BAN210: Predictive Analytics

Final Project Report

Design and Implementation of Predictive Analytics on Credit Card dataset using SAS Enterprise Miner

Predictive Analytics	
Professor	Uzair Ahmad
Name	Jinalben Patel
Student ID	135354215
Date	15th April 2022

BAN210: Predictive Analytics

Introduction

The goal of this project is to develop and build the necessary components of a data pipeline for predictive modelling on the credit card dataset. The dataset concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. This dataset is interesting because there is a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. There are also a few missing values.

SAS Enterprise Miner has been used to perform all the operations.

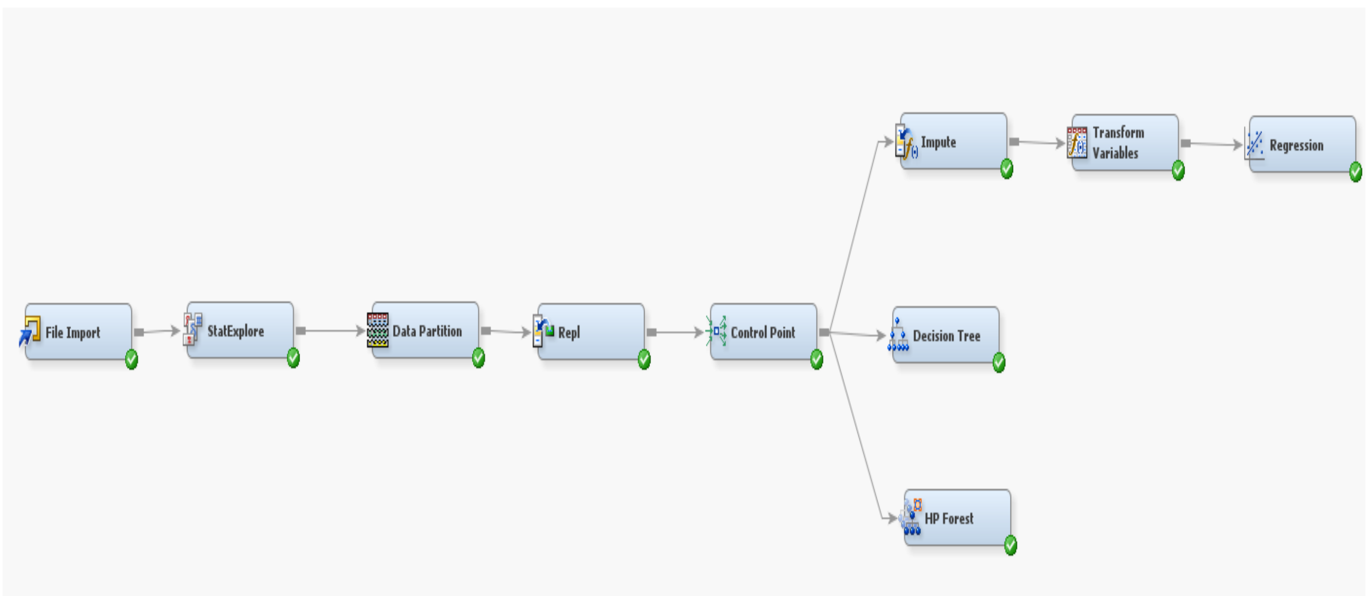
Dataset description:

Variable	Attribute Information
A1	b, a
A2	continuous
A3	continuous
A4	u, y, l, t
A5	g, p, gg
A6	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff
A7	v, h, bb, j, n, z, dd, ff, o
A8	continuous
A9	t, f
A10	t, f
A11	continuous
A12	t, f

BAN210: Predictive Analytics

A13	g, p, s
A14	continuous
A15	continuous
A16	+,- (class attribute)

Diagram:



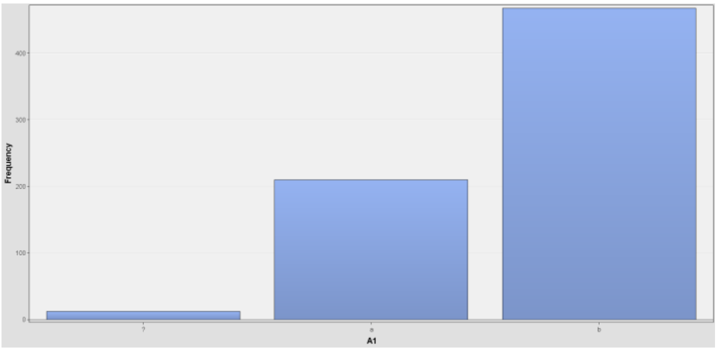
File Import:

It helps to import data from your local computer. The File Import node allows you to customize the data conversion process by selecting the file to import and setting the metadata (such as table and variable roles) that Enterprise Miner needs to perform data mining activities.

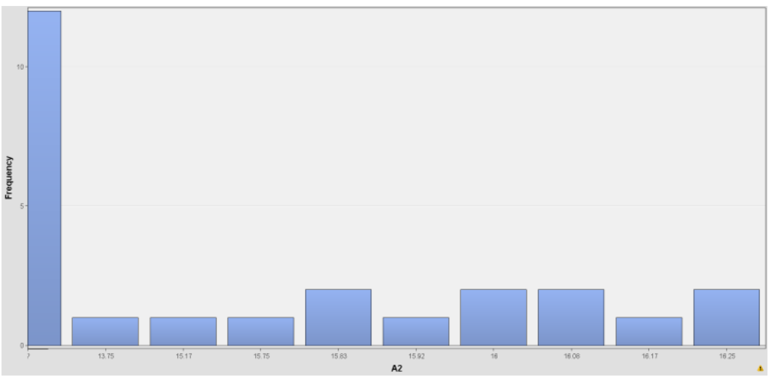
BAN210: Predictive Analytics

Exploring the dataset variables:

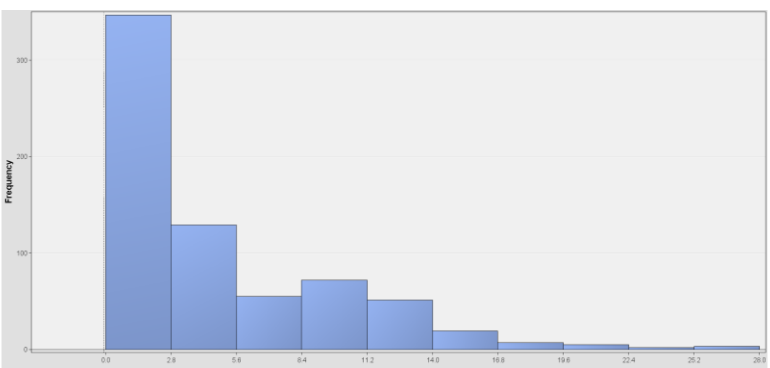
A1: This is a discrete variable with the presence of missing values



A2: This is a discrete variable with the presence of missing values

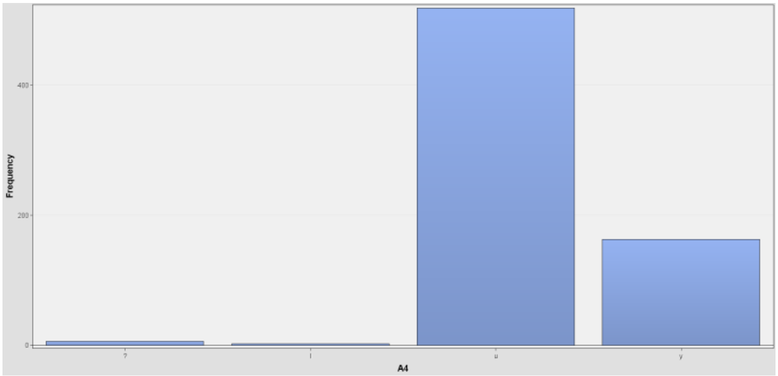


A3: This is a continuous variable with the presence of missing values

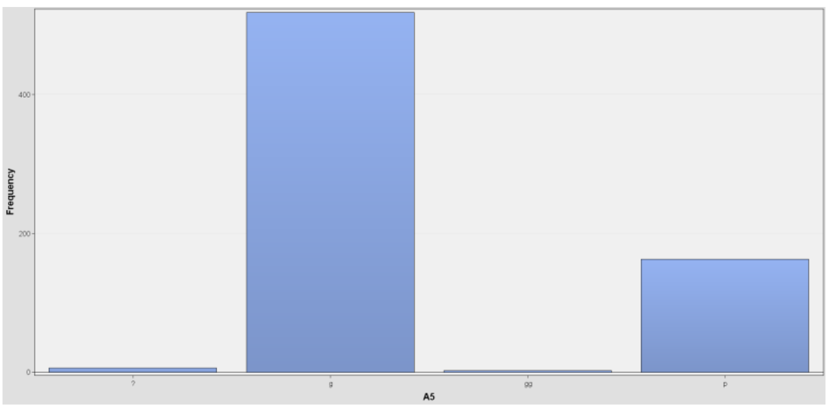


BAN210: Predictive Analytics

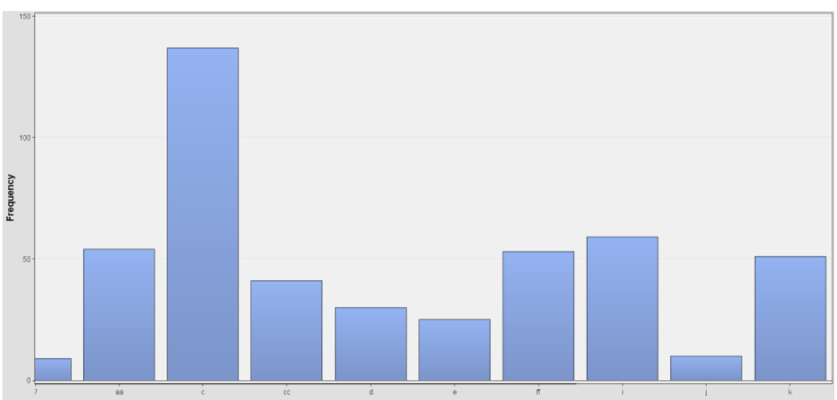
A4: This is a discrete variable with the presence of missing values



