DUBLIN INSTITUTE OF TECHNOLOGY

Title: Bank Marketing

Data Mining

(Assignment 1)

Lecturer: Brendan Tierney

NAME: **ASHIS SAHU**

STUDENT ID: **D17129721**

COURSE: ***DT-228A (DATA ANALYTICS) 2018-19***

ABSTRACT

This paper is proposed to build data mining models using SAS enterprise miner which will predict the success of bank telemarketing campaign where the clients subscribing to long-term deposit plan or not, is predicted. The dataset used is a Portuguese bank dataset which was processed in SAS enterprise miner diagram for the model is presented below in figure 1. The data mining (DM) model used for predicting the success of bank telemarketing campaign are decision tree, neural network, autoneural network, support vector machine, random forest and linear regression. The comparison and discussion of model has been presented. Comparison of the model used in this paper and the paper by Sergio Moro, Paulo Cortez , and Paulo Rita has been proposed. Several key attributes has been discussed briefly which affet the outcome variables and how SAS enterprise miner does feature selection, engineering and uses DM model for optimal results are discussed.

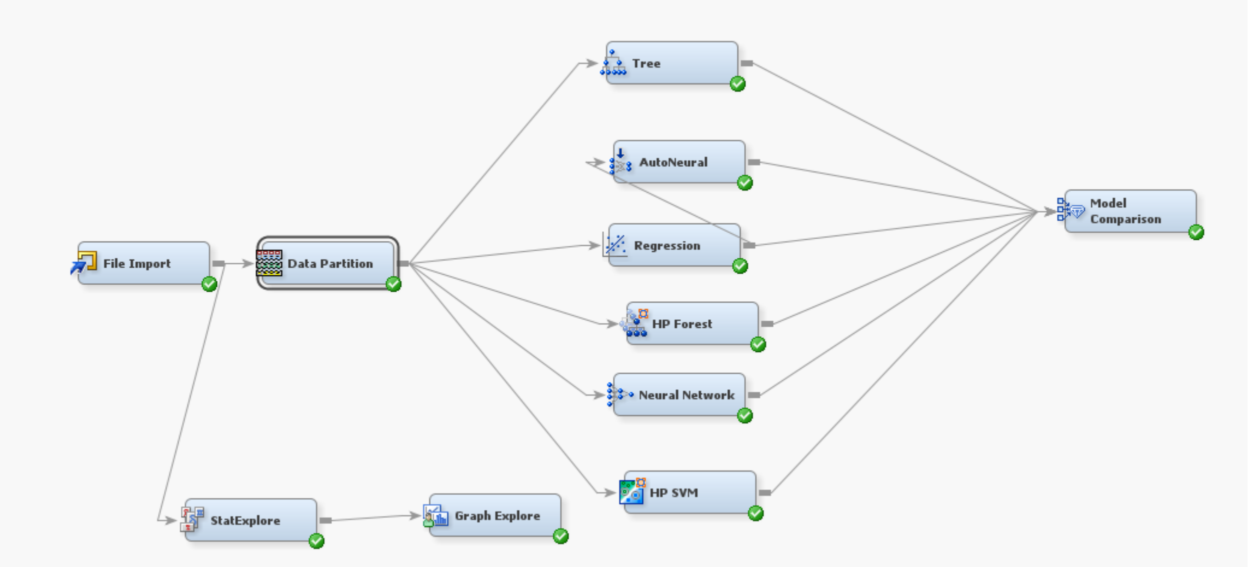


FIGURE 1 DIAGRAM

1. DATASET DESCRIPTION

The data is collected from the UCI ML Repository website (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The data set is used in a marketing campaign for a Portuguese bank institution for selling of long term deposit. The bank conducts marketing campaigns. The purpose of this dataset was to identify customers who will be most likely to subscribe to a term deposit scheme on the basis of previous marketing campaigns. There were two versions of the data set given in the website; the second data set containing 20 input attributes was used for this paper. It contains a 10% sample (4,119 records) of the entire data set which was used by Sergio Moro, Paulo Cortez , and Paulo Rita in their paper.

1.1 DESCRIPTIVE ANALYTICS AND FINDINGS

Varaibles

1.1.1 CONTACT

Contact variable is a nominal variable and input variable to outcome variable y in dataset. It has two values cellular (64.01%) and telephone (35.99%) in the dataset

1.1.2 JOB

Job variable is a categorical variable in data set, with variables such as "admin.", "blue-collar", "entrepreneur", “housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed".

1.1.3 MARITAL

Marital is categorical variable in the data set, with variables such as "divorced","married","single", “unknown”.

1.1.4 AGE

Age is numerical variable in the data set. Age has the following statistics(Mean =40, sd= 10.37, Min = 18, Maximum =88.) Skewness =0.72 and kurtosis = 0.56. The skewness and kurtosis are under acceptable range of +/-2 ,so Age is considered as a variable attaining normal distribution as per (George, Mallery,2010).

1.1.5 EDUCATION

Education is a categorical variable in the dataset with variables such as "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", “unknown”.

1.1.6 DEFAULT

Default is categorical variable in the data set with variables such "no","yes", “unknown”. It determines the client is defaulter or not.

1.1.7 HOUSING

Housing determines the client having a housing loan or not. It is a categorical variable in dataset, with values “no", "yes".

1.1.8 LOAN

Loan determines the client has personal loan or not. It is a categorical variable in the dataset with values "no", "yes".

1.1.9 MONTH

Month is a categorical variable which determine the last contact month of the year, it has month of the year.

1.1.10 DAY_OF_WEEK

day_of_week is a categorical variable in the dataset with values "mon","tue","wed","thu","fri".

