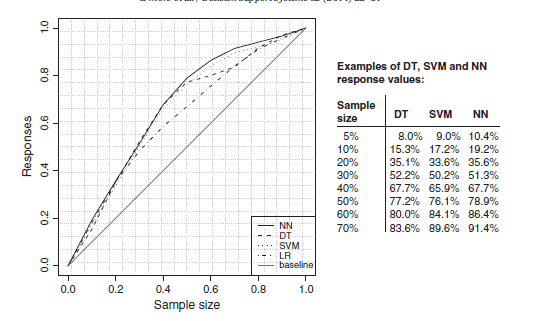


Fig: ROC of refereance paper

**Reference:**

[1]S.S. Haykin, Neural networks and Learning Machines, Prentice Hall, 2009.

[2] Philip Kotler, Kevin Lane Keller, Framework for Marketing Management, 5th edition

Pearson, 2012.

[3] David Arnott, Graham Pervan, Eight key issues for the decision support systems discipline,

Decision Support Systems 44 (3) (2008) 657–672.

[4] Refaat, M. (2007). Data preparation for data mining using SAS. Amsterdam: Morgan Kaufmann Publishers.

Base research paper

[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].