Lab Report Assignment

P2777638

4/19/24

CISP5302

Contents

Abstract

This research is based on an analysis of a bank dataset provided by a Portuguese financial institution. The dataset includes direct marketing campaigns that are mostly carried out over the phone. Using SAS Enterprise Miner, the primary task is to create a technical report that includes the best model for determining the factors influencing bank term deposit subscription. The Decision Tree was outperformed by the logistic regression model. Prior to producing Logistic Regression and Decision Trees, SAS Enterprise Miner is used in the study to perform EDA, data imputation, data transformation, and variable selection. The paper concludes with recommendations and potential applications of data mining in the future.

Business Problem

A Portuguese financial institution intends to enhance its direct marketing to term deposit customers by utilizing the BANK dataset, which consists of direct marketing campaigns that are mainly carried out over the phone. The objectives include determining the target audience's key styles and actions to generate a direct reaction or level of engagement. Initially, the variables are cleaned, processed, transformed, and analysed using exploratory analysis. Predictive modelling employs a variety of techniques, all of which are assessed according to performance criteria.

Methodology - SEMMA framework

A well-liked data mining methodology called SEMMA provides a methodical approach for gleaning insightful information from massive datasets.
SEMMA is divided into five main stages:

1. Sample: It is essential to select an appropriate sample size and make sure it accurately represents the total data.

2. Explore: Explore includes methods such as computing summary statistics, visualizing data, and spotting early patterns or trends.

3. Modify: Before raw data can be used for modeling, it frequently needs to be prepared. In this stage, missing values, inconsistencies, and outliers are addressed in order to clean up the data. To enhance the model's performance, data transformation techniques may also be used.

4. Model: This is where primary data mining takes place. Based on the particular issue we're attempting to solve, we will choose and construct a model. For the bank dataset, we will choose classification model.

5. Assess: Lastly, we analyze the model's effectiveness and usefulness. This includes metrics, graph that are relevant to the issue at hand.

Data Exploration

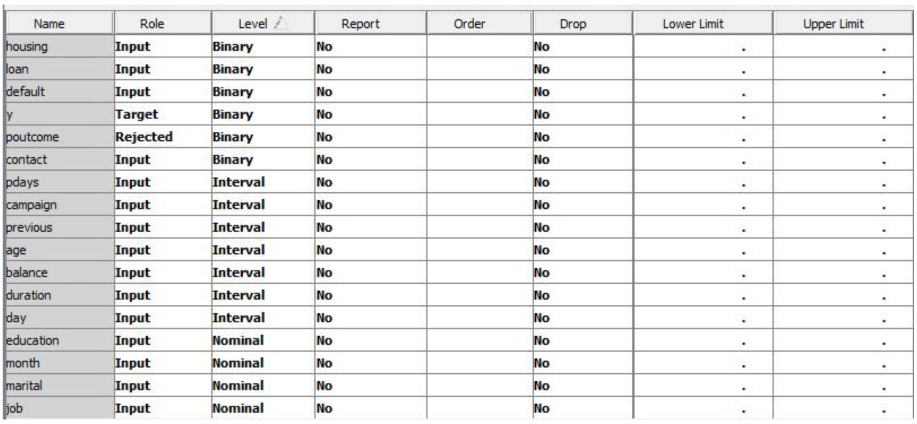
Model Roles: There are seventeen variables in all. The variables are all input variables, with the exception of y, which is the target label and poutcome, the rejected variable.

Data Type/Measurement levels: The three types of measurement that we have are Binary, Interval, and Nominal.

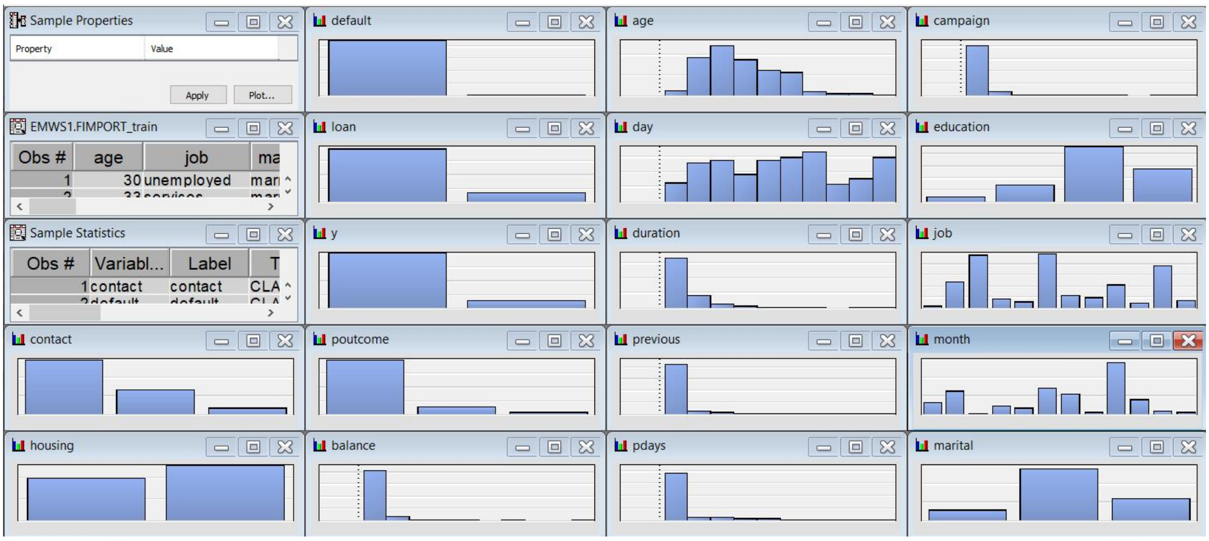
Housing, loan, default, Y, and contract are the binary variables because they only have two different non-missing levels. They're all character variables, too.

The following seven variables-pdays, campaign, previous, age, balance, duration, and day-are all numeric variables in the SAS datasets with more than 20 distinct levels, making them interval measurement level variables.

Month, work, marital status, and education have all been given nominal values. The groups of categorical data those variables represent. Additionally, they have between two and twenty non-missing distinct levels.



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
housing	Input	Binary	No		No	.	.
loan	Input	Binary	No		No	.	.
default	Input	Binary	No		No	.	.
y	Target	Binary	No		No	.	.
poutcome	Rejected	Binary	No		No	.	.
contact	Input	Binary	No		No	.	.
pdays	Input	Interval	No		No	.	.
campaign	Input	Interval	No		No	.	.
previous	Input	Interval	No		No	.	.
age	Input	Interval	No		No	.	.
balance	Input	Interval	No		No	.	.
duration	Input	Interval	No		No	.	.
day	Input	Interval	No		No	.	.
education	Input	Nominal	No		No	.	.
month	Input	Nominal	No		No	.	.
marital	Input	Nominal	No		No	.	.
job	Input	Nominal	No		No	.	.



Sample Properties

Property	Value

EMWS1.FIMPORT_train

Obs #	age	job	ma
1	30	unemployed	mar

Sample Statistics

Obs #	Variabl...	Label	T
1	contact	contact	CLA

Balanced/Imbalanced dataset: Upon analyzing the target variable Y, it turns out that there is an imbalance in the dataset. The percentage of clients who do not subscribe is 88.48%, while 11.52% subscribe.

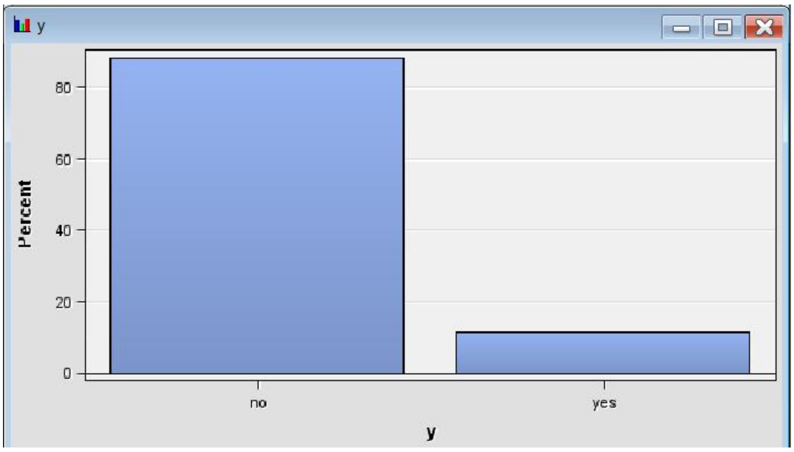
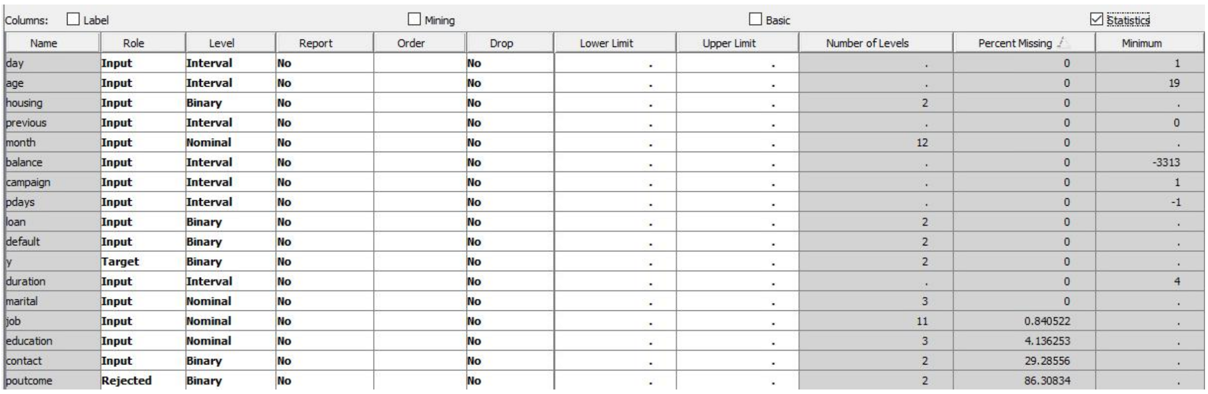


Fig: Imbalanced data

Missing data: There are missing values for the following four variables: contract, job, education, and poutcome. Poutcome has more than 50%, or approximately 86.3%, missing values out of all of them. Poutcome variable has been assigned as rejected because of this. The remaining three variables have less than 50% of missing values, so we have kept them.