1.1.11 DURATION

Duration is a numeric variable in the dataset which determine last contact duration, in seconds. (Mean =257.09, sd= 262.152, Min = 5, Maximum =3643.) Skewness =0.72 and kurtosis = 0.56. The skewness and kurtosis are under acceptable range of +/-2, so duration is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.12 CAMPAIGN

Campaign is the number of contacts performed during this campaign and for this client, it is a numeric variable. Campaign has the following statistics (Mean =40, sd= 10.37, Min = 18, Maximum =88.). Skewness =4.1788 and kurtosis = 25.87. The skewness and kurtosis are not under acceptable range of +/-2 ,so we cannot consider variable campaign as attaining normal distribution as per (George, Mallery,2010).

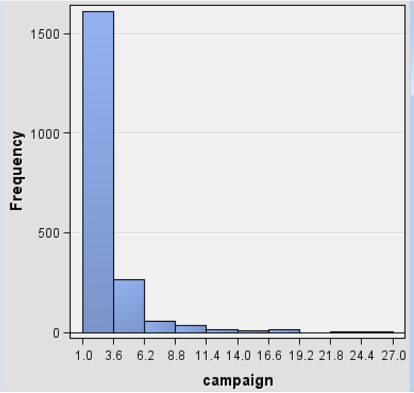


FIGURE 2 DISTRIBUTION OF CAMPAIGN VARAIBLE

1.1.13 PDAYS

Pdays is the number of days that passed by after the client was last contacted from a previous campaign. Its is a interval variable (Mean =956.56, sd= 200.87, Min = 0, Maximum =999). Skewness =-4.52 and kurtosis = 18.49. The skewness and kurtosis are not under acceptable range of +/-2 ,so we cannot consider variable pdays as attaining normal distribution as per (George, Mallery,2010).

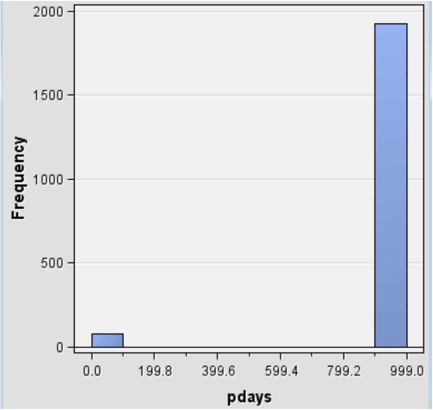


FIGURE 3 DISTRIBUTION OF PDAYS

1.1.14 PREVIOUS

Previous is the number of contacts performed before this campaign and for this client. It is a numeric variable in data set with (Mean =0.19, sd= 0.53, Min = 0, Maximum =5). Skewness =3.69 and kurtosis = 17.26. The skewness and kurtosis are not under acceptable range of +/-2 ,so we cannot consider variable previous as attaining normal distribution as per (George, Mallery,2010).

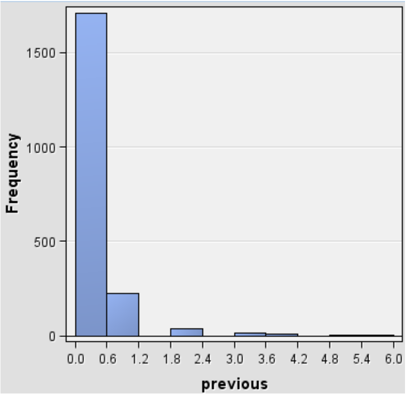


FIGURE 4 DISTRIBUTION OF PREVIOUS

1.1.15 POUTCOME

Poutcome is the outcome of the previous marketing campaign. It is the categorical a variable with values "failure", "nonexistent", "success".

1.1.16 EMP.VAR.RATE

emp.var.rate is an employment variation rate of quarterly indicator. It is a numerical variable. Skewness =-0.70 and kurtosis = -1.07. The skewness and kurtosis are under acceptable range of +/-2 ,so emp.var.rate is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.17 CONS.PRICE.IDX

cons.price.idx is a consumer price index monthly indicator. It is a numeric variable. Skewness =0.218 and kurtosis =- 0.84. The skewness and kurtosis are under acceptable range of +/-2 ,so cons price index is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.18 CONS.CONF.IDX

cons.conf.idx is a consumer confidence index - monthly indicator. It is numeric variable. Skewness =0.28 and kurtosis = -0.31. The skewness and kurtosis are under acceptable range of +/-2 ,so cons conf index is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.19 EURIBOR3M

euribor3m is the3 month rate of daily indicator. It is a numeric variable. Skewness =-0.69 and kurtosis = -1.4. The skewness and kurtosis are under acceptable range of +/-2 ,so duration is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.20 NR.EMPLOYED

nr.employed is the number of employees in quarterly indicator. It is numeric variable. Skewness =-1.05 and kurtosis = 0.0003. The skewness and kurtosis are under acceptable range of +/-2 ,so duration is considered as attaining normal distribution as per (George, Mallery,2010).

1.1.21 Y (OUTCOME VARAIBLE)

Y is the outcome variable in the bank marketing dataset. It represents the client subscribed a term deposit or not. It is a binary variable in the dataset presents "yes","no".