A5: This is a discrete variable with the presence of missing values

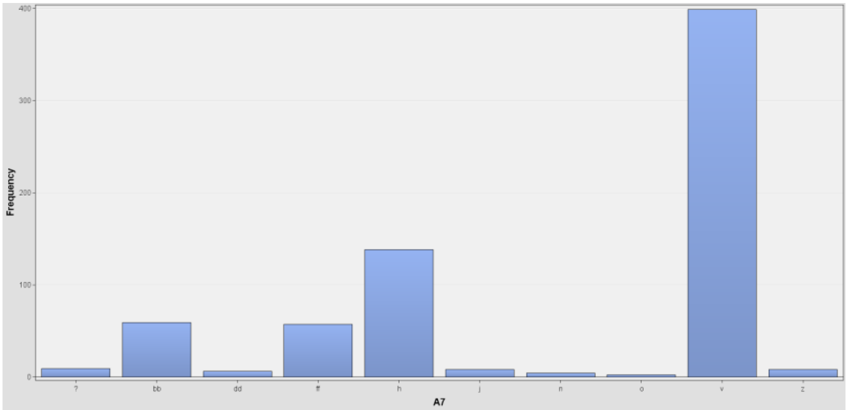


A6: This is a discrete variable with the presence of missing values

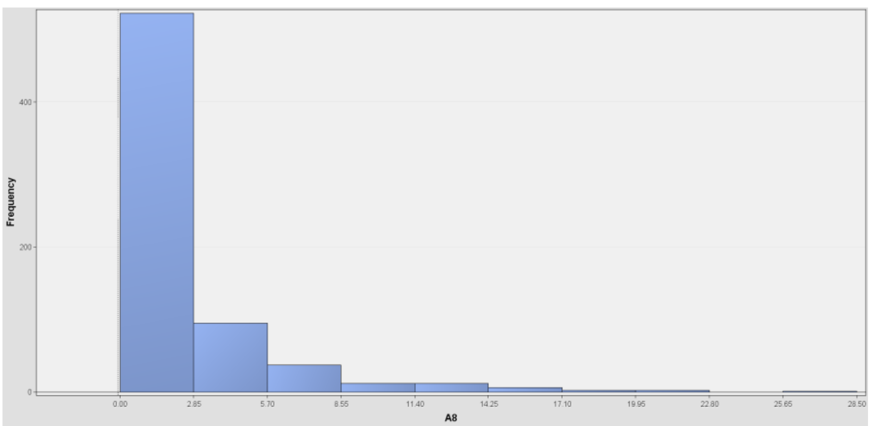


BAN210: Predictive Analytics

A7: This is a discrete variable with the presence of missing values



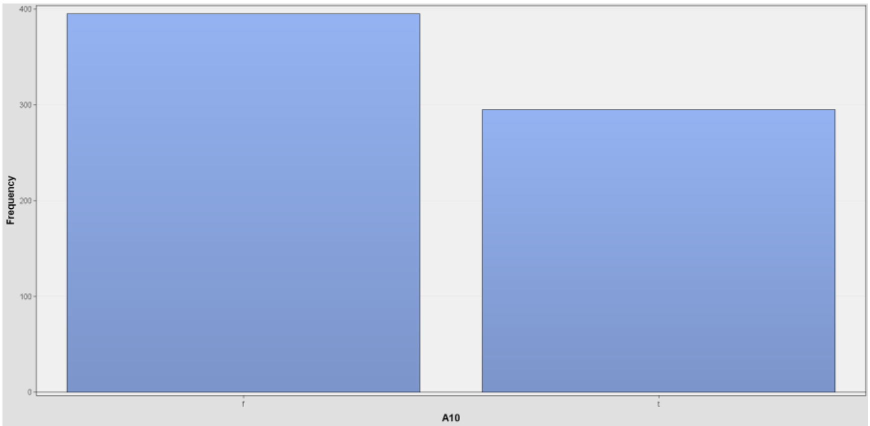
A8: This is a continuous variable.



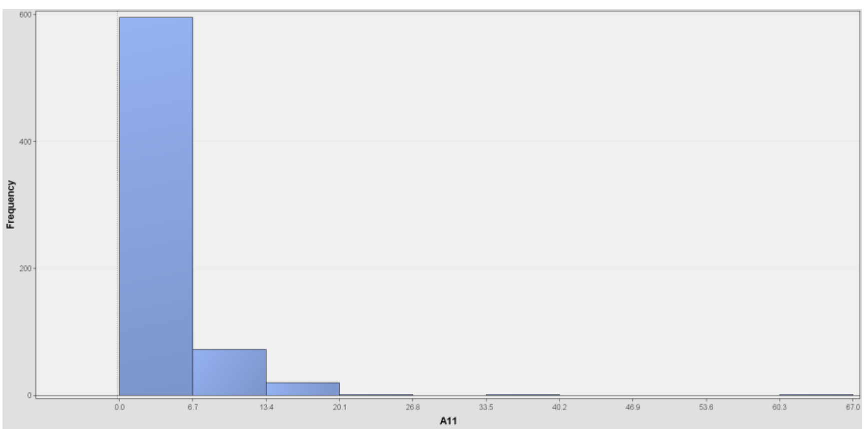
A9: This is a discrete variable.

BAN210: Predictive Analytics

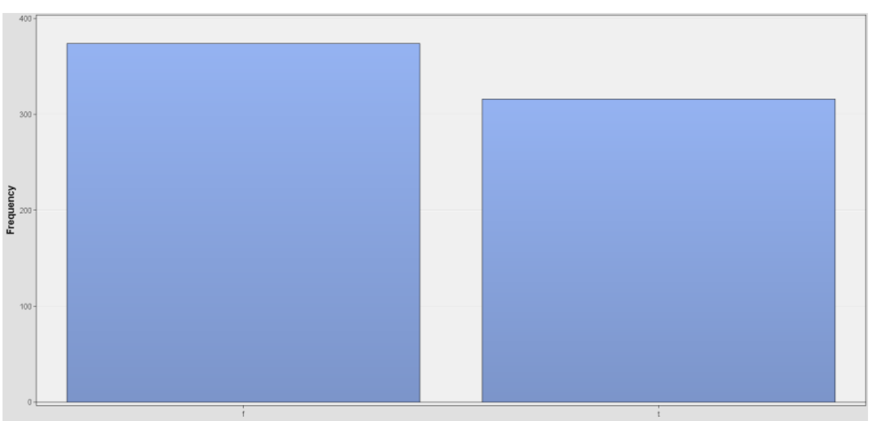
A10: This is a discrete variable.



A11: This is a continuous variable.

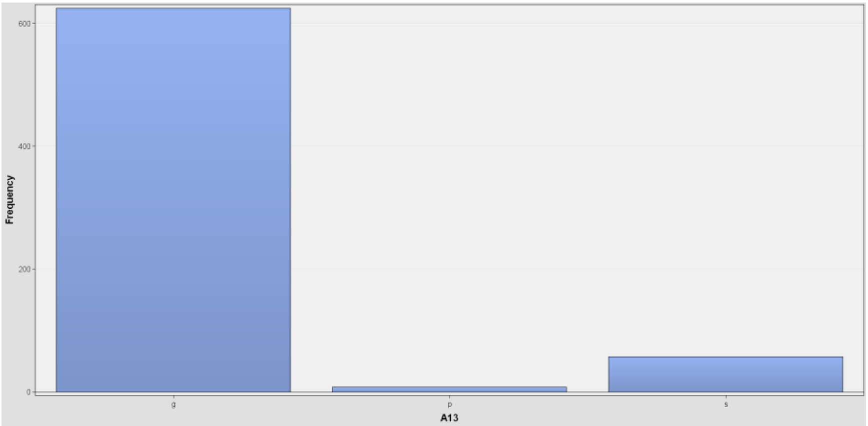


A12: This is a discrete variable.