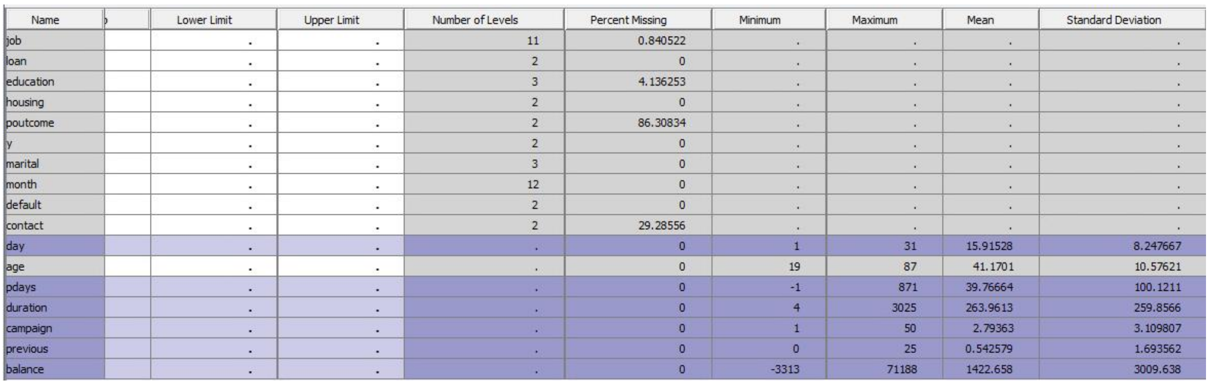


Columns: □ Label □ Mining □ Basic ☑ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum
day	Input	Interval	No		No	.	.	.	0	1
age	Input	Interval	No		No	.	.	.	0	19
housing	Input	Binary	No		No	.	.	2	0	.
previous	Input	Interval	No		No	.	.	.	0	0
month	Input	Nominal	No		No	.	.	12	0	.
balance	Input	Interval	No		No	.	.	.	0	-3313
campaign	Input	Interval	No		No	.	.	.	0	1
pdays	Input	Interval	No		No	.	.	.	0	-1
loan	Input	Binary	No		No	.	.	2	0	.
default	Input	Binary	No		No	.	.	2	0	.
y	Target	Binary	No		No	.	.	2	0	.
duration	Input	Interval	No		No	.	.	.	0	4
marital	Input	Nominal	No		No	.	.	3	0	.
job	Input	Nominal	No		No	.	.	11	0.840522	.
education	Input	Nominal	No		No	.	.	3	4.136253	.
contact	Input	Binary	No		No	.	.	2	29.28556	.
poutcome	Rejected	Binary	No		No	.	.	2	86.30834	.

Fig: Missing data

Variance and standard deviation: A variable is considered to have a high standard deviation if its value is greater than half of the mean. This indicates that the variable's values are dispersed over a large range, which will make it challenging to effectively interpret the data. Transforming the variable can help address this issue and make the data more suitable for analysis. Day, pday, duration, campaign, previous, and balance are needed to be transformed.



Name		Lower Limit	Upper Limit	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation
job		.	.	11	0.840522	.	.	.	.
loan		.	.	2	0	.	.	.	.
education		.	.	3	4.136253	.	.	.	.
housing		.	.	2	0	.	.	.	.
poutcome		.	.	2	86.30834	.	.	.	.
y		.	.	2	0	.	.	.	.
marital		.	.	3	0	.	.	.	.
month		.	.	12	0	.	.	.	.
default		.	.	2	0	.	.	.	.
contact		.	.	2	29.28556	.	.	.	.
day		.	.	.	0	1	31	15.91528	8.247667
age		.	.	.	0	19	87	41.1701	10.57621
pdays		.	.	.	0	-1	871	39.76664	100.1211
duration		.	.	.	0	4	3025	263.9613	259.8566
campaign		.	.	.	0	1	50	2.79363	3.109807
previous		.	.	.	0	0	25	0.542579	1.693562
balance		.	.	.	0	-3313	71188	1422.658	3009.638

Fig: Standard Deviation

Skewness: This Bank dataset have five variables(pday, duration, campaign, previous, and balance) which are positively skewed. These variables need to be transformed close to the normal distribution.

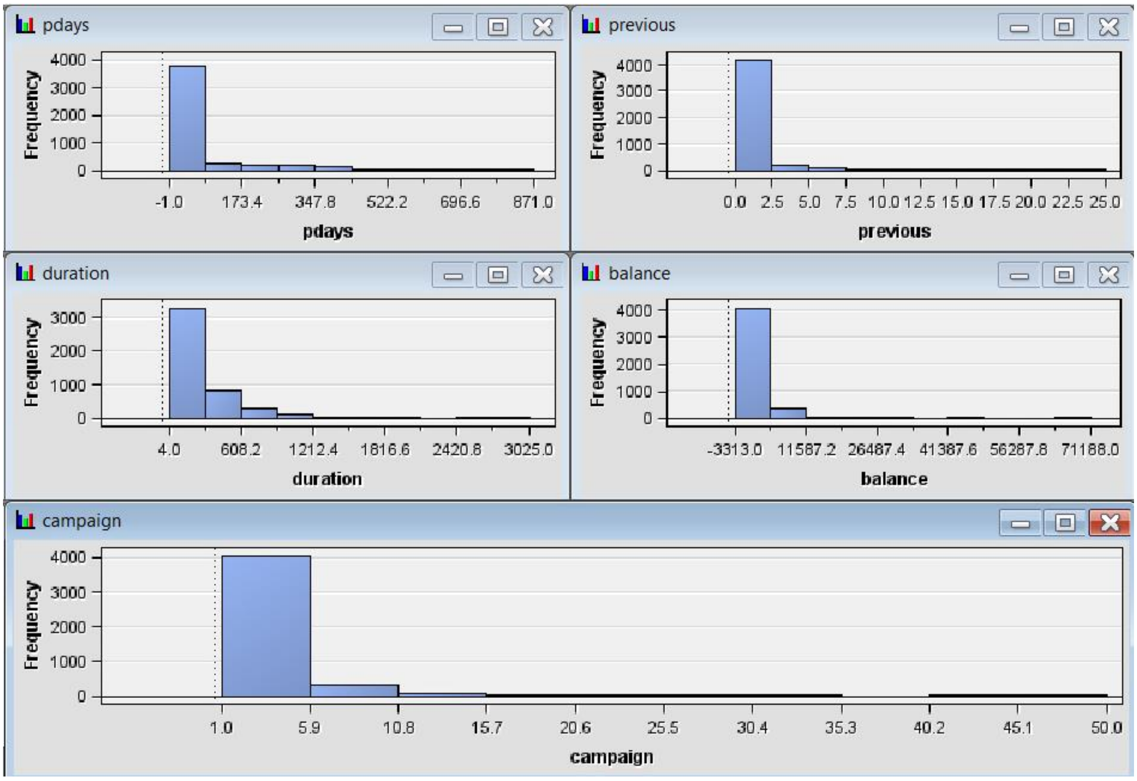


Fig: Skewed Data

Outliers: Variables such as campaign, balance, age and duration have outliers detected using Box plot.

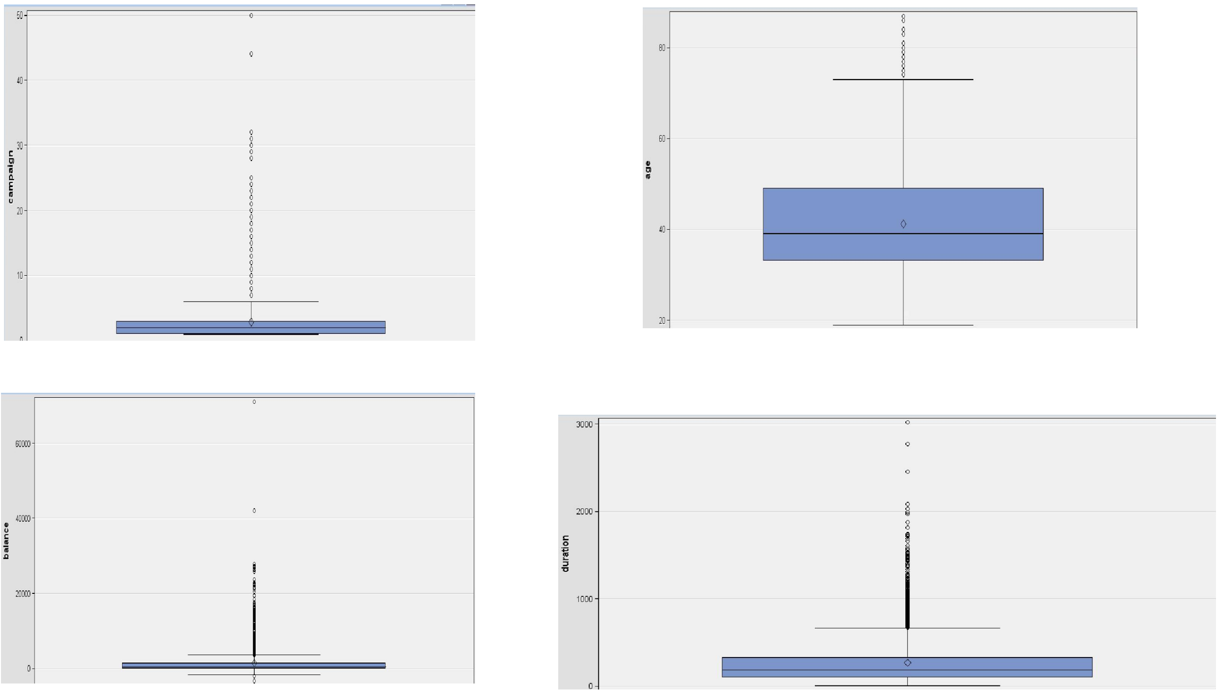


Fig: Outliers Boxplot

Multicollinearity: To detect correlation among predicted variables, we can go with matrix option from the Graph Explorer 01. There are 7 variables in this figure which will help us to get know about correlation with each other.