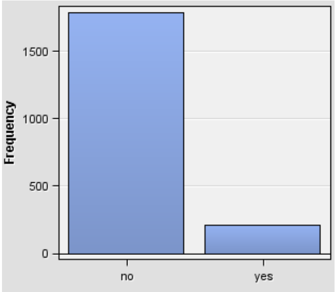


FIGURE 5 OUTCOME 'Y' HISTOGRAM

1.2 CHI-SQUARE PLOT

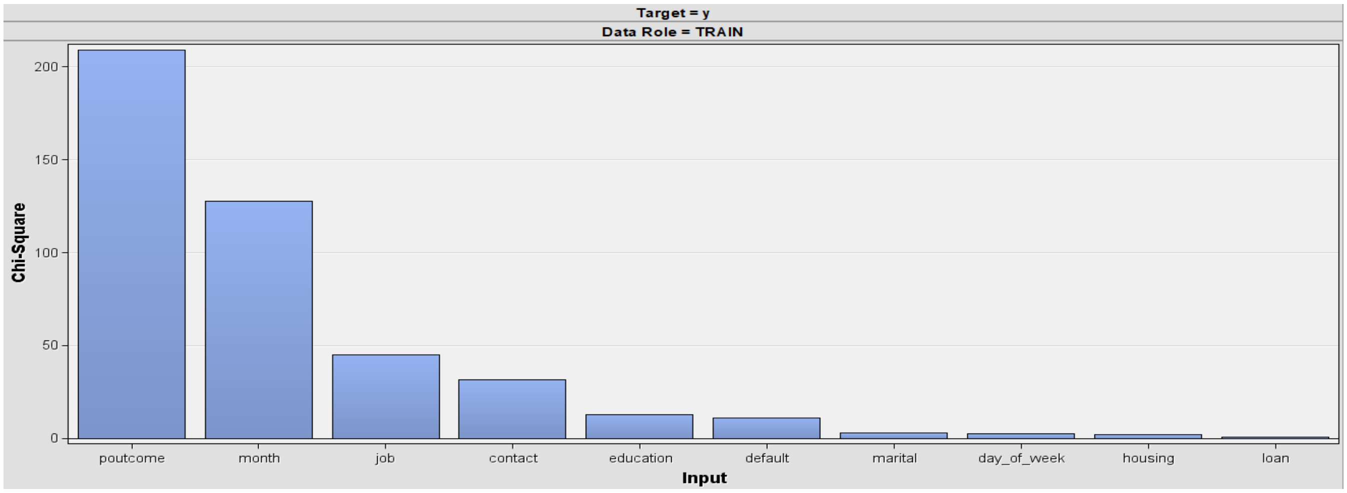


FIGURE 6 CHI-SQUARE PLOT OF CATEGORICAL VARAIBLE

The above figure 4, Chi-square plot of categorical variables in the data set with respect to outcome. The chi-square plot gives the association between two variables. Through the chi-square we observed that poutcome, month, job and contact is significantly associated with outcome variable ‘y’.

1.3 INTERVAL VARAIBLE SCALED MEAN DEVIATION WITH RESPECT TO OUTCOMES

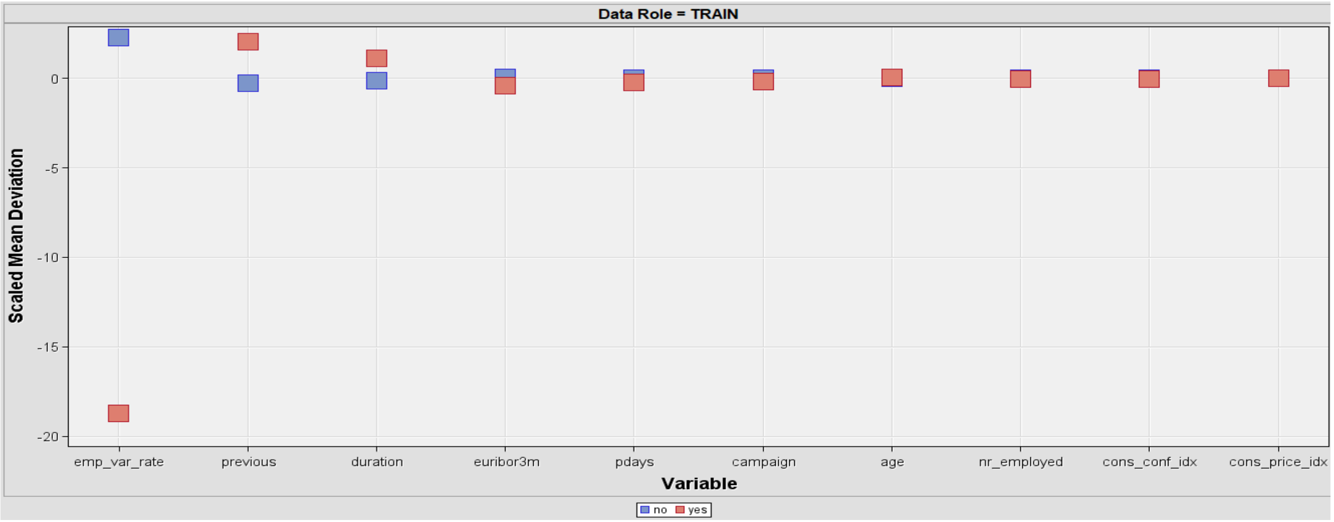


FIGURE 7 INTERVAL VARAIBLES SCALED MEAN DEVIATION

The above figure shows, the mean deviation in interval variables with respect to the outcome variable ‘y’. It has been observed that emp_var_rate, previous, duration, euribor3m, pdays, and campaign has significant deviation to the outcome variable which can be predictor variables for the outcome variable.

2. DATA PROCESSING

2.1 DATA CLEANING

The data set consist of several missing values. These unknown variables are considered can have a significant effect on the conclusions that can be drawn from the data and the data mining model need all data to be present in order to work. Therefore, the data set having several missing values in some categorical attributes, all coded with the "unknown" label. ‘unknown’ is defined as a categorical feature which will have a definite number of possibilities. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features having missing values have been replaced with a new category ‘unknown’. This strategy will add more information into the dataset which will result in the change of variance.