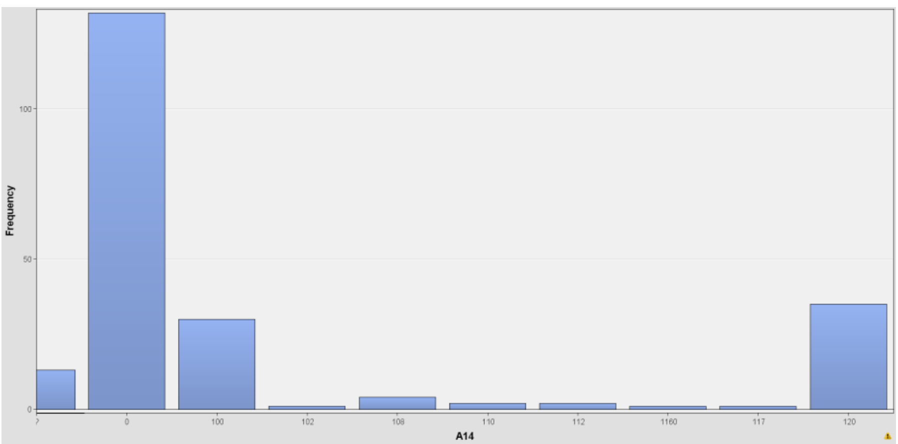


BAN210: Predictive Analytics

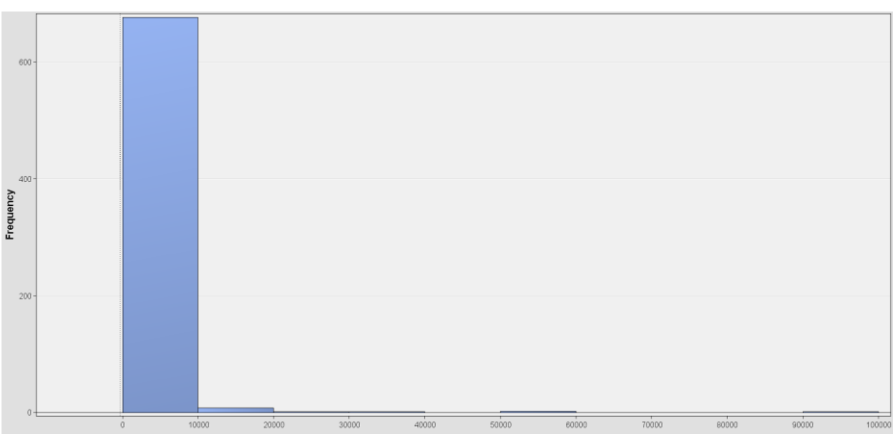
A13: This is a discrete variable.



A14: This is a discrete variable with the presence of missing values

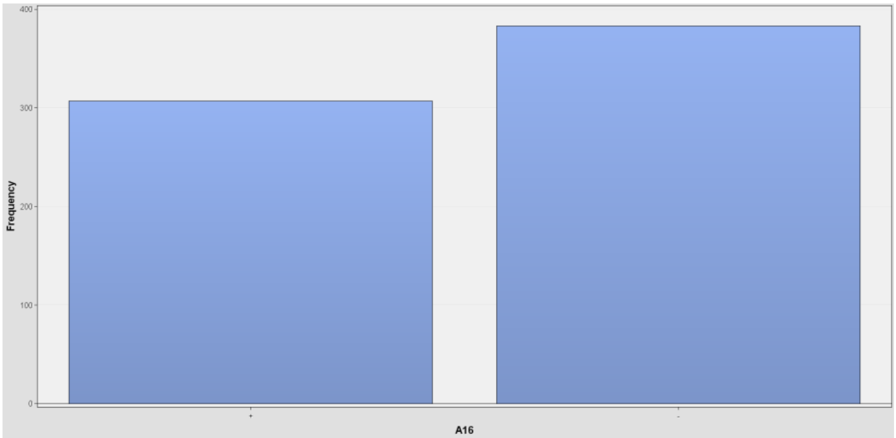


A15: This is a continuous variable.



BAN210: Predictive Analytics

A16: This is the target binary class variable.



Stat Explore:

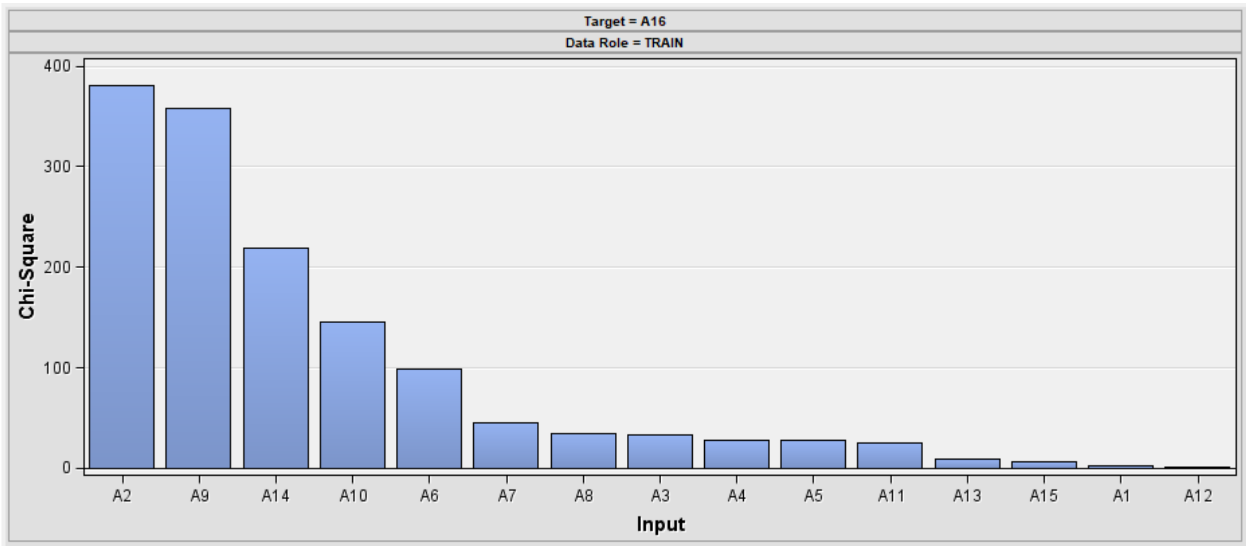
The StatExplore node generates summarization statistics.

Note: The Internal Variables property in the Properties Panel has been set to yes. When calculating the Chi-squared statistics for interval variables, Enterprise Miner distributes the internal variables into five bins, then determines the Chi-squared values for the binned variables when you run the node.

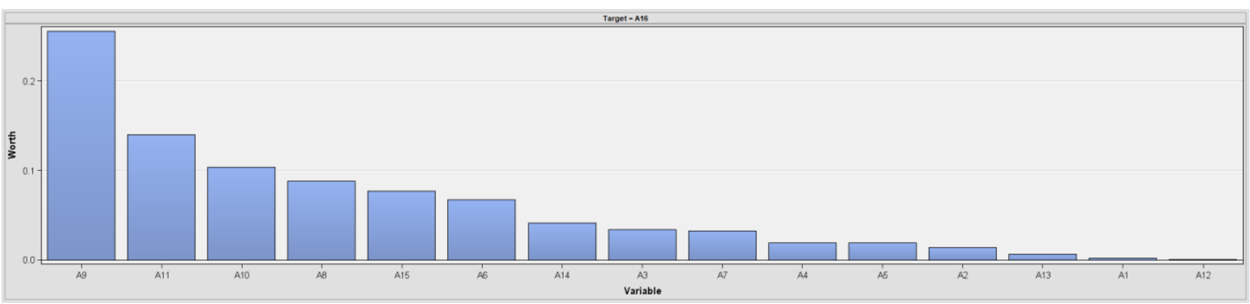
Property	Value
General	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	...
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	Yes
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No

BAN210: Predictive Analytics

The plot shows top 20 variables by their Chi-square statistics.



The variable worth plots order the input variables by their worth in predicting the target variable.



A9 variable has the highest worth in predicting the target variable.

Data Partition:

This node is used to divide the data into train, validation, and test set.

Training data is reserved for preliminary model fitting

Validation data is reserved to empirically test the model without overfitting the model.

Test data is reserved for an optional final assessment of the model.