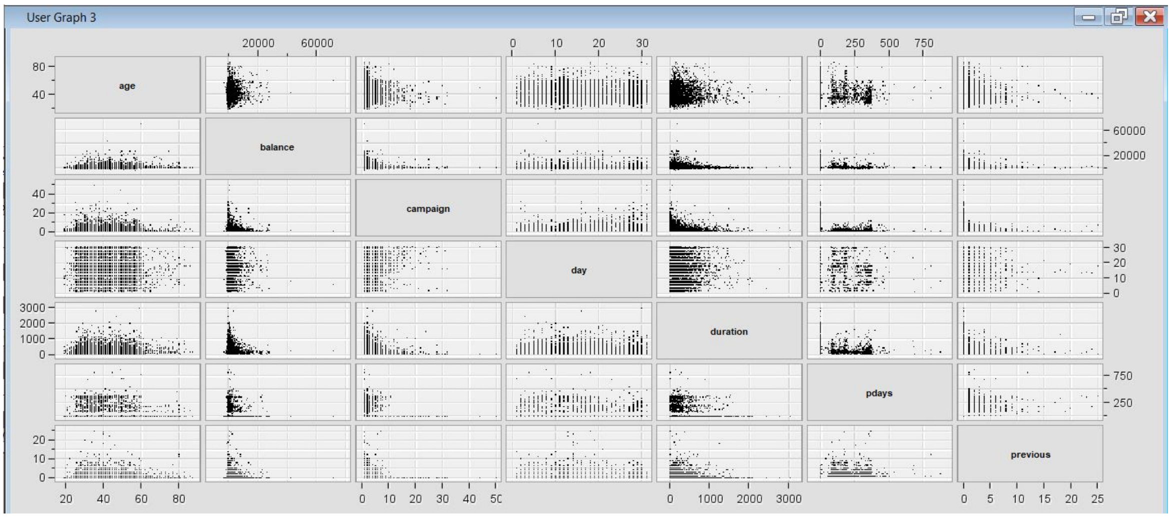
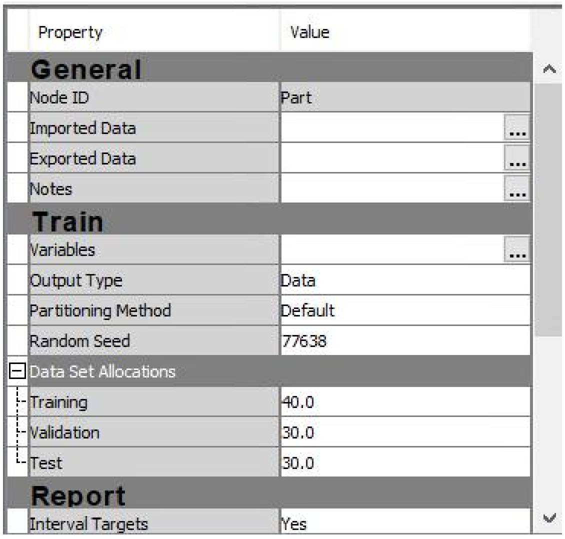


Fig: Matrix (Multicollinearity)

Data partition creation of model sets

Data partition is a necessary step before using data mining and machine learning model . The bank dataset is divided into three categories: 40% for training, 30% for validation, and 30% to facilitate model training, evaluation, and validation. For the random seed generator, I used the last five digit(77638) of my DMU student ID number.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	77638
⊟ Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0
Report	
Interval Targets	Yes

Fig: Data partition Property

Data Modification

In order to preserve the quality of the data, we must first deal with missing values. Three variables have values that are missing. The missing values are handled by the impute nodes from the modify. I chose the tree as the default input method for the property's class and interval variables. Next, we have data with a high standard deviation, positively skewed data, and outliers. To change the data distribution, I used the transform variables node from the modify. Moreover, I have changed the distribution's nature by using mathematical formulas. I have utilized the variable's square root and natural logarithm (Log10).

TR_duration = LOG10(duration)
TR_campaign = LOG10(campaign)
TR_pdays = SQRT(pdays)
TR_previous = Log10(previous)
TR_balance = LOG10(SQRT(balance))
TR_day = SQRT(day)

Data Modelling

I have selected regression and decision tree models. I have used more than four variation for each model.

Regression

We have two different regression: Liner regression and Logistic Regression. I have chosen logistic regression as a first model for this project.

Development of models

I have used 10 regression variation.

The first three variation: default, backward and forward have been built and connected to the Impute node which was used to handle missing values.

Next three variation: default, backward, and stepwise have been connected to the transform node which was used to handle the data distribution.

There are four variable selections node with two variation of target model: R-square and chi-square. First two is connected to impute and second two is with transform node.

These four variable selections are connected to four regression model with backward variation.

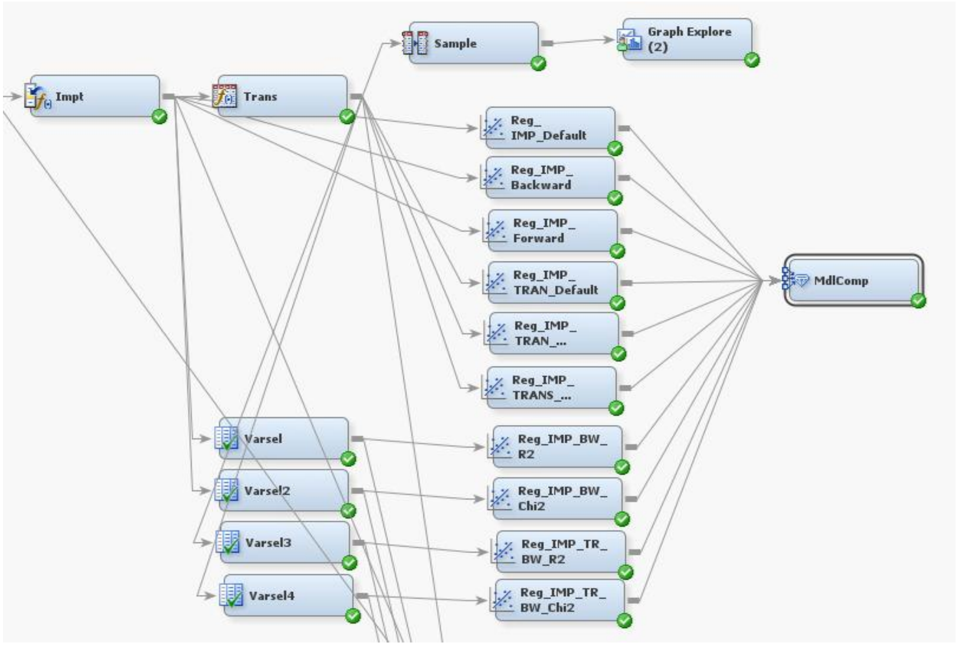


Fig: Regression Model

Model performance

Table 01: Regression models performance

Model variations	ROC Index	Cu. Lift	Scope %	True – %	False – %	True + %	False + %
Reg Imp Default	**0.853**	3.238	20	71.153	96.413	28.847	3.587
Reg Imp BW	0.846	3,366	20	71.79	96.83	28.205	3.169
Reg Imp FW	0.848	3.366	20	8.681	91.319	52.27	47.727
Reg Imp Trans Default	0.631	1.91	20	79.48	97.83	20.51	2.169
Reg Imp Trans BW	0.631	1.91	20	79.48	97.83	20.51	2.169
Reg Imp Trans SW	0.796	2.757	20	54.487	89.407	45.512	10.592
Reg Imp BW R2	0.809	2.917	20	83.33	97.164	16.67	2.836
Reg Imp BW Chi2	0.85	3.33	20	73.08	96.67	26.92	3.33
Reg Imp Trans BW R2	0.686	2.28	20	95.51	99.17	4.49	0.834
Reg Imp Trans BW Chi2	0.848	3.14	20	67.94	96.74	32.06	3.26

From the table, Reg Imp Default model has the best ROC index among all the models.

Now we have cumulative lift chart to observe the model performance. At the 20% scope, the cumulative lift is 32.38 that means the model performance is more than 3 times better than baseline model and almost 2 times better at 40%.