2.2 FEATURE SELECTION

In SAS enterprise miner, the feature selection process is automatically done by the enterprise miner, though the enterprise miner finds the variable worth of each variable with respect to outcome variable ‘y’. The variables nr_employed, cons_conf_idx, duration, euribor3m, emp_var_rate, cons_pice_idx, pdays, previous, month, age, job, contact, education, default, campaign are having significant variable worth with the outcome variable 'y', which would be used by the enterprise miner in the process of feature selection. Therefore, no manual interventions are required for feature selection in enterprise miner.

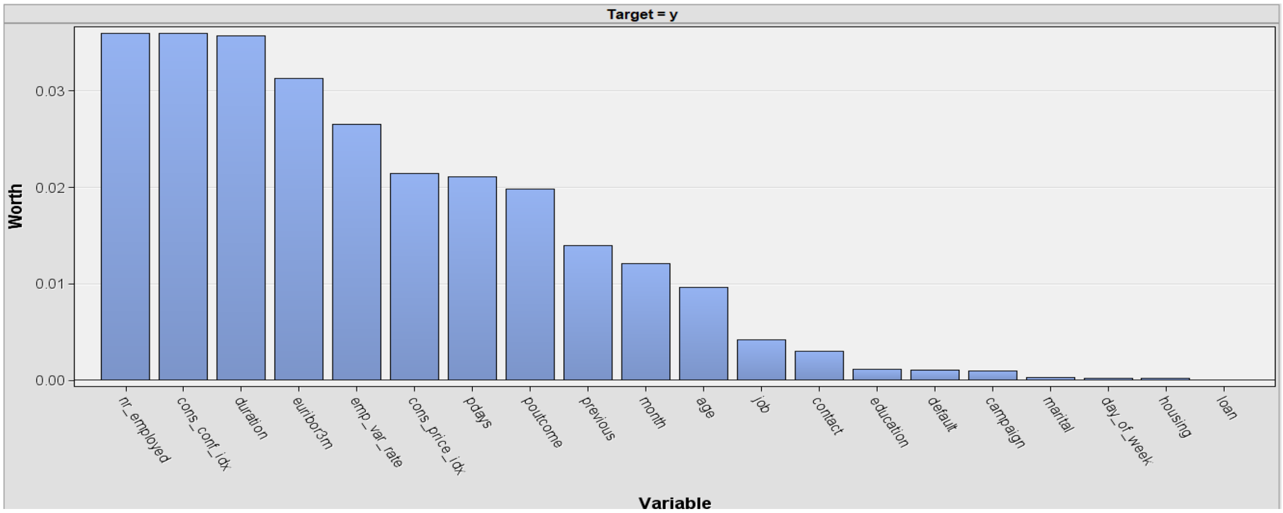


FIGURE 8 VARAIBLE WORTH WITH RESPECT TO TARGET

2.3 FEATURE ENGINEERING

There are various interval variables which are skewed or kurtosis such as pdays, duration, previous and campaign; which cannot attain normal distribution and can cause variance in output of model. Therefore these variables are normalized by SAS enterprise miner in order to be used in data mining model without causing variance in the output of the models or affecting the accuracy of the models.

2.4 DATA PARTITION

Data partitioning has been done on the bank marketing data set with 70 -30 partition, where 70% is training data and 30% is validation data. The imported dataset has 4119 records, after 70-30 partition with a seed variable of 12345; Train partition has 2883 records and validation partition has 1236 records.

3. DATA MINING ALGORITHM

There are 6 data mining algorithm used in this paper, which has been described below.

3.1 DECISION TREE

Decision tree is the most popular and powerful algorithm used for classification. It works on the information gain principle. The tree algorithm *learns* by splitting the source set into subsets based on information gain principle. This process is repeated on each derived node in recursion called *recursive partitioning*. The decision tree is also used for exploratory knowledge discovery, as classifier does not require any domain knowledge. The algorithm builds classification or regression models in the tree structure; therefore it is named as decision tree.

Entropy is found by:

$$\mathrm{H}(T) = \mathrm{I}_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$$

where p1, p2,….,pj are fractions that add up to 1 and represent the percentage of each class present in the child node that results from a split in the tree.

$$\overbrace{IG(T,a)}^{\text{Information Gair}} = \overbrace{\mathrm{H}(T)}^{\text{Entropy (parer t)}} - \overbrace{\mathrm{H}(T|a)}^{\text{Weighted Sum of Entropy (Children)}}$$

$$= -\sum_{i=1}^{J} p_i \log_2 p_i - \sum_{a} p(a) \sum_{i=1}^{J} -\Pr(i|a) \log_2 \Pr(i|a)$$

In decision tree, Information gain is used for each feature split on at each step of building the tree. Default configuration has been used in decision tree, which gives the optimal output by the SAS enterprise miner.