Property	Value
General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	15.0
Test	15.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	4/14/22 7:22 PM
Run ID	52aec388-d5f8-e841-8637-d086204fdfe1
Last Error	
Last Status	Complete
Last Run Time	4/15/22 7:26 PM
Run Duration	0 Hr. 0 Min. 3.02 Sec.
Grid Host	
User-Added Node	No

Note:

Training set = 70%

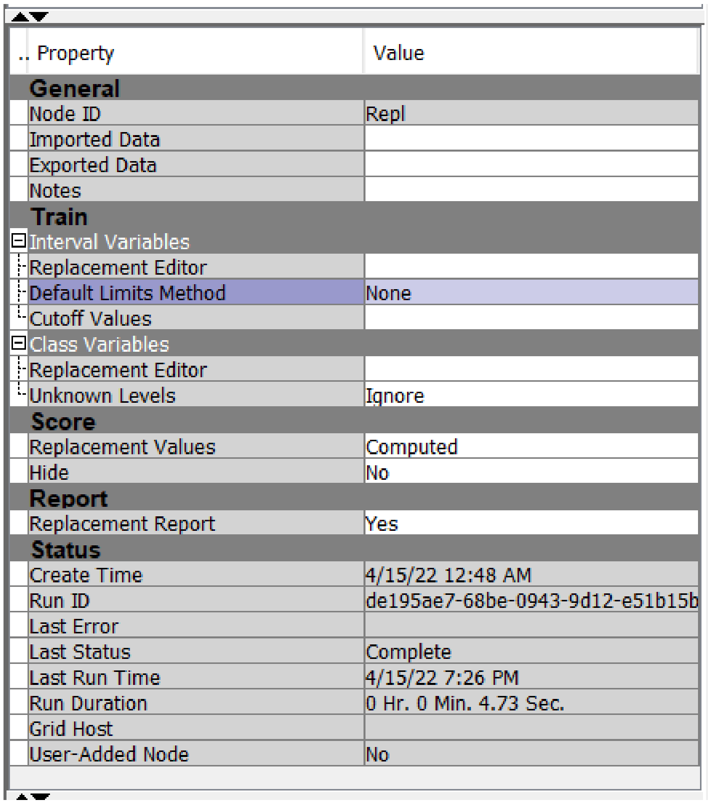
Validation set = 15%

Test set = 15%

Replacement:

The Replacement node belongs to the Modify category of the SAS SEMMA (Sample, Explore, Modify, Model, and Assess) data mining process.

BAN210: Predictive Analytics



Property	Value
General	
Node ID	Repl
Imported Data	
Exported Data	
Notes	
Train	
Interval Variables	
Replacement Editor	
Default Limits Method	None
Cutoff Values	
Class Variables	
Replacement Editor	
Unknown Levels	Ignore
Score	
Replacement Values	Computed
Hide	No
Report	
Replacement Report	Yes
Status	
Create Time	4/15/22 12:48 AM
Run ID	de195ae7-68be-0943-9d12-e51b15b
Last Error	
Last Status	Complete
Last Run Time	4/15/22 7:26 PM
Run Duration	0 Hr. 0 Min. 4.73 Sec.
Grid Host	
User-Added Node	No

Select the Default Limits Method to none for the interval variables. None indicates that no interval variable values should be replaced. The default setting of Standard Deviations from the Mean would enforce a range of values for each interval variable, which is not suitable for this example.

We will specify the replacement values for the class variables in the dataset.

BAN210: Predictive Analytics

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
A1	b		182	C	b	.
A1	a		90	C	a	.
A1	?		3	C	?	.
A1	_UNKNOWN_	_DEFAULT_	.	C		.
A10	f		158	C	f	.
A10	t		117	C	t	.
A10	_UNKNOWN_	_DEFAULT_	.	C		.
A12	f		144	C	f	.
A12	t		131	C	t	.
A12	_UNKNOWN_	_DEFAULT_	.	C		.
A13	g		250	C	g	.
A13	s		23	C	s	.
A13	p		2	C	p	.
A13	_UNKNOWN_	_DEFAULT_	.	C		.
A14	0		54	C	0	.
A14	120		18	C	120	.
A14	100		16	C	100	.
A14	200		16	C	200	.
A14	160		14	C	160	.
A14	80		13	C	80	.
A14	280		9	C	280	.
A14	240		8	C	240	.
A14	140		6	C	140	.
A14	180		6	C	180	.
A14	300		6	C	300	.
A14	260		5	C	260	.
A14	340		5	C	340	.
A14	60		4	C	60	.
A14	220		3	C	220	.
A14	320		3	C	320	.
A14	360		3	C	360	.
A14	?		3	C	?	.
A14	110		2	C	110	.
A14	112		2	C	112	.
A14	129		2	C	129	.
A14	144		2	C	144	.
A14	164		2	C	164	.
A14	176		2	C	176	.
A14	216		2	C	216	.
A14	290		2	C	290	.
A14	352		2	C	352	.

OK Cancel

BAN210: Predictive Analytics

For the variables with ? level represent observations that have missing values. In the Replacement value column, we have entered _MISSING_ for those values as shown below:

Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value	Frequency Count	Type	Character Unformatted Value	Numeric Value
A1	b		182	C	b	.
A1	a		90	C	a	.
A1	?	_MISSING_	3	C	?	.

The following class variables have been replaced and their count has been mentioned in the output below:

```
Replacement Values for Class Variables

                                       Character
                Formatted              Unformatted     Numeric     Replacement
Variable          Value       Type        Value          Value        Value        Label

  A1                ?           C           ?              .          _blank_
  A14               ?           C           ?              .          _blank_
  A2                ?           C           ?              .          _blank_
  A4                ?           C           ?              .          _blank_
  A5                ?           C           ?              .          _blank_
  A6                ?           C           ?              .          _blank_
  A7                ?           C           ?              .          _blank_
```

```
Replacement Counts

Obs     Variable     Role      Label     Train     Validation     Test

 1        A1         INPUT                 7           5            0
 2        A14        INPUT                 9           1            3
 3        A2         INPUT                 7           4            1
 4        A4         INPUT                 4           0            2
 5        A5         INPUT                 4           0            2
 6        A6         INPUT                 5           2            2
 7        A7         INPUT                 5           2            2
```

The replaced new variables are prefixed with REP.

Control Point:

A control point makes it easier to distribute connections amongst several interconnected nodes in a process flow step. It has the potential to lessen the number of connections formed.

Models:

1. **Logistic Regression:**
    - **Impute –**

BAN210: Predictive Analytics

Impute values to use as replacement for missing values in the input data. We replace missing data because Regression and Neural Network models ignore observations that contain missing values. This reduces the size of the training dataset, which can weaken the predictive power of those types of models. However missing values are not problematic for decision trees.

In the Class Variables section of the Impute node Train properties, select Tree Surrogate from the list as the default input method

Train	
Variables	
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Tree Surrogate
Default Target Method	None
Normalize Values	Yes

In the Interval Variables section of the Impute node Train properties, click Default Input Method, and select Median from the list.

The values of missing interval variables are replaced by median of the non-missing values. The median statistic is less sensitive to extreme values than mean or midrange statistics.

Interval Variables	
Default Input Method	Median
Default Target Method	None

Output:

```
Imputation Summary
Number Of Observations


                                                                                         Number of
Variable    Impute      Imputed     Impute                Measurement                      Missing
  Name      Method      Variable     Value      Role         Level           Label        for TRAIN

REP_A1      TREESURR    IMP_REP_A1     .       INPUT        NOMINAL    Replacement: A1           7
REP_A14     TREESURR    IMP_REP_A14    .       INPUT        NOMINAL    Replacement: A14          9
REP_A2      TREESURR    IMP_REP_A2     .       INPUT        NOMINAL    Replacement: A2           7
REP_A4      TREESURR    IMP_REP_A4     .       INPUT        NOMINAL    Replacement: A4           4
REP_A5      TREESURR    IMP_REP_A5     .       INPUT        NOMINAL    Replacement: A5           4
REP_A6      TREESURR    IMP_REP_A6     .       INPUT        NOMINAL    Replacement: A6           5
REP_A7      TREESURR    IMP_REP_A7     .       INPUT        NOMINAL    Replacement: A7           5
```

```
Variable Distribution Training Data

         Number of
           Missing    Number of     Percent of
Obs      for TRAIN    Variables      Variables

 1               9        1           14.2857
 2               7        2           28.5714
 3               5        2           28.5714
 4               4        2           28.5714
```

The Imputed variables in SAS results are identified by the prefix IMP_

- **Transform –**

  Transforming the data can improve model response. Transforming the data tends to stabilize variance, remove nonlinearity, improve additivity, and counter non-normality.

  Note: All the original variables are rejected, and imputed variables are carried forward for the analysis.