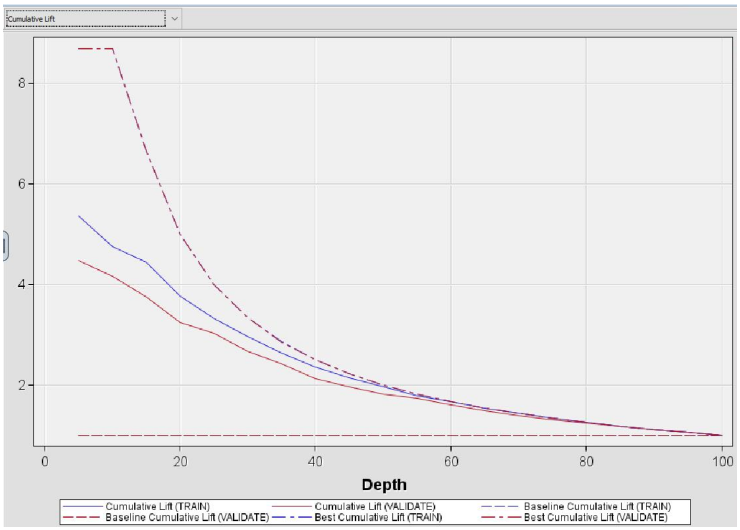


Fig: Cumulative Lift

The % response chart shows that the model performance is good. Approximately at the scope 30% depth, the model is no longer predictive.

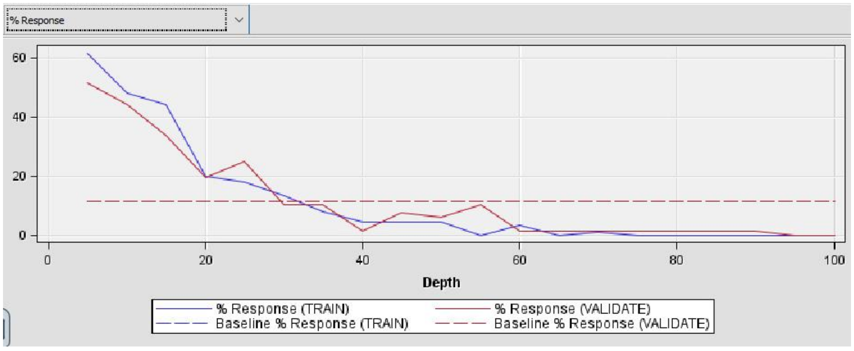


Fig: % Response chart

We know the higher AUC(The area under the ROC curve) means the better performance. All the models are predictive.

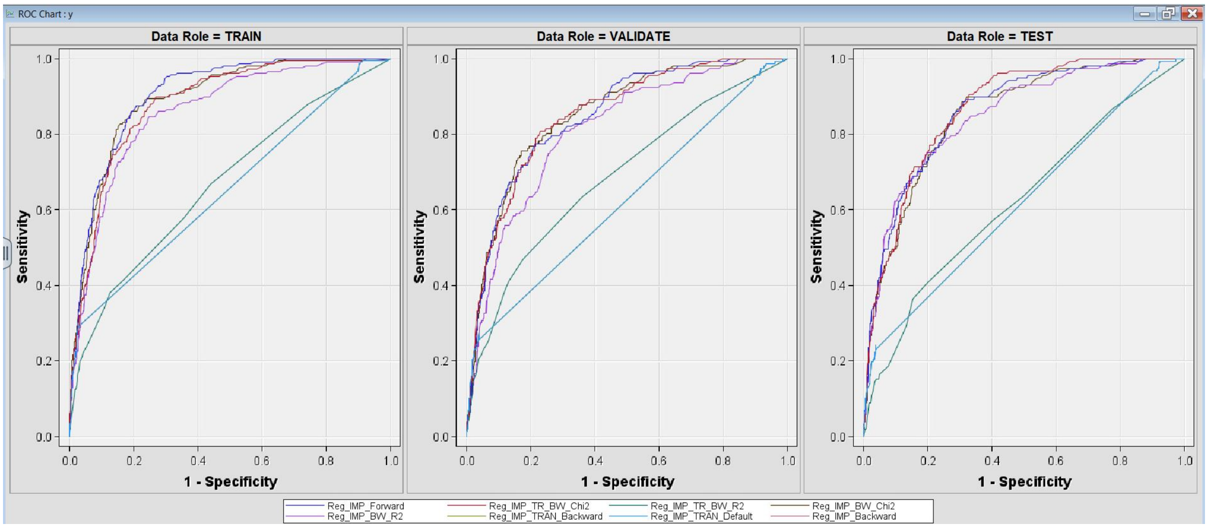


Fig: ROC Curve

Chosen Regression equation

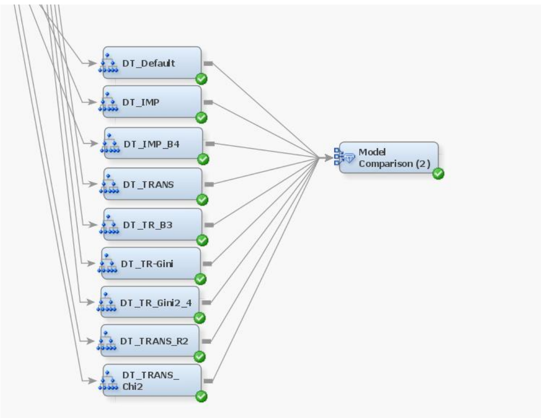
Logit(p) = -3.6682 + 1.1796 * IMP_ job + 0.00477 * duration + 0.2977 * housing + 0.3975 * loan + 1.2827 * month (dec) - 1.2466 * month (jan) - 0.4545 * month (jul) - 0.9024 * month (jun) + 1.9667 * month (mar) - 1.099 * month (may) -1.059 * month (nov) + 1.3621 * month (oct) + 0.00404 * pdays + 0.0825 * previous

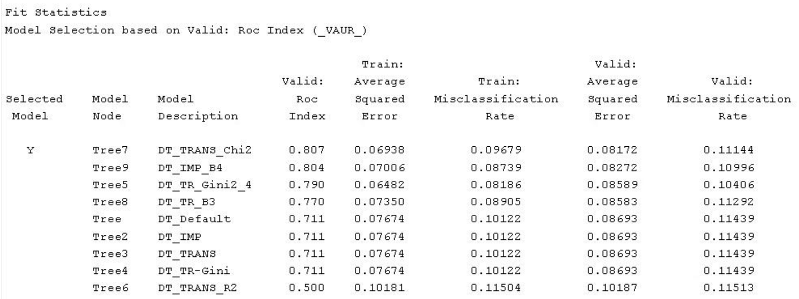
Decision Tree

A decision tree is a popular data mining technique used for predictive modelling and classification tasks. It's a graphical representation of a tree-like structure where each node represents a predictive variable, a decision rule, and each leaf node represents a class label or a continuous value.

Development of models

9 Decision Tree variations have been created. The first model is connected to the data partition. The second and third are with impute. The maximum branch is changed to 4 for the third model. The next four modules are connected to the transform node and the last three among them are modified. The last two are with the two variable selections: R-square and Chi-square.





```
Fit Statistics
Model Selection based on Valid: Roc Index (_VAUR_)

                                                     Train:                              Valid:
                                          Valid:     Average         Train:              Average           Valid:
Selected    Model     Model               Roc        Squared    Misclassification       Squared     Misclassification
 Model      Node      Description         Index      Error           Rate               Error           Rate

   Y        Tree7     DT_TRANS_Chi2       0.807      0.06938        0.09679             0.08172         0.11144
            Tree9     DT_IMP_B4           0.804      0.07006        0.08739             0.08272         0.10996
            Tree5     DT_TR_Gini2_4       0.790      0.06482        0.08186             0.08589         0.10406
            Tree8     DT_TR_B3            0.770      0.07350        0.08905             0.08583         0.11292
            Tree      DT_Default          0.711      0.07674        0.10122             0.08693         0.11439
            Tree2     DT_IMP              0.711      0.07674        0.10122             0.08693         0.11439
            Tree3     DT_TRANS            0.711      0.07674        0.10122             0.08693         0.11439
            Tree4     DT_TR-Gini          0.711      0.07674        0.10122             0.08693         0.11439
            Tree6     DT_TRANS_R2         0.500      0.10181        0.11504             0.10187         0.11513
```

Fig: Decision Tree

Performance of models

Table 2: Decision tree models performance

Model variations	**ROC Index**	**Cu. Lift**	**Scope %**	**True – %**	**False – %**	**True + %**	**False + %**
DT_Default	0.711	2.699	20	84.61	98.081	15.384	1.9183
DT_IMP	0.711	2.699	20	84.61	98.081	15.384	1.9183
DT_IMP_B4	0.804	3.244	20	56.41	94.912	43.589	5.088
DT_Trans	0.711	2.699	20	84.61	98.081	15.384	1.9183
DT_TR_B3	0.77	2.862	20	60.257	95.079	39.743	4.9208
DT_TR_Gini	0.711	2.699	20	84.61	98.081	15.384	1.9183
DT_TR_Gini2_4	0.79	3.189	20	59.616	95.996	40.384	4.0033
DT_TRANS_R2	0.5						
DT_Trans_Chi2	**0.807**	3.227	20	79.49	97.749	20.512	2.251

DT Trans Chi2 performs better than any other variations of Decision tree model. DT_TRANS_R2 are dropped because it didn't perform well.

At 20% scope, the model performs more than 3 time better than baseline model and approx. 2 times better at 40% scope.