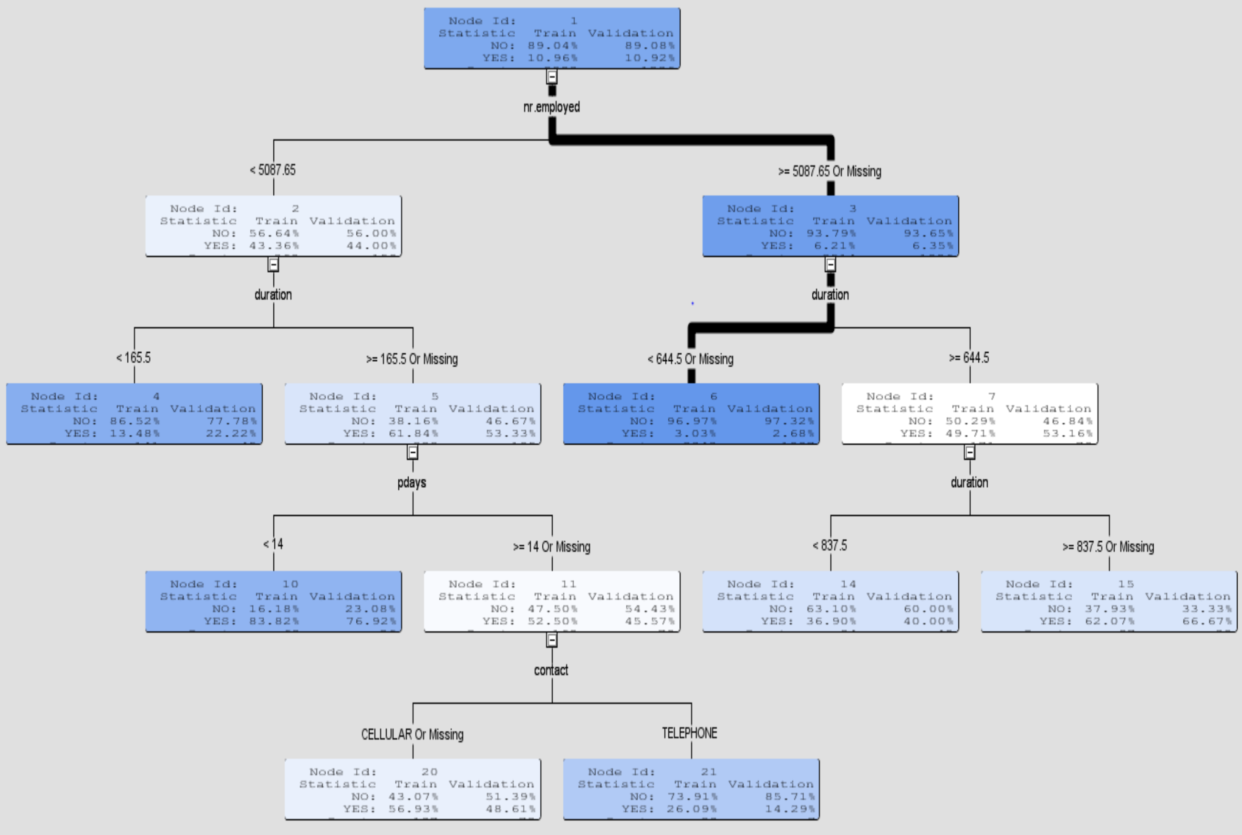


FIGURE 9 DECISION TREE

The variable of importance for decision tree are duration, nr_employed, pdays, and contact. The variable of importance are calculated due to the split of each node in the decision tree and determined by occurrence of the split. The table of variable importance for split in decision tree is given below:

Variable name	Variable Importance
Duration	1.00
nr_employed	0.87
Pdays	0.28
Contact	0.18

3.2 NEURAL NETWORK

A neural network is composition of back propagation network and multi-layer preceptor. Neural networks also called black box model, composed of layers of computational units called neurons. These neurons are connected in different layers. The output is classified when data transformed by network of neurons. Each neuron holds some weight, sums results with other neurons in different layers and then normalizes the output with an activation function. Neural network is an iterative learning process where the network trains by adjusting the weights each time there is occurrence to predict the correct class label output. Advantages of neural networks are their high tolerance to noisy data and classification where they haven’t been trained already. The below figure 9 shows the structure of neural network, where the hidden is the layers of neurons and distributed by weights which are formed due to occurrence of variables. That means the higher the occurrence of variables in the model, higher the weight of the neuron of variable. Default configuration has been used in neural network node, which gives the optimal output by the SAS enterprise miner.