Variables - Trans

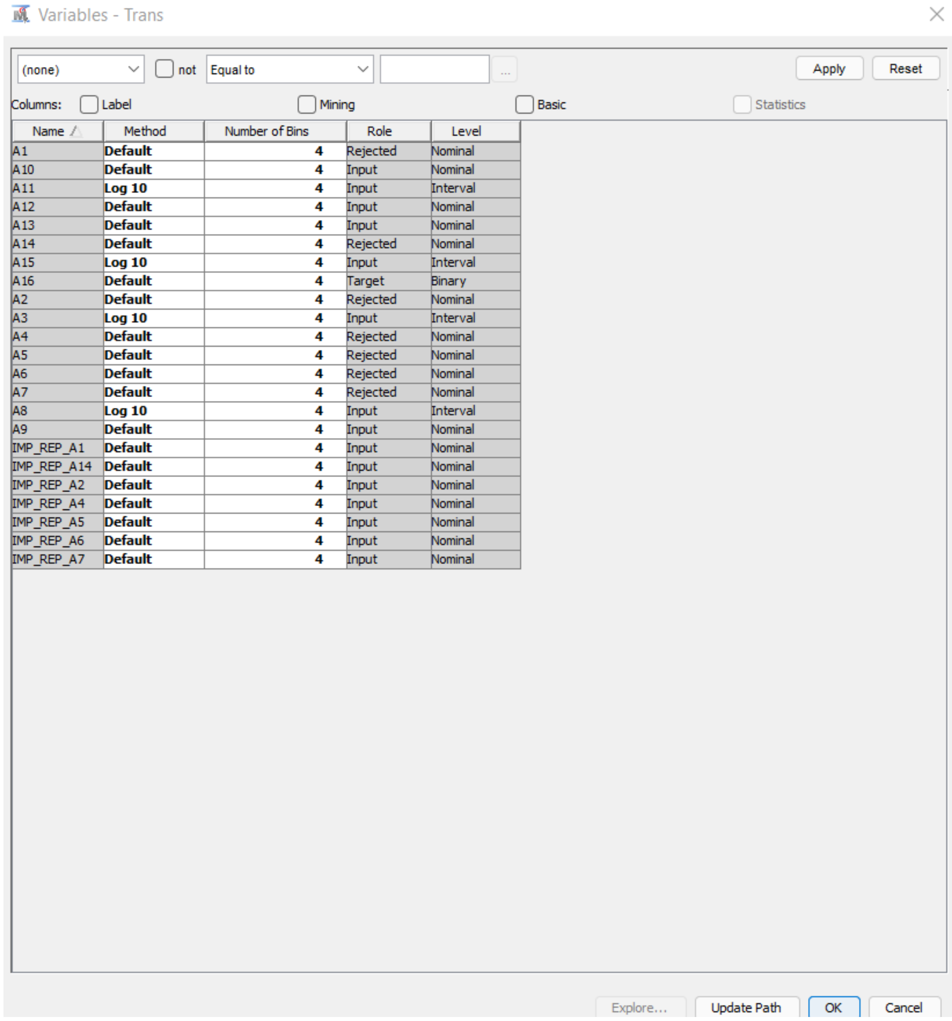
(none) | not | Equal to | Apply | Reset

Columns: Label | Mining | Basic | Statistics

Name	Method	Number of Bins	Role	Level
A1	Default	4	Rejected	Nominal
A10	Default	4	Input	Nominal
A11	Default	4	Input	Interval
A12	Default	4	Input	Nominal
A13	Default	4	Input	Nominal
A14	Default	4	Rejected	Nominal
A15	Default	4	Input	Interval
A16	Default	4	Target	Binary
A2	Default	4	Rejected	Nominal
A3	Default	4	Input	Interval
A4	Default	4	Rejected	Nominal
A5	Default	4	Rejected	Nominal
A6	Default	4	Rejected	Nominal
A7	Default	4	Rejected	Nominal
A8	Default	4	Input	Interval
A9	Default	4	Input	Nominal
IMP_REP_A1	Default	4	Input	Nominal
IMP_REP_A14	Default	4	Input	Nominal
IMP_REP_A2	Default	4	Input	Nominal
IMP_REP_A4	Default	4	Input	Nominal
IMP_REP_A5	Default	4	Input	Nominal
IMP_REP_A6	Default	4	Input	Nominal
IMP_REP_A7	Default	4	Input	Nominal

BAN210: Predictive Analytics

Based on the histogram of the variables, following transformations have been applied.



Name	Method	Number of Bins	Role	Level
A1	Default	4	Rejected	Nominal
A10	Default	4	Input	Nominal
A11	Log 10	4	Input	Interval
A12	Default	4	Input	Nominal
A13	Default	4	Input	Nominal
A14	Default	4	Rejected	Nominal
A15	Log 10	4	Input	Interval
A16	Default	4	Target	Binary
A2	Default	4	Rejected	Nominal
A3	Log 10	4	Input	Interval
A4	Default	4	Rejected	Nominal
A5	Default	4	Rejected	Nominal
A6	Default	4	Rejected	Nominal
A7	Default	4	Rejected	Nominal
A8	Log 10	4	Input	Interval
A9	Default	4	Input	Nominal
IMP_REP_A1	Default	4	Input	Nominal
IMP_REP_A14	Default	4	Input	Nominal
IMP_REP_A2	Default	4	Input	Nominal
IMP_REP_A4	Default	4	Input	Nominal
IMP_REP_A5	Default	4	Input	Nominal
IMP_REP_A6	Default	4	Input	Nominal
IMP_REP_A7	Default	4	Input	Nominal

Log10 transformations have been applied to selected interval variables as they have skewed distribution.

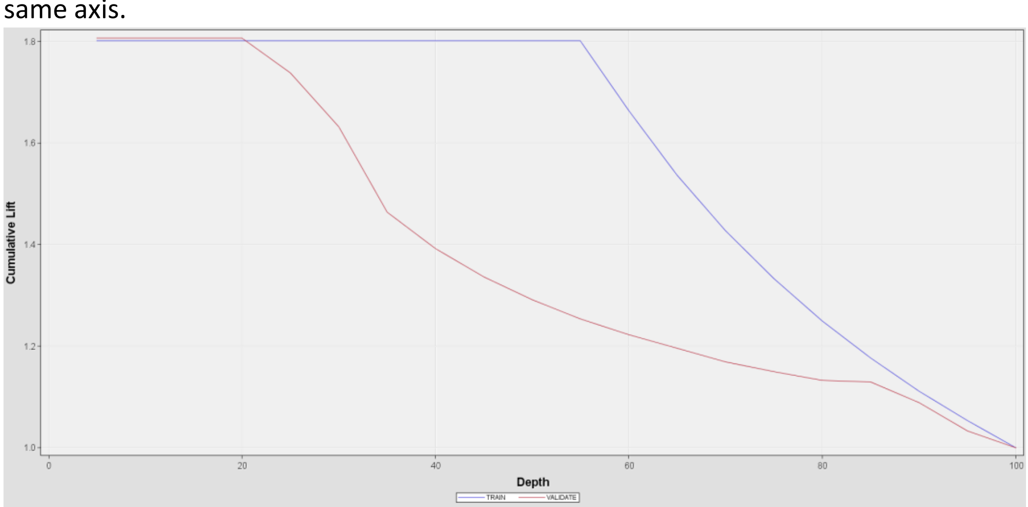
- **Regression –**
  Regression node is used to fit both linear and logistic regression. This is the problem of Logistic regression. Input coding property used to specify the coding method that you want to use with class variables. Generalized linear model (GLM) specifying how to interpret coefficients for categorical variables.

BAN210: Predictive Analytics

Property Selection

Property	Value
Variables	...
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	GLM
Model Selection	
Selection Model	None
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...

The score rankings overlay plot displays both train and validate statistics on the same axis.



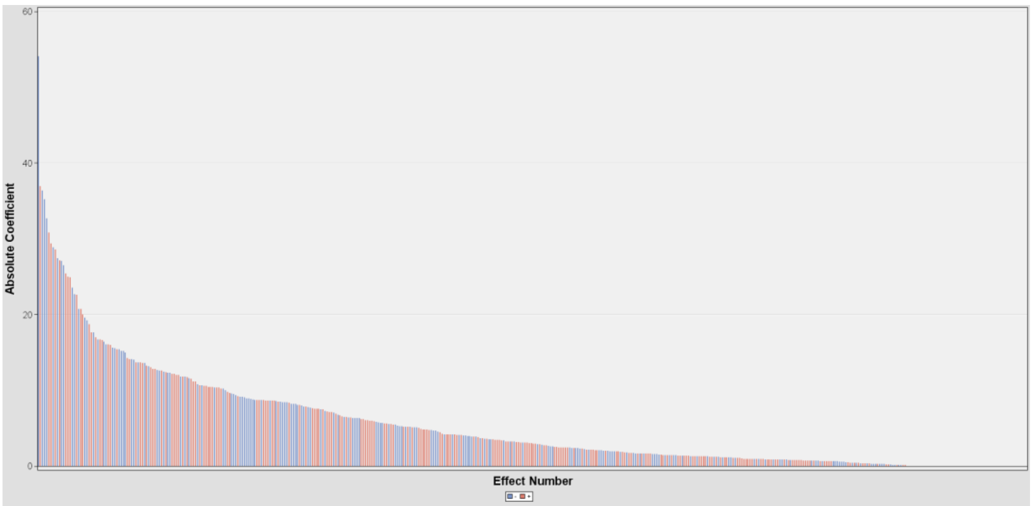
BAN210: Predictive Analytics

Table of the fit statistics from the model

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
A16		AIC	Akaike's Information Criterion	818.6583	.	.
A16		ASE	Average Squared Error	1.972E-5	0.212237	0.181349
A16		AVERR	Average Error Function	.0006843	1.026811	0.937902
A16		DFE	Degrees of Freedom for Error	72	.	.
A16		DFM	Model Degrees of Freedom	409	.	.
A16		DFT	Total Degrees of Freedom	481	.	.
A16		DIV	Divisor for ASE	962	206	212
A16		ERR	Error Function	0.658256	211.523	198.8352
A16		FPE	Final Prediction Error	.0002437	.	.
A16		MAX	Maximum Absolute Error	0.049497	1	1
A16		MSE	Mean Square Error	.0001317	0.212237	0.181349
A16		NOBS	Sum of Frequencies	481	103	106
A16		NW	Number of Estimate Weights	409	.	.
A16		RASE	Root Average Sum of Squares	0.00444	0.460691	0.425851
A16		RFPE	Root Final Prediction Error	0.015612	.	.
A16		RMSE	Root Mean Squared Error	0.011477	0.460691	0.425851
A16		SBC	Schwarz's Bayesian Criterion	2526.588	.	.
A16		SSE	Sum of Squared Errors	0.018968	43.72073	38.44595
A16		SUMW	Sum of Case Weights Times Freq	962	206	212
A16		MISC	Misclassification Rate	0	0.339806	0.301887