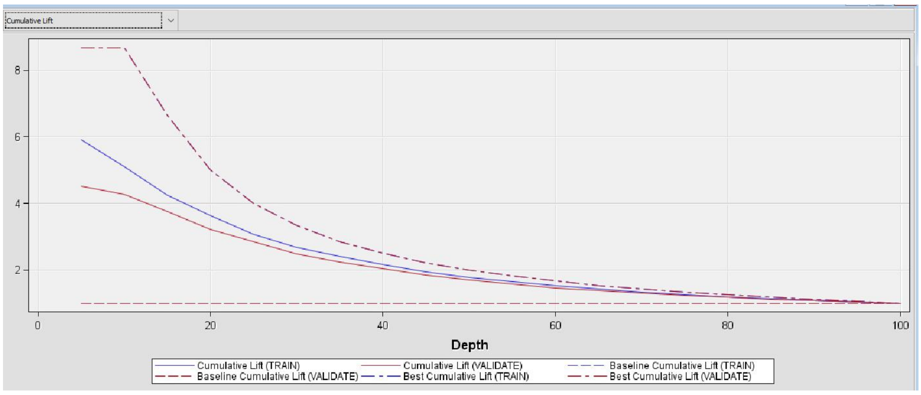


Fig: Cumulative

Approx. At 25% depth, the model lost it's predictability. Hence, the model provided good performance.

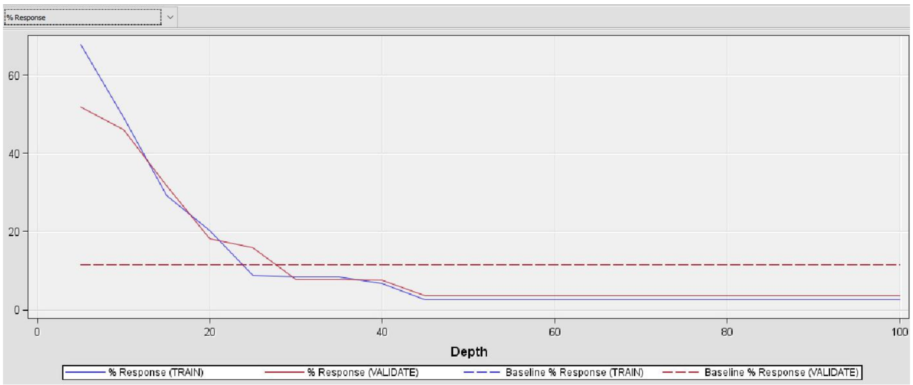


Fig: % Response

From the validate ROC chart, DT_Trans_Chi2 provided better performance.

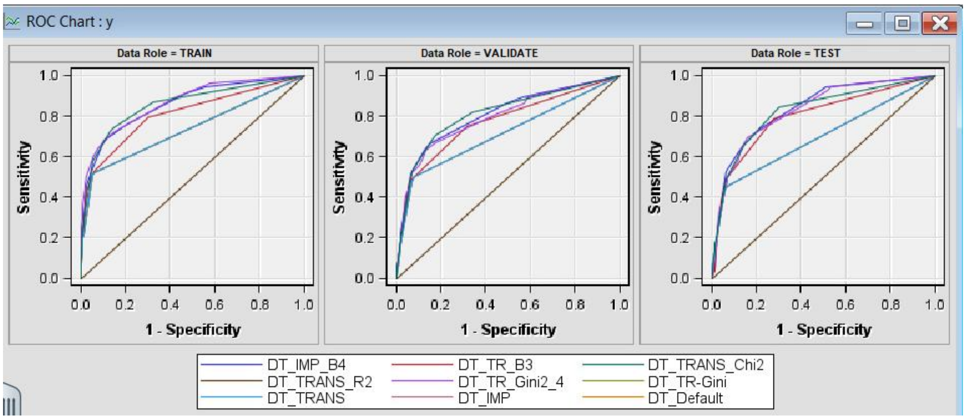
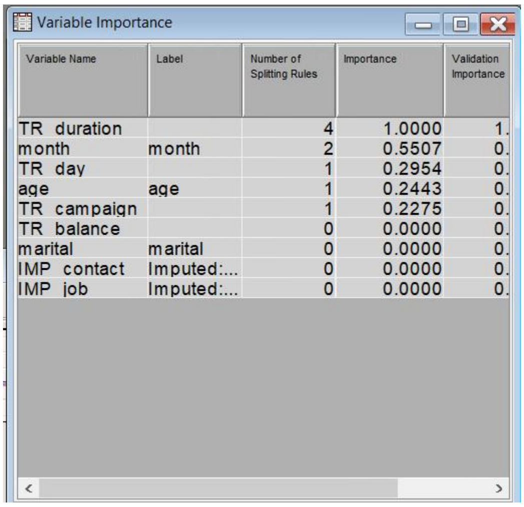


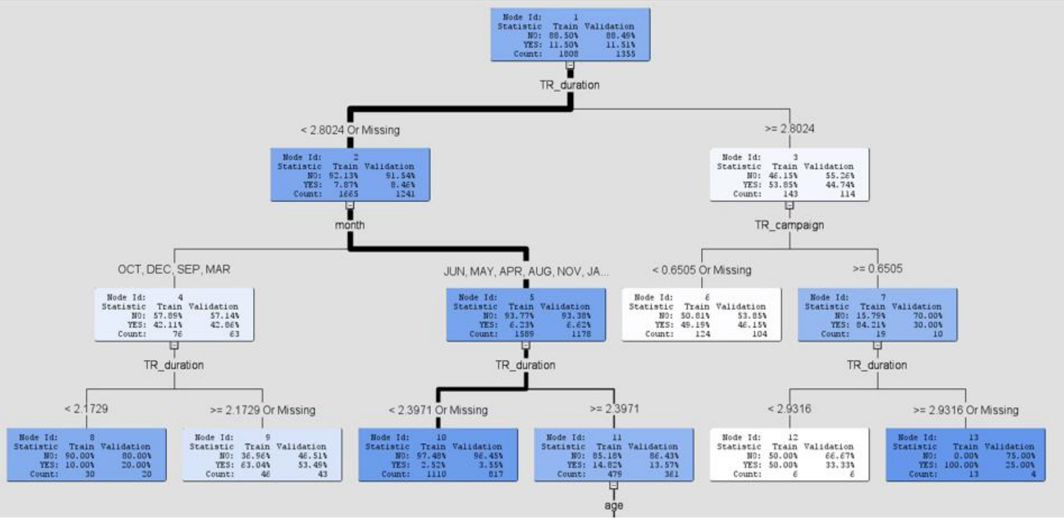
Fig: ROC Curve

We have got five important variable: TR_duration, month, TR_day, age and TR_campaign.



Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance
TR_duration		4	1.0000	1.
month	month	2	0.5507	0.
TR_day		1	0.2954	0.
age	age	1	0.2443	0.
TR_campaign		1	0.2275	0.
TR_balance		0	0.0000	0.
marital	marital	0	0.0000	0.
IMP_contact	Imputed:...	0	0.0000	0.
IMP_job	Imputed:...	0	0.0000	0.

Critical path of best model



Node Id: 1 Statistic Train Validation NO: 88.50% 88.49% YES: 11.50% 11.51% Count: 1808 1355
TR_duration
< 2.8024 Or Missing
>= 2.8024
Node Id: 2 Statistic Train Validation NO: 92.13% 91.54% YES: 7.87% 8.46% Count: 1665 1241
Node Id: 3 Statistic Train Validation NO: 46.15% 55.26% YES: 53.85% 44.74% Count: 143 114
month
TR_campaign
OCT, DEC, SEP, MAR
JUN, MAY, APR, AUG, NOV, JA...
< 0.6505 Or Missing
>= 0.6505
Node Id: 4 Statistic Train Validation NO: 57.89% 57.14% YES: 42.11% 42.86% Count: 76 63
Node Id: 5 Statistic Train Validation NO: 93.77% 93.38% YES: 6.23% 6.62% Count: 1589 1178
Node Id: 6 Statistic Train Validation NO: 50.81% 53.85% YES: 49.19% 46.15% Count: 124 104
Node Id: 7 Statistic Train Validation NO: 15.79% 70.00% YES: 84.21% 30.00% Count: 19 10
TR_duration
TR_duration
TR_duration
< 2.1729
>= 2.1729 Or Missing
< 2.3971 Or Missing
>= 2.3971
< 2.9316
>= 2.9316 Or Missing
Node Id: 8 Statistic Train Validation NO: 90.00% 80.00% YES: 10.00% 20.00% Count: 30 20
Node Id: 9 Statistic Train Validation NO: 36.96% 46.51% YES: 63.04% 53.49% Count: 46 43
Node Id: 10 Statistic Train Validation NO: 97.48% 96.45% YES: 2.52% 3.55% Count: 1110 817
Node Id: 11 Statistic Train Validation NO: 85.18% 86.43% YES: 14.82% 13.57% Count: 479 361
Node Id: 12 Statistic Train Validation NO: 50.00% 66.67% YES: 50.00% 33.33% Count: 6 6
Node Id: 13 Statistic Train Validation NO: 0.00% 75.00% YES: 100.00% 25.00% Count: 13 4
age

Fig: Critical Path

Decision rule for the critical path :

____

Node = 6

____

if TR_duration >= 2.80243 or MISSING
AND TR_campaign < 0.65051 or MISSING
then

Tree Node Identifier = 6
Number of Observations = 124
Predicted: y=yes = 0.49
Predicted: y=no = 0.51