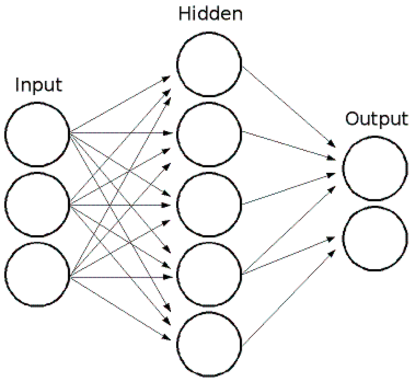


FIGURE 10 NEURAL NETWORK

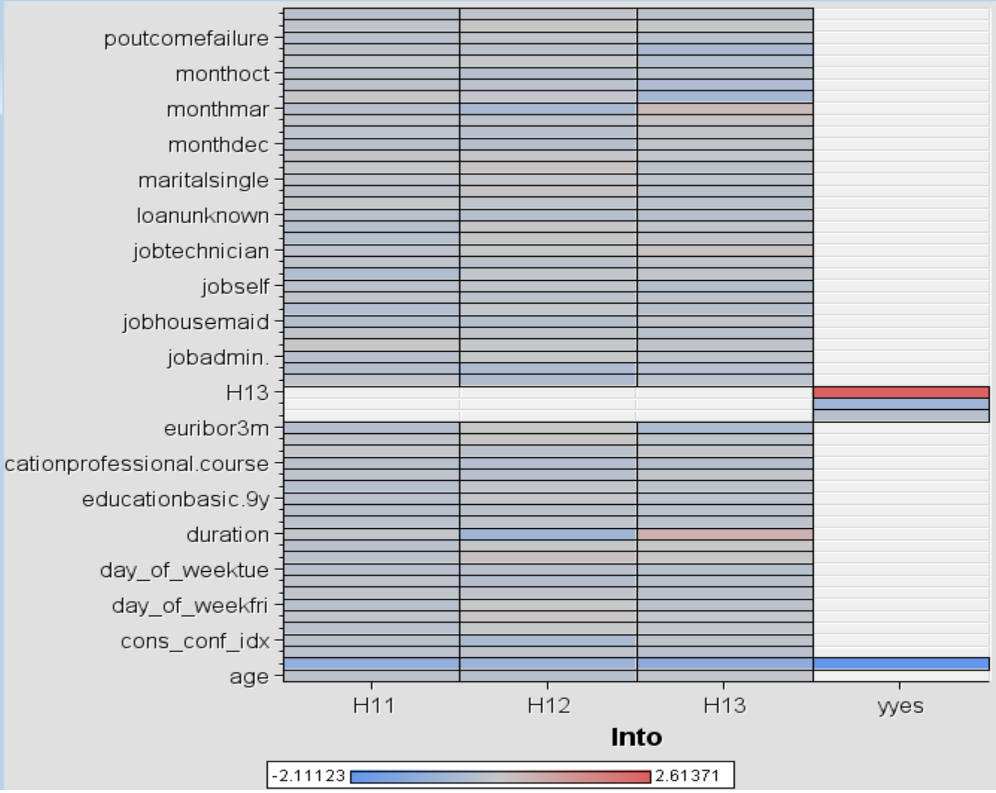


FIGURE 11 WEIGHT DISTRIBUTION OF NN

The above figure 10, shows the weight distribution of Neural network for the target variable 'Y' . It illustrates that the variable poutcome, month, duration, day_of_week, age, cons_conf_idx, euribor3m, education and job has variable weights in the neural network model.

3.3 AUTO NEURAL NETWORK

AutoNeural is a data mining model in SAS enterprise miner which is an optimized version of neural network which optimizes itself to give more accurate result. AutoNeural node belongs to the Model category in the SAS data mining process of Sample, Explore, Modify, Model, and Assessment. AutoNeural node is used an automated tool to help to find optimal configurations for a neural network model. Hidden nodes are added one at a time. The default combination functions and error functions of neural network are used in autoneural network. Default configuration has been used in auto neural node, which gives the optimal output by the SAS enterprise miner.

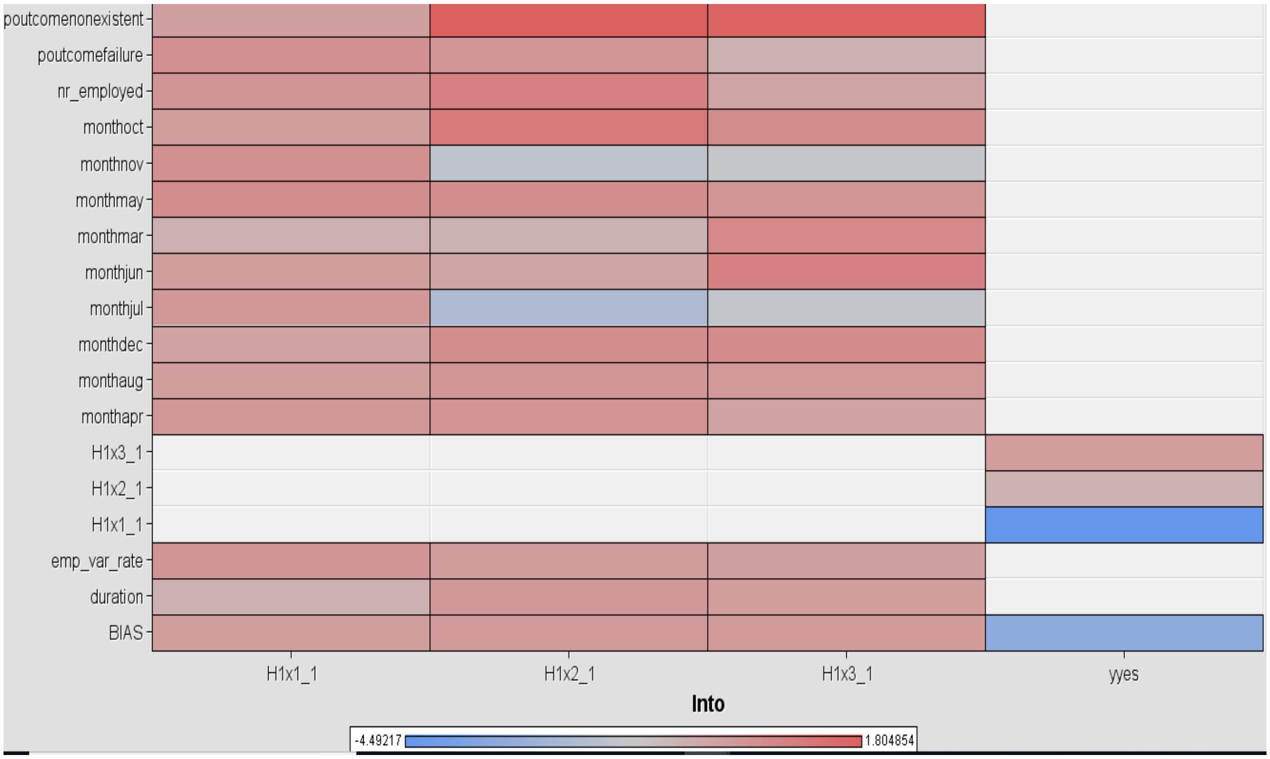


FIGURE 12 WEIGHT DISTRIBUTION OF AUTONEURAKL NODE

The above figure 11, shows the weight distribution of Neural network for the target variable ‘Y’ . It illustrates that the variable poutcome, month, duration, day_of_week, nr_employed and job has variable weights in the auto neural network model.

3.4 SVM

A support vector machine is a supervised learning algorithm that works on classification and regression analysis . SVM can work as binary and non-binary linear classifier. A Support Vector Machine (SVM) is a classifier which classifies by a separating hyperplane. Default configuration has been used in SVM node, which gives the optimal output by the SAS enterprise miner. It uses kernel trick technique to transform data and then based on this transformations it finds an optimal boundary by using support vectors between the possible outputs to predict the outcome variable.