Effects plot displays a bar graph of the absolute values of the coefficients in the final model



Event Classification Table:

```
Event Classification Table

Data Role=TRAIN Target=A16 Target Label=' '

  False       True      False       True
Negative    Negative   Positive   Positive

     .         214         .         267


Data Role=VALIDATE Target=A16 Target Label=' '

  False       True      False       True
Negative    Negative   Positive   Positive

     4          15        31          53
```

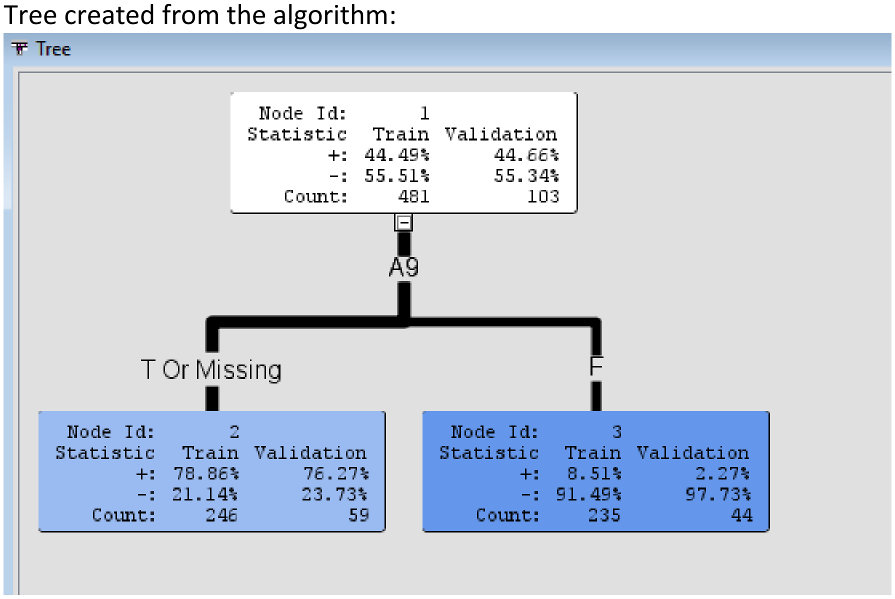
2. Decision Tree:

Properties selected:

BAN210: Predictive Analytics

Property	Value
General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
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	.
Split Search	

Tree created from the algorithm:



Fit Statistics:

Fit Statistics	Statistics Label	Train	Validation	Test
NOBS	Sum of Frequencies	481.000	103.000	106.000
MISC	Misclassification Rate	0.150	0.146	0.123
MAX	Maximum Absolute Error	0.915	0.915	0.915
SSE	Sum of Squared Errors	118.612	23.732	21.747
ASE	Average Squared Error	0.123	0.115	0.103
RASE	Root Average Squared Error	0.351	0.339	0.320
DIV	Divisor for ASE	962.000	206.000	212.000
DFT	Total Degrees of Freedom	481.000	.	.

Event Classification Table:

Event Classification Table

Data Role=TRAIN Target=A16 Target Label=' '

False Negative	True Negative	False Positive	True Positive
52	194	20	215

Data Role=VALIDATE Target=A16 Target Label=' '

False Negative	True Negative	False Positive	True Positive
14	45	1	43

3. **Random Forest:**

Note: Random Forest algorithm has been implemented using the HP Forest node

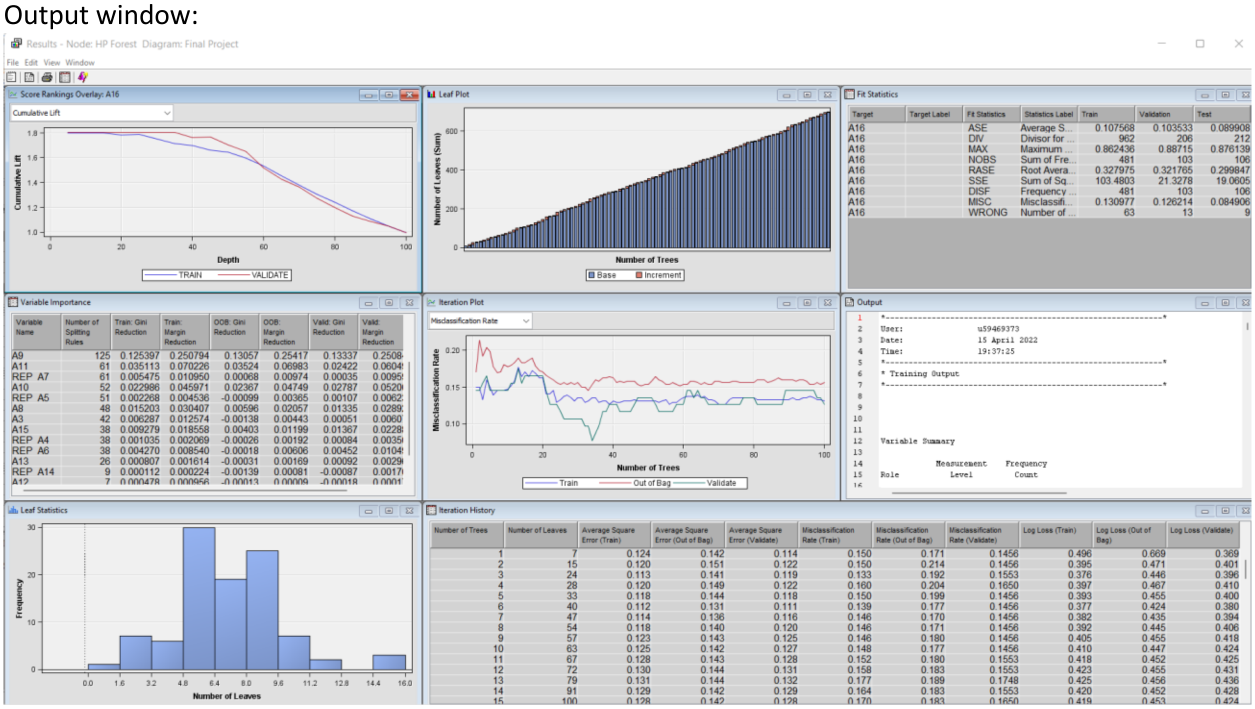
Properties selected:

BAN210: Predictive Analytics

Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	20
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split Se	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in Node	1.0E-5
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes

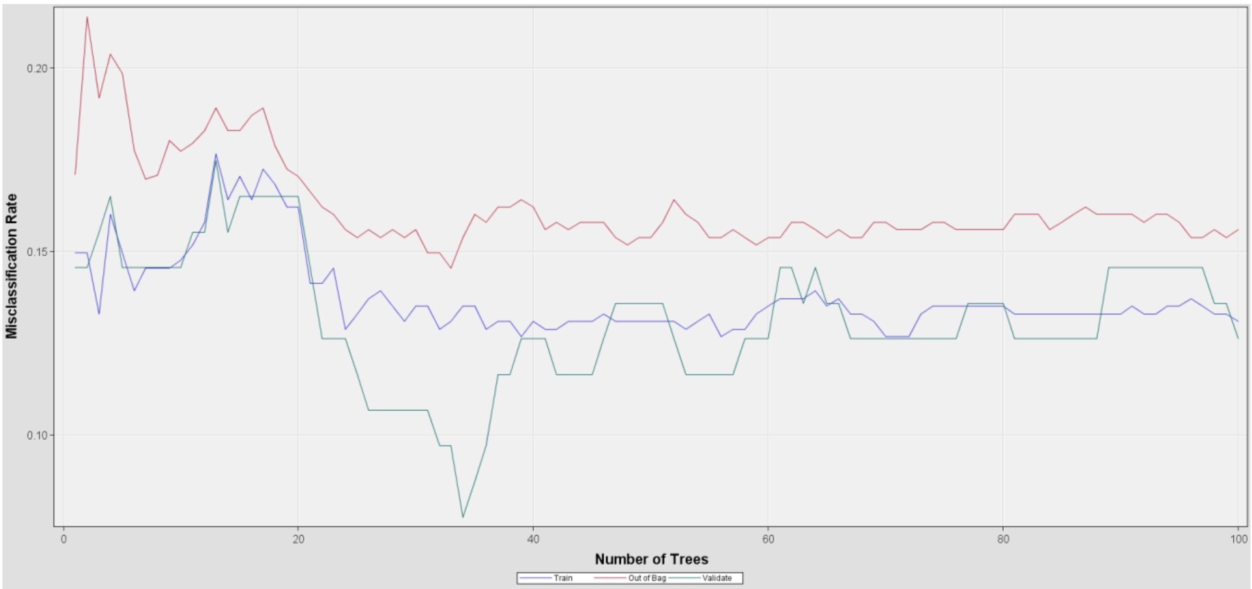
General

Output window:



BAN210: Predictive Analytics

The misclassification rate plots begin to flatten after more than 20 trees are added to the forest. That is, once the forest contains 20 trees, adding more trees does not have a significant effect on the misclassification rate of the model. Using this information to specify a reasonable value for the Maximum Number of Trees property.