Target path of interest

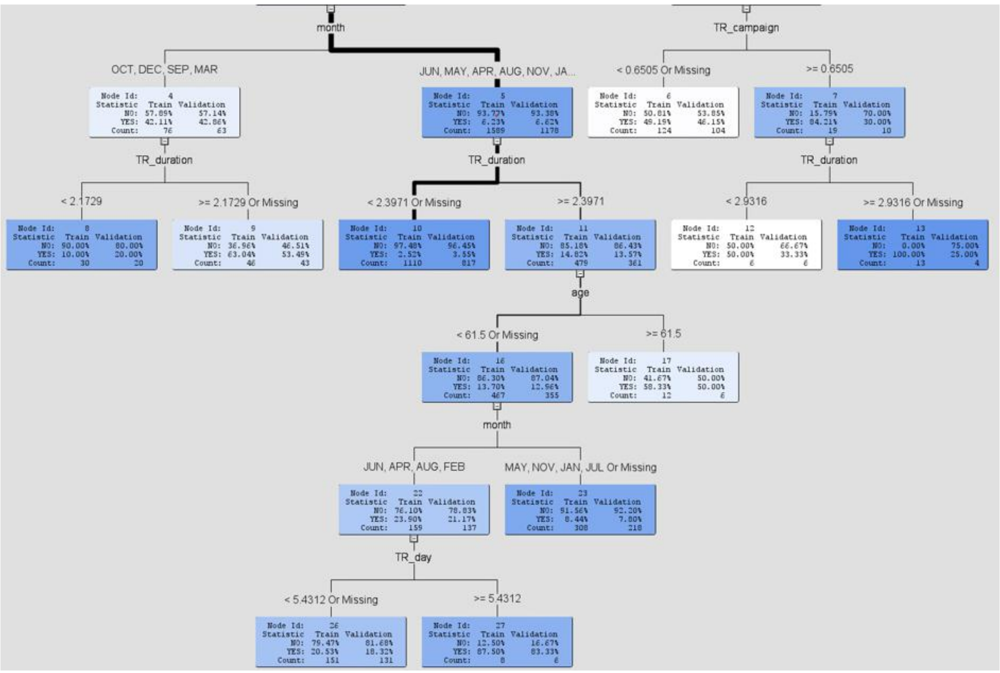

month
TR_campaign
OCT, DEC, SEP, MAR
JUN, MAY, APR, AUG, NOV, JA...
< 0.6505 Or Missing
>= 0.6505
Node Id: 4 Statistic Train Validation NO: 57.89% 57.14% YES: 42.11% 42.86% Count: 76 63
Node Id: 5 Statistic Train Validation NO: 93.77% 93.38% YES: 6.23% 6.62% Count: 1589 1178
Node Id: 6 Statistic Train Validation NO: 50.81% 53.85% YES: 49.19% 46.15% Count: 124 104
Node Id: 7 Statistic Train Validation NO: 15.79% 70.00% YES: 84.21% 30.00% Count: 19 10
TR_duration
TR_duration
TR_duration
< 2.1729
>= 2.1729 Or Missing
< 2.3971 Or Missing
>= 2.3971
< 2.9316
>= 2.9316 Or Missing
Node Id: 8 Statistic Train Validation NO: 90.00% 80.00% YES: 10.00% 20.00% Count: 30 20
Node Id: 9 Statistic Train Validation NO: 36.96% 46.51% YES: 63.04% 53.49% Count: 46 43
Node Id: 10 Statistic Train Validation NO: 97.48% 96.45% YES: 2.52% 3.55% Count: 1110 817
Node Id: 11 Statistic Train Validation NO: 85.18% 86.43% YES: 14.82% 13.57% Count: 479 361
Node Id: 12 Statistic Train Validation NO: 50.00% 66.67% YES: 50.00% 33.33% Count: 6 6
Node Id: 13 Statistic Train Validation NO: 0.00% 75.00% YES: 100.00% 25.00% Count: 13 4
age
< 61.5 Or Missing
>= 61.5
Node Id: 16 Statistic Train Validation NO: 86.30% 87.04% YES: 13.70% 12.96% Count: 467 355
Node Id: 17 Statistic Train Validation NO: 41.67% 50.00% YES: 58.33% 50.00% Count: 12 6
month
JUN, APR, AUG, FEB
MAY, NOV, JAN, JUL Or Missing
Node Id: 22 Statistic Train Validation NO: 76.10% 78.83% YES: 23.90% 21.17% Count: 159 137
Node Id: 23 Statistic Train Validation NO: 91.56% 92.20% YES: 8.44% 7.80% Count: 308 218
TR_day
< 5.4312 Or Missing
>= 5.4312
Node Id: 26 Statistic Train Validation NO: 79.47% 81.68% YES: 20.53% 18.32% Count: 151 131
Node Id: 27 Statistic Train Validation NO: 12.50% 16.67% YES: 87.50% 83.33% Count: 8 6

Fig: Target Path

Overfitting analysis

Validation error became steady after 10 leaves and started increasing at 13 leaves that means it showed overfitting in the model. Hence, we can say 10 leaves is enough for this model.

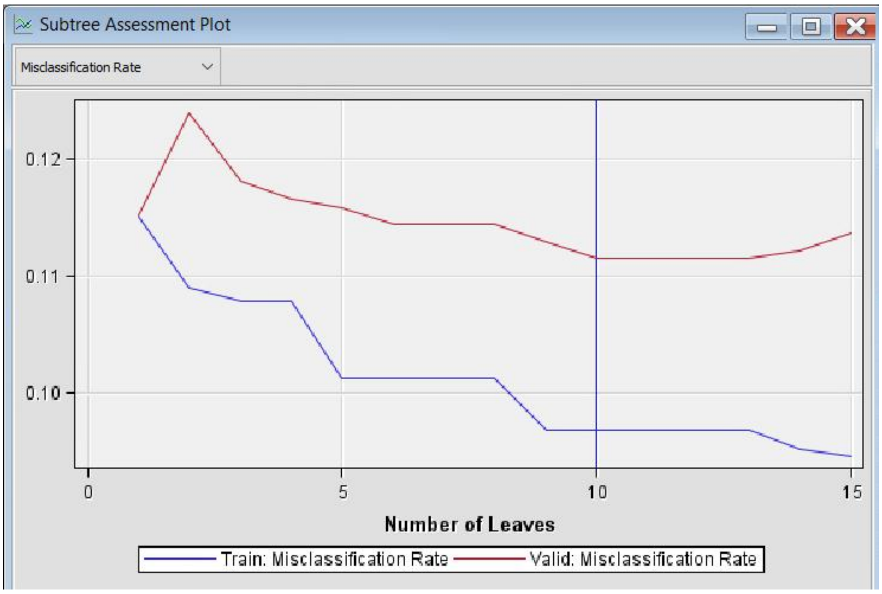


Fig: Misclassification rate

Analysis of the best model

Table X Summary results of the best performing models

Model	ROC Index	Cu. Lift	Scope %	True – %	False – %	True + %	False + %
Reg Imp Default	**0.853**	3.238	20	71.153	96.413	28.847	3.587
DT_Trans_Chi2	0.807	3.227	20	79.49	97.749	20.512	2.251

At the 20% scope, the cumulative value of Reg Imp Default and DT Trans Chi2 are 3.237179 and 3.227148 respectively. They have 2.668 and 2.492 at the 30% depth scope. Hence, Regular Impute Default model performed better than Decision tree Chi-square.

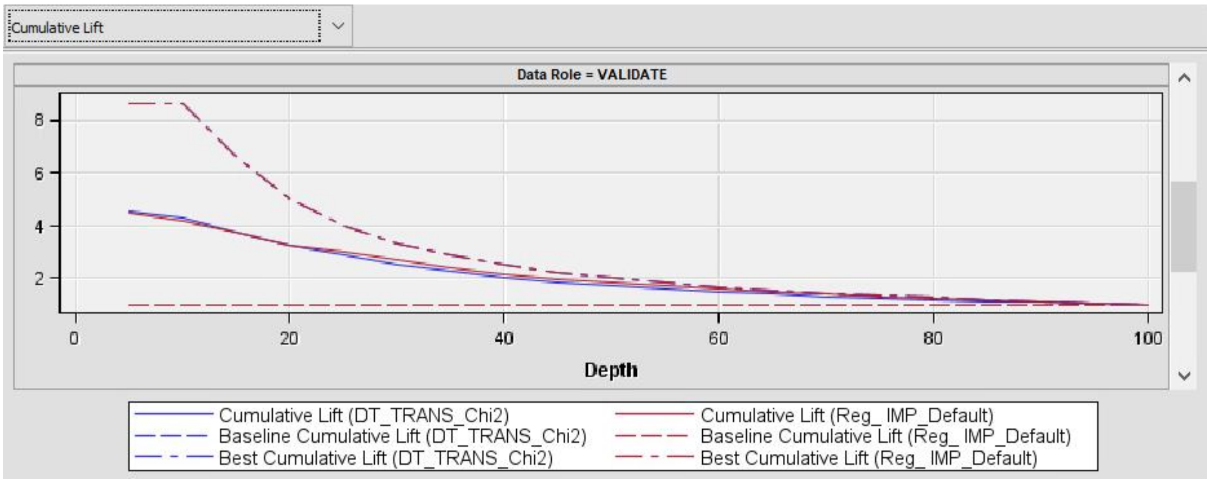


Fig: Cumulative Lift

The AUC of Reg_Imp_Default is more closer to 1 which indicates Reg_Imp_Default provided better performance than DT_Trans_Chi2.