3.5 RANDOM FOREST

Random forests or random decision forests are data mining model for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class. Random tree builds multiple decision trees and merges them together to get more accurate and stable prediction.

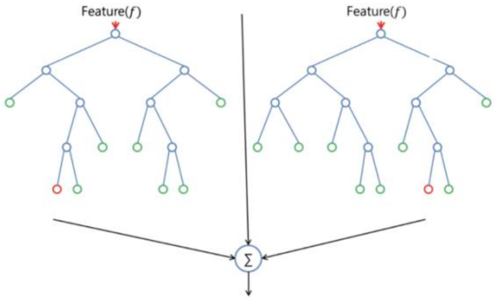


FIGURE 13 RANDOM FOREST

Several times, it shows less accuracy due to over fitting model. The random forest data mining model has been used in SAS enterprise miner with default configuration for the optimal output.

3.6 LINEAR REGRESSION

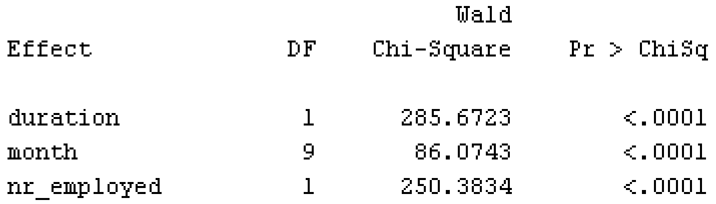
Linear regression is the most simple and popular data mining model. Linear regression model fits a linear equation to observed data. The linear regression has been devised from the line equation;

$$Y=mx+c$$

When a linear regression model is fitted, the equation formed is given by;

$$Y= b0+b1x1+b2x2+\ldots+bnxn + E0$$

The varaible of worth found by Chi-Square probability for linear regression model is given by



```
                                   Wald
Effect               DF     Chi-Square     Pr > ChiSq

duration              1       285.6723         <.0001
month                 9        86.0743         <.0001
nr_employed           1       250.3834         <.0001
```

The varaibles duration, month, nr_employed are the varaibles of worth for linear regression model which has been undertaken SAS enterprise miner. The default configuration for linear regression has been used in SAS enterprise miner for optimal output.

4. MODEL STATISTICS

4.1 CONFUSION MATRIX

A confusion matrix is a table of statistics which is used to describe the performance of a classification model (binary classifier) on a set of data where outcome is known. Various statistical measure could be derived from the confusion matrix such as specificity, error-rate, sensitivity, accuracy, and precision.

TABLE 1 CONFUSION MATRIX OF DATA MINING MODELS

Data Mining Model	True Positive TP (a)	True Negative TN (d)	False Positive FP (c)	False Negative FN (b)	Accuracy of Model
Neural	71	1053	48	64	71.85194
Regression	60	1070	31	75	60.8657
Auto Neural	69	1067	34	66	69.86327
Decision tree	81	1045	56	54	81.84547
SVM	34	1080	21	101	34.87379
Random Forest	21	1091	10	114	21.88269

The above shown table 1 is the confusion matrix of the entire DM models used in this paper.

Total accuracy is given by,

Accuracy = TP+TN/(TP+TN+FP+FN)

Sensitivity, specificity and other statistical measures formulas is given by;

Sensitivity	$\frac{A}{a + b}$	**Specificity**	$\frac{d}{c + d}$
Positive Likelihood Ratio	$\frac{\text{Sensitivity}}{1 - \text{Specificity}}$	**Negative Likelihood Ratio**	$\frac{1 - \text{Sensitivity}}{\text{Specificity}}$
Positive Predictive Value	$\frac{a}{a + c}$	**Negative Predictive Value**	$\frac{d}{b + d}$

From the above table 1, it can be observed that decision tree is the best performing model with an overall accuracy of 81.84%. The least performing model is the random forest with 21.88% of accuracy.

4.2 MEAN SQUARED ERROR AND MISCLASSIFICATIO RATE

In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an model measures the average of the squares of the errors—that is, the average squared difference between actual and estimated values. MSE is a risk function, corresponding to the expected value of the squared error loss. The below table (2) shows, the average squared error and misclassification rate of all the data mining models. Misclassification rate determines how often the model predicted misclassified the outcome variable.

TABLE 2 AVERAGE SQUARED ERROR AND MISCLASSIFICATION RATE

Data Mining Model	Average Squared Error	Misclassification rate
AutoNeural	0.057	0.809
Neural Network	0.059	0.090
Regression	0.060	0.085
Decision tree	0.062	0.089
Random Forest	0.063	0.100
SVM	0.084	0.098

From table 2, it is clearly observed that SVM has the highest squared error with 0.084 and Random forest has the highest misclassification rate with 0.100.

4.3 ROC CURVE STATISTICS

The below figure (13) shows the ROC curve of decision tree, Linear Regression, Neural Network, Support vector machine, AutoNeural network. From the Roc curve it could be observed that, Decision tree is performing better than all other models in consideration. The ROC curve of Decision tree is closer to the baseline model, than other data mining model. And all other DM models perform the same when ROC curve is taken into consideration.