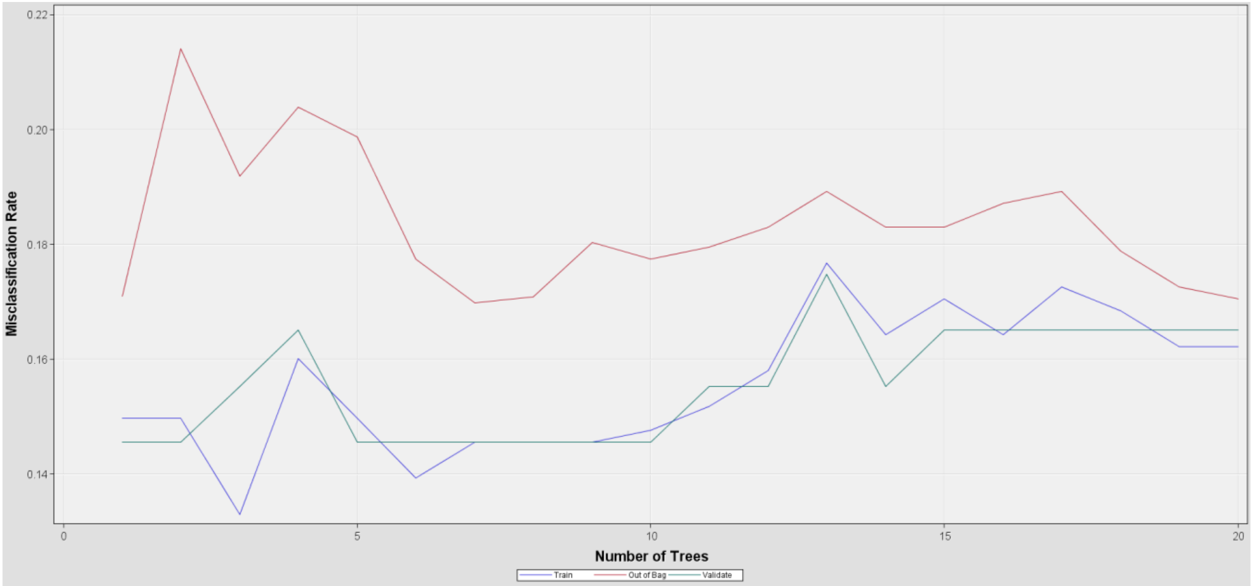


Update the maximum number of trees to 20

Property	Value
General	
Node ID	HPDMForest
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Tree Options	
Maximum Number of Trees	20
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sample	0.6
Number of Obs in Each Sample	.
Splitting Rule Options	
Maximum Depth	50
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Consider in Split Se	.
Significance Level	0.05
Max Categories in Split Search	30
Minimum Category Size	5
Exhaustive	5000
Node Options	
Method for Leaf Size	Default
Smallest Percentage of Obs in Node	1.0E-5
Smallest Number of Obs in Node	1
Split Size	.
Use as Modeling Node	Yes

BAN210: Predictive Analytics

Misclassification rate on selected Trees

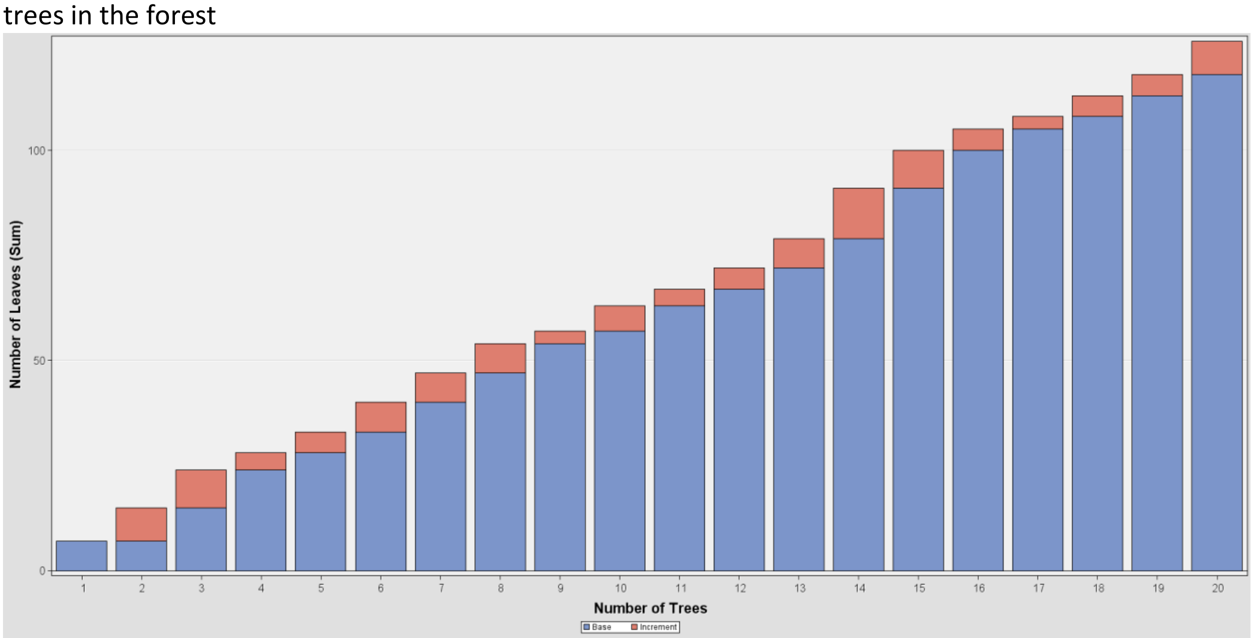


Fit Statistics table displays statistics for the training, validation, and test data sets

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
A16		ASE	Average Squared Error	0.12444	0.121507	0.101762
A16		DIV	Divisor for ASE	962	206	212
A16		MAX	Maximum Absolute Error	0.84125	0.870512	0.854221
A16		NOBS	Sum of Frequencies	481	103	106
A16		RASE	Root Average Squared Error	0.35276	0.348579	0.319002
A16		SSE	Sum of Squared Errors	119.7109	25.03052	21.57358
A16		DISF	Frequency of Classified Cases	481	103	106
A16		MISC	Misclassification Rate	0.162162	0.165049	0.113208
A16		WRONG	Number of Wrong Classifications	78	17	12

Leaf plot displays the total number of leaves in the forest plotted against the number of trees in the forest



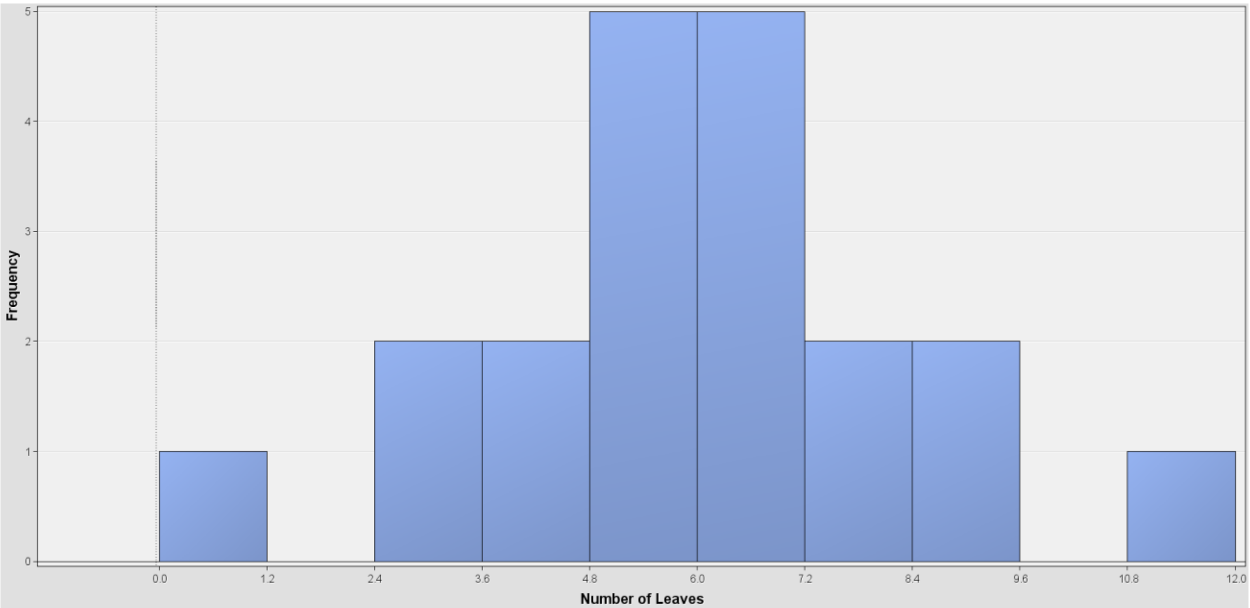
BAN210: Predictive Analytics

Variable importance is a table with information about each variable's worth to the model