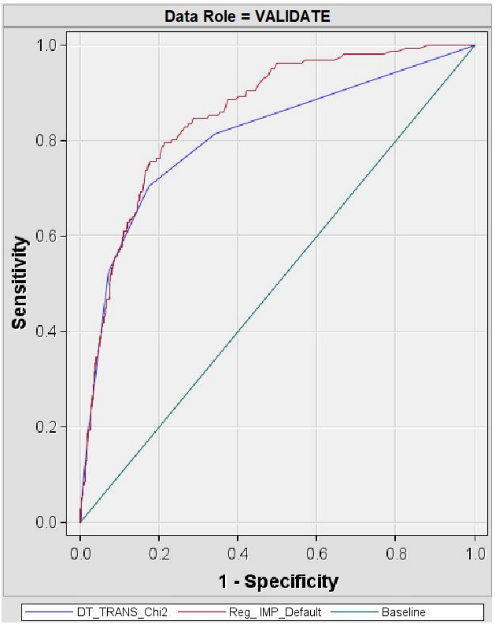
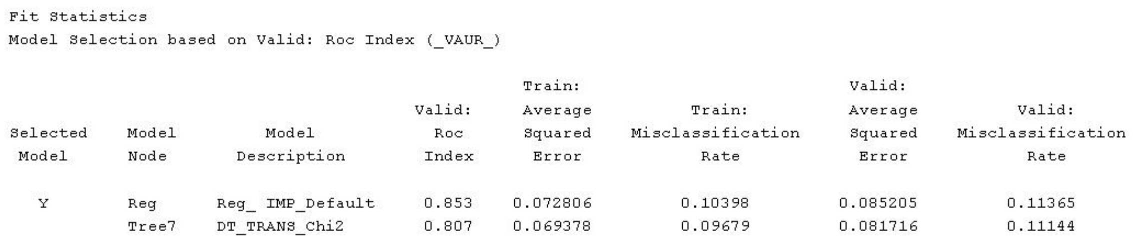


Fig: ROC Curve



```
Fit Statistics
Model Selection based on Valid: Roc Index (_VAUR_)

                                                        Train:                                  Valid:
                                             Valid:    Average          Train:                 Average           Valid:
Selected      Model          Model            Roc      Squared     Misclassification          Squared     Misclassification
 Model        Node        Description        Index      Error            Rate                  Error            Rate

   Y          Reg        Reg_ IMP_Default    0.853    0.072806         0.10398               0.085205          0.11365
              Tree7      DT_TRANS_Chi2       0.807    0.069378         0.09679               0.081716          0.11144
```

- Discussion on the breadth of areas of application and research in data mining

Finding hidden patterns and making data-driven decisions in a variety of fields have made data mining essential. Although conventional methods such as regression and classification still hold significant value, predictive modeling is a dynamic field. These innovative and exciting trends are pushing the envelope:

1. Explainable AI (XAI): Although black-box models are accurate, their reasoning may be ambiguous. By revealing these models' inner workings, XAI approaches hope to increase confidence and comprehension in the predictions they make. This is important in fields like medicine, where diagnoses must be comprehensible.
2. Ensemble Learning: More reliable and broadly applicable predictions can be produced by combining several models with different learning styles. The accuracy and efficiency of ensemble methods, such as stacking and neural network architectures, are being pushed to new limits.
3. Generative Modeling: These models are able to produce new, realistic samples as well as learn the underlying distribution of the data. This can be used to create artificial data to train models in situations where privacy is a concern, or it can even be used to design new medications or materials.
4. Causal Inference: Methods of causal inference go beyond correlation and aim to identify cause-and-effect connections in data. This makes predictions and interventions more reliable. Applications include anything from analyzing the effects of policy changes to launching focused marketing campaigns.
5. Integration with Domain Knowledge: A growing emphasis is being placed on integrating domain knowledge into prediction models. This may entail using symbolic reasoning in conjunction with machine learning algorithms or incorporating past knowledge about the relationships between variables. As a result, models are produced that are both accurate and consistent with knowledge from the actual world.

Conclusion

We have successfully assisted a Portuguese financial institution in improving its direct marketing strategies for term deposit subscriptions by analysing the bank dataset using the SEMMA framework. To deal with missing values, we used impute, and to deal with outliers and data distribution, we used transformation formula. Additionally, variable selection was employed prior to predictive models, which aids in their simplification and optimization. Regression Imp default offered the best predictive performance with higher cumulative lift and higher area under the (AUC) in the ROC curve when compared to the DT Trans Chi2 model. That means The higher the cumulative lift, the better the model's ability to discriminate between positive and negative instances, making it valuable for decision-making tasks such as targeting marketing efforts, prioritizing leads, or identifying potential risks. This research will help the institution increase their profit and business possibilities.

Recommendation

While doing exploratory data analysis, we need to be careful about outliers, high standard variation and skewness. As there is no specific formula, the best approach depends on the specifics of the data and analysis goals. We can use both logarithmic transformation and square root transformation for variance. For the skewness, we can use logarithmic transformation.

My Reflections on the process – What did I learn from this exercise?

Here's what I learnt:

- The SEMMA Framework: This methodical approach to data mining projects, which is industry standard, organizes the procedure from selecting data to deploying models. This framework serves as the foundation for SAS Enterprise Miner, so I was able to gain experience in all five stages: sample, explore, modify, model, and assess.
- Data Preparation: In order to create high-quality models, data must be cleaned and transformed. Together with data transformation methods, SAS Enterprise Miner provides tools for managing outliers, inconsistent data, and missing values.
- Predictive Modeling: Up until now, I've studied three modeling approaches. There are more, including clustering (assembling related data points), regression (predicting continuous values), classification (predicting categories), and more.
- Model Evaluation: The software offers tools to evaluate the models' performance and spot possible problems like overfitting. It helps in selecting the ideal model for the particular issue.
- Finding Trends and Patterns: SAS Enterprise Miner assists in locating hidden relationships and patterns in the data. Numerous business applications, including

risk assessment, fraud detection, customer segmentation, and targeted marketing campaigns, can benefit from this knowledge.

References and Bibliography

I. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data mining: practical machine learning tools and techniques. Morgan Kaufmann https://www.researchgate.net/publication/251275816_Witten_IH_Frank_E_Data_Mining_Practical_Machine_Learning_Tools_and_Techniques

II. Parr-Rudd, O. (2014). Business analytics using SAS Enterprise Guide and SAS Enterprise Miner: a beginner's guide. SAS Institute. ISBN 9781612907833.

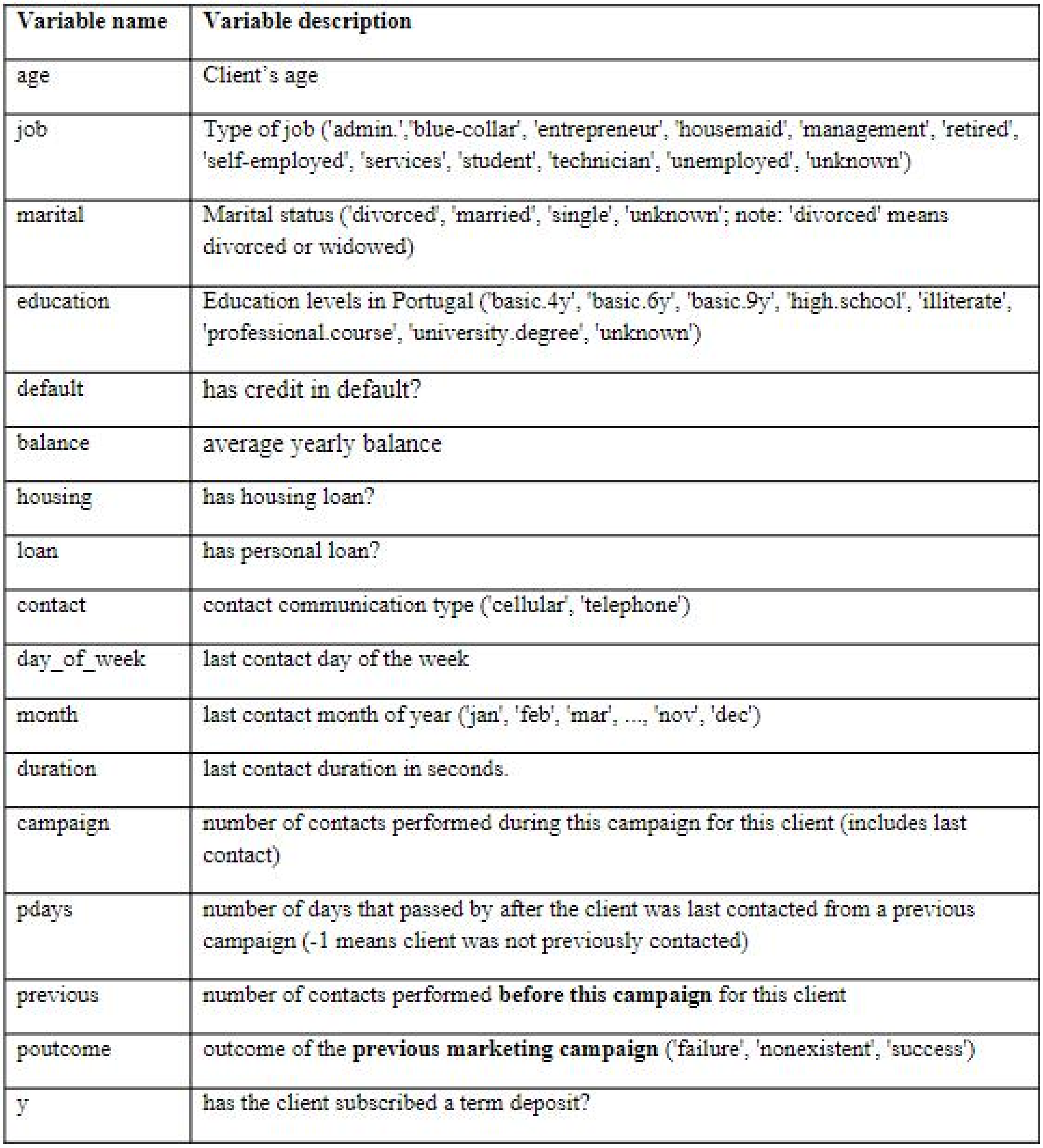
III. Patidar, D. M., & Jain, S. (2017). Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications (2nd ed.). Wiley. ISBN 9781118116197