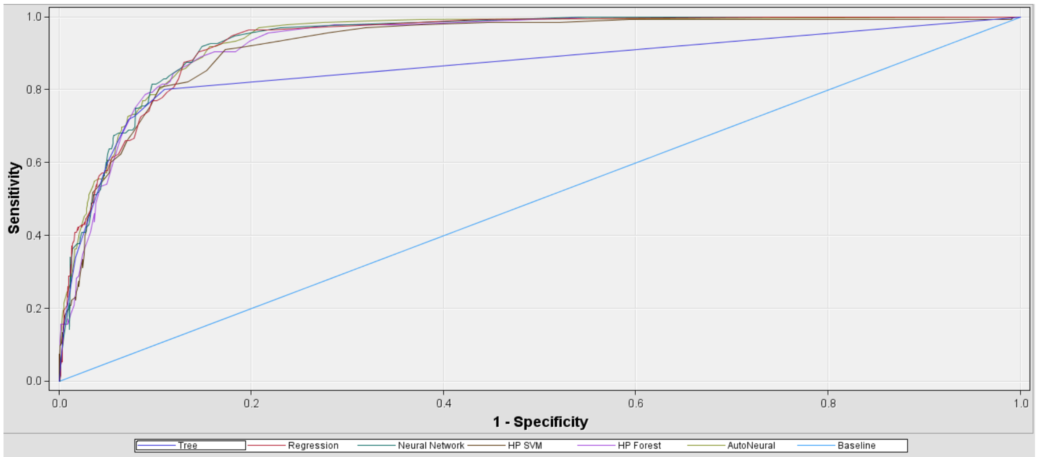
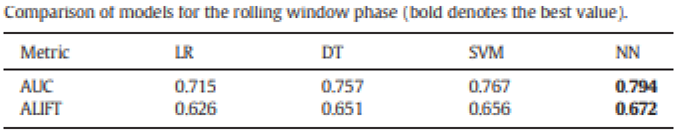


FIGURE 14: ROC CURVE OF DM MODEL(S)

4.4 RESULTS PRESENTED BY PRIMARY PAPER

The research paper by Sérgio Moro, Paulo Cortez , Paulo Rita on their research paper ’A data-driven approach to predict the success of bank telemarketing’ in which a data mining approach is proposed to predict the success of selling bank long-term deposits through telemarketing calls. The data addressed by the researchers are collected from 2008 to 2013 of the Portuguese bank marketing dataset. They analyzed the whole dataset with 150 features including product, bank clients, and social-economic attributes which has several client information. They used a semi-automatic feature selection in the modeling phase with feature reduction they performed the data mining modeling with 22 features. They used four data mining models i.e, logistic regression, neural network, decision trees and support vector machine and compared these models to analyze which data mining performs best on the bank marketing dataset. Using two metrics, area of the receiver operating characteristic curve (AUC) and area of the LIFT cumulative curve (ALIFT), the four models were tested , and results were produced which data mining performs the best. The neural network performed the best results (AUC = 0.8 and ALIFT = 0.7) according to the test performed by Sergio Moro, Paulo Cortez , Paulo Rita , allowing to reach 79% of the subscribers by selecting the half better classified clients.

TABLE 3 COMPARISON OF MODEL PRESENTED



Comparison of models for the rolling window phase (bold denotes the best value).

Metric	LR	DT	SVM	NN
AUC	0.715	0.757	0.767	**0.794**
ALIFT	0.626	0.651	0.656	**0.672**

5. RESULTS & DISCUSSION

The decision tree has been the best performing model in this paper and random forest has been the worst performing model. It is observed a huge difference in accuracy of decision tree and random forest which is unusual as random forest is formation of multiple decision trees. Therefore it is significant that, the random forest is having the least accuracy due to over fitting of model. Support vector machine accuracy is also significantly lower than linear regression; it is also due to over fitting on model. Though autoneral is an optimized version of neural network in SAS enterprise miner, it performs with lesser accuracy than the neural network.

According to the results produced on SAS enterprise miner, illustrates that decision tree performs the best on the validation datasets with an accuracy of 81.84%. That means 81.84% of the time the bank can contact the selected clients who will subscribe the long term deposit plan. In contrast, the best selected model according to Sergio Moro, Paulo Cortez , Paulo Rita proposed on their research paper; Neural network is the best model with area under curve AUC value = 0.8 and cumulative LIFT curve, ALIFT value = 0.7) . According to their paper, the neural network model can perform, allowing reaching 79% of the subscribers by selecting the half better classified clients. Whereas, the decision tree model proposed in this paper can predict the subscriber of bank deposit term with an accuracy of 81%, which is relatively better. Even, decision tree are better in terms of an algorithm as its not a black box algorithm which has a justified output and eligible to be used to certain extent according to EU General data protection regulation (EU GDPR). The difference in proposed model by this paper and by Sérgio Moro, Paulo Cortez , Paulo Rita paper is due to difference in their data set as we have taken a set of sample data which had only 21 variables but, the paper proposed by Sérgio Moro, Paulo Cortez , Paulo Rita has 150 variables as input and they have gone undergone feature selection and feature reduction which is slightly different from out data. Therefore, the results proposed in this paper are different from their paper.

6. CONCLUSION

With the banking industry growing, optimizing telemarketing by targeting particular customers is the goal of the banks want to accomplish, under growing pressure to reduce costs and increase profits. This paper is proposes that decison tree is the best performing data mining models using SAS enterprise miner which will predict the success of bank telemarketing campaign where the clients subscribing to the long-term deposit can be predicted with an overall accuracy of 81.87%. The comparison and discussion of model has been presented in this paper and the results proposed by Sérgio Moro, Paulo Cortez , Paulo Rita (2014) in their paper 'A data-driven approach to predict the success of bank telemarketing' has also been addressed and the contrast features between this and the authors paper has also been discussed. All key attributes has been discussed briefly which affects the outcome variables in prediction and how SAS enterprise miner does feature selection, engineering and uses DM model for optimal results are discussed in this paper.

7. REFERENCE

1. George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a.) Boston: Pearson.
2. Sérgio Moro, Paulo Cortez , Paulo Rita (2014). A data-driven approach to predict the success of bank telemarketing. *Elsevier*, 62, 22-31, doi:10.1016/j.dss.2014.03.001