Variable Importance

Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
A10	17	0.032682	0.065365	0.03641	0.070395	0.03782	0.072544	
A9	15	0.068566	0.137131	0.06700	0.131821	0.06717	0.130904	
A11	14	0.040804	0.081607	0.03836	0.076048	0.02240	0.061609	
A8	10	0.016978	0.033955	0.00917	0.024961	0.01728	0.034281	
REP_A5	10	0.003197	0.006393	-0.00114	0.003627	0.00099	0.006998	Replacement: A5
REP_A7	8	0.004539	0.009077	0.00226	0.010123	0.00063	0.008133	Replacement: A7
A3	6	0.005295	0.010590	-0.00052	0.003239	0.00224	0.005821	
REP_A4	6	0.001447	0.002894	-0.00036	0.002289	0.00194	0.005239	Replacement: A4
REP_A6	6	0.008264	0.016528	0.00379	0.011489	0.01241	0.019737	Replacement: A6
A13	5	0.000997	0.001994	-0.00051	0.000713	0.00075	0.001928	
A15	5	0.008236	0.016473	0.00677	0.014984	0.01667	0.025942	
A12	3	0.001522	0.003044	0.00048	0.001624	-0.00021	0.000717	
REP_A14	1	0.000000	0.000000	-0.00040	0.001931	0.00002	0.002146	Replacement: A14
REP_A1	0	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	Replacement: A1
REP_A2	0	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000	Replacement: A2

Leaf Statistics is a histogram that displays the distribution for the number of leaves in each tree.

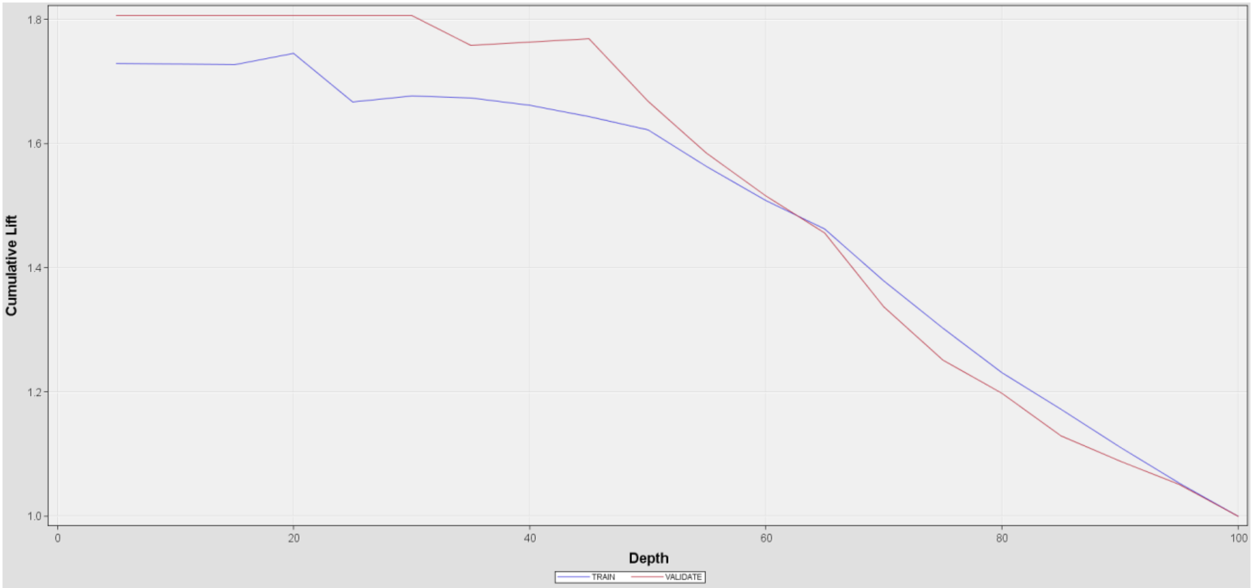


Statistics history on each iteration image

Iteration History

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)	Misclassification Rate (Train)	Misclassification Rate (Out of Bag)	Misclassification Rate (Validate)	Log Loss (Train)	Log Loss (Out of Bag)	Log Loss (Validate)
1	7	0.124	0.142	0.114	0.150	0.171	0.146	0.496	0.669	0.369
2	15	0.120	0.151	0.122	0.150	0.214	0.146	0.395	0.471	0.401
3	24	0.113	0.141	0.119	0.133	0.192	0.155	0.376	0.446	0.396
4	28	0.120	0.149	0.122	0.160	0.204	0.165	0.397	0.467	0.410
5	33	0.118	0.144	0.118	0.150	0.199	0.146	0.393	0.455	0.400
6	40	0.112	0.131	0.111	0.139	0.177	0.146	0.377	0.424	0.380
7	47	0.114	0.136	0.116	0.146	0.170	0.146	0.382	0.435	0.394
8	54	0.118	0.140	0.120	0.146	0.171	0.146	0.392	0.445	0.406
9	57	0.123	0.143	0.125	0.146	0.180	0.146	0.405	0.455	0.418
10	63	0.125	0.142	0.127	0.148	0.177	0.146	0.410	0.447	0.424
11	67	0.128	0.143	0.128	0.152	0.180	0.155	0.418	0.452	0.425
12	72	0.130	0.144	0.131	0.158	0.183	0.155	0.423	0.455	0.431
13	79	0.131	0.144	0.132	0.177	0.189	0.175	0.425	0.456	0.436
14	91	0.129	0.142	0.129	0.164	0.183	0.155	0.420	0.452	0.428
15	100	0.128	0.142	0.128	0.170	0.183	0.165	0.419	0.453	0.424
16	105	0.129	0.142	0.129	0.164	0.187	0.165	0.423	0.453	0.427
17	108	0.132	0.144	0.132	0.173	0.189	0.165	0.430	0.459	0.434
18	113	0.129	0.142	0.129	0.168	0.179	0.165	0.423	0.454	0.426
19	118	0.127	0.139	0.125	0.162	0.173	0.165	0.417	0.446	0.417
20	126	0.124	0.135	0.122	0.162	0.170	0.165	0.412	0.438	0.409

The score rankings overlay plot displays both train and validate statistics on the same axis



Event Classification Table:

```
Event Classification Table

Data Role=TRAIN Target=A16 Target Label=' '

  False       True       False       True
Negative    Negative    Positive    Positive

   22          158         56          245


Data Role=VALIDATE Target=A16 Target Label=' '

  False       True       False       True
Negative    Negative    Positive    Positive

    7          36          10           50
```

BAN210: Predictive Analytics

Conclusion

The project has been successfully completed all the Predictive modelling steps on the Credit Card dataset. This includes

- File Import
- Stats Explore
- Data partition
- Replacement and Imputation (required for regression)
- Control Point
- Models: Decision tree, Random Forest, Logistic Regression

Let's find the classification parameters like Precision, Recall and F1 score calculated for all the three above models below (Validation set):

Precision is the ratio between the True Positives and all the Positives.

For our problem statement, that would be the measure of tweets that we correctly identify as positive out of all the examples.

The **recall** is the measure of our model correctly identifying True Positives. Thus, for all the examples that are actually positive, recall tells us how many we correctly identified as having as positive.

For our dataset we need to have correct identification of both positive and negative examples hence both precision and recall are equally important. In such cases, we use something called F1-score. F1-score is the Harmonic mean of the Precision and Recall:

Note: **Precision** = True Positives / (True Positives + False Positives)

= TP / (TP+FP)

Recall = True Positives / (True Positives + False Negatives)

= TP / (TP+FN)

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

BAN210: Predictive Analytics

Parameters	Regression	Decision Tree	Random Forest
Precision	0.630	0.977	0.833
Recall	0.929	0.754	0.877
F1 score	0.750	0.851	0.854
Misclassification rate	0.340	0.146	0.165
Average squared error	0.212	0.115	0.122

A classification technique with the highest accuracy and precision with the lowest misclassification rate and average squared error is the most intelligent classifier for prediction purposes. Please find the observations below:

- Decision tree has highest Precision and its average square error is also lowest amongst all the three models.
- Regression has high Recall and its Misclassification rate and Average squared error is also high compared to other model.
- Random Forest model has the highest model score, but its misclassification and average squared error is higher than Decision Tree model.

Based on the comparison of all the above parameters, I conclude that Decision Tree is the best model for the Credit Card dataset.

References:

https://documentation.sas.com

https://www.youtube.com/watch?v=IVHdZwf1nlw

https://www.youtube.com/watch?v=fJDuKQVVj50

https://www.youtube.com/watch?v=kmrpr8LmMz4

BAN210: Predictive Analytics

Declaration

I, **Jinalben Patel** declare that the attached assignment is my own work in accordance with the Seneca Academic Policy. I have not copied any part of this assignment, manually or electronically, from any other source including web sites, unless specified as references. I have not distributed my work to other students.