IV. Frank E. Harrell (2015). Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Springer DOI:10.1007/978-3-319-19425-7

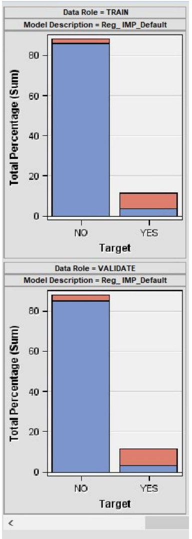
V. Faiz Maazouzi & Halima Bahi (2012). Decision Tree Network Data mining approach. DOI:10.13140/2.1.1047.3921

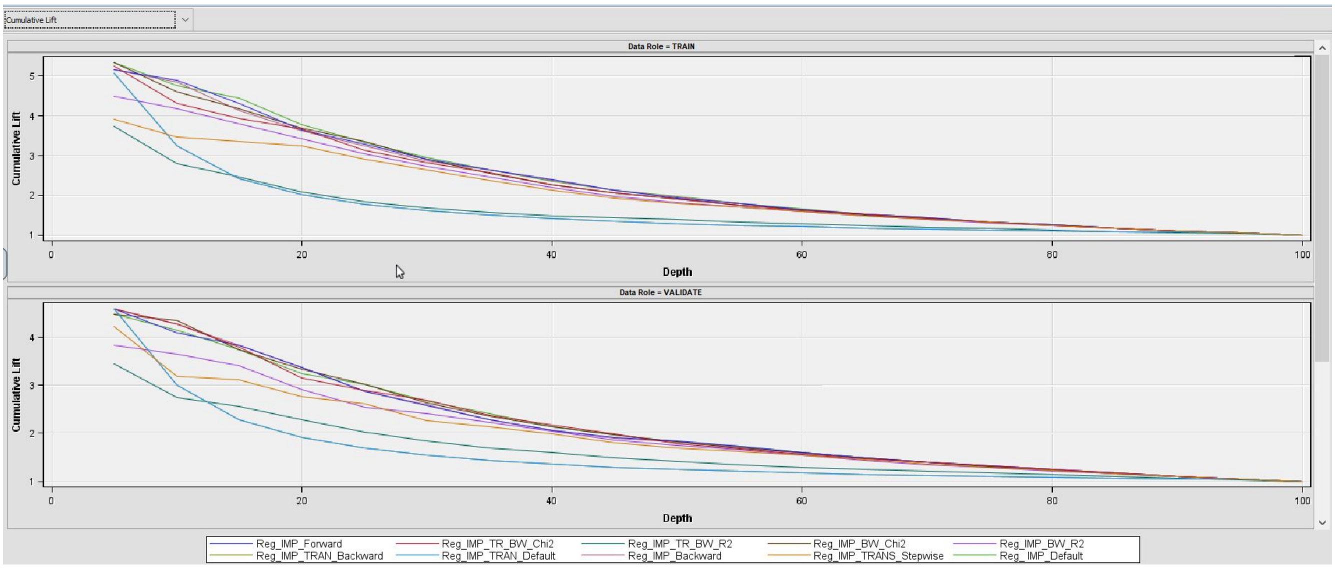
Appendix

Data Mining Roles

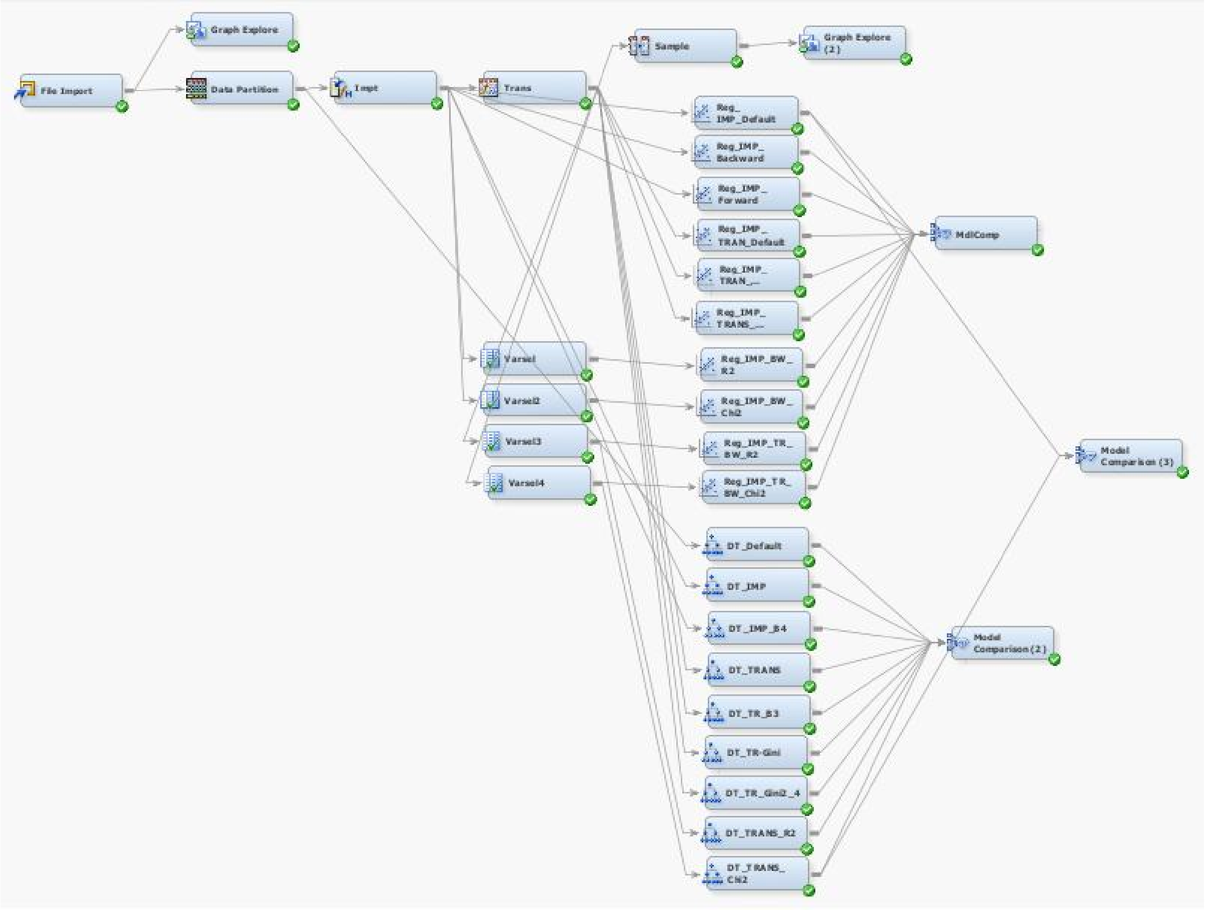


Variable name	**Variable description**
age	Client's age
job	Type of job ('admin.','blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
marital	Marital status ('divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
education	Education levels in Portugal ('basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
default	has credit in default?
balance	average yearly balance
housing	has housing loan?
loan	has personal loan?
contact	contact communication type ('cellular', 'telephone')
day_of_week	last contact day of the week
month	last contact month of year ('jan', 'feb', 'mar', ..., 'nov', 'dec')
duration	last contact duration in seconds.
campaign	number of contacts performed during this campaign for this client (includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted)
previous	number of contacts performed **before this campaign** for this client
poutcome	outcome of the **previous marketing campaign** ('failure', 'nonexistent', 'success')
y	has the client subscribed a term deposit?


Data Role = TRAIN
Model Description = Reg_IMP_Default
Total Percentage (Sum)
0
20
40
60
80
NO
YES
Target
Data Role = VALIDATE
Model Description = Reg_IMP_Default
Total Percentage (Sum)
0
20
40
60
80
NO
YES
Target


Cumulative Lift
Data Role = TRAIN
Cumulative Lift
1
2
3
4
5
0
20
40
60
80
100
Depth
Data Role = VALIDATE
Cumulative Lift
1
2
3
4
0
20
40
60
80
100
Depth
Reg_IMP_Forward
Reg_IMP_TR_BW_Chi2
Reg_IMP_TR_BW_R2
Reg_IMP_BW_Chi2
Reg_IMP_BW_R2
Reg_IMP_TRAN_Backward
Reg_IMP_TRAN_Default
Reg_IMP_Backward
Reg_IMP_TRANS_Stepwise
Reg_IMP_Default

Workflow diagram


File Import
Graph Explore
Data Partition
Impt
Trans
Sample
Graph Explore (2)
Reg_IMP_Default
Reg_IMP_Backward
Reg_IMP_Forward
Reg_IMP_TRAN_Default
Reg_IMP_TRAN_...
Reg_IMP_TRANS_...
Varsel
Varsel2
Varsel3
Varsel4
Reg_IMP_BW_R2
Reg_IMP_BW_Chi2
Reg_IMP_TR_BW_R2
Reg_IMP_TR_BW_Chi2
MdlComp
Model Comparison (3)
DT_Default
DT_IMP
DT_IMP_B4
DT_TRANS
DT_TR_B3
DT_TR-Gini
DT_TR_Gini2_4
DT_TRANS_R2
DT_TRANS_Chi2
Model Comparison(2)