DMBA ASSIGNMENT

AY21/22 Apr Semester

DECLARATION

I declare that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism."

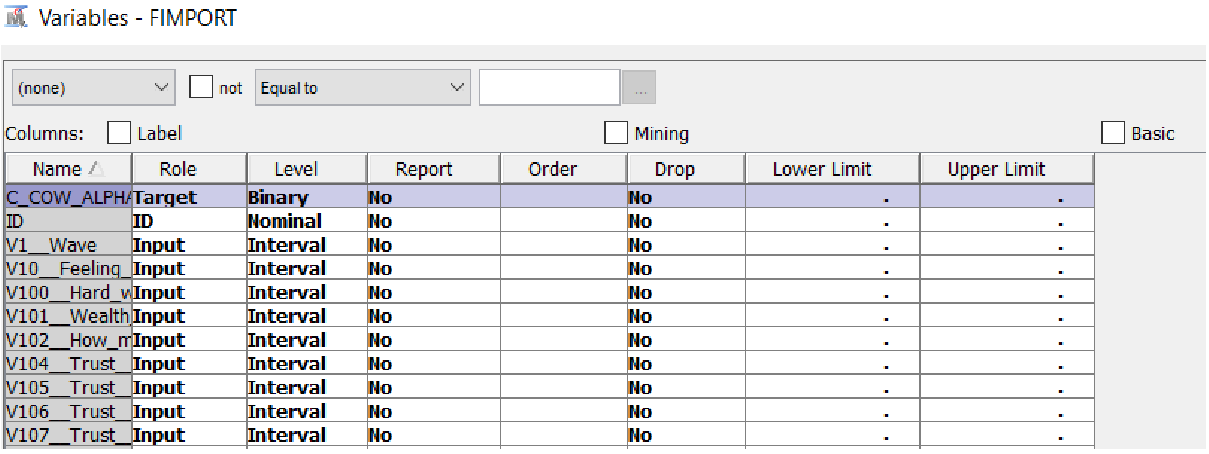
- ❏ I Agree (Please right-click on box, select tick)

My Information

Name (as in matriculation card)	Joyce Teng Min Li
Admin Number	1907675A
Practical Group (e.g. P01)	P02
Task selected (A or B)	Task C

Performance of Cluster and Association Analysis / Predictive Modelling Task (40 marks, 20%)

File Import



Variables - FIMPORT

(none) □ not Equal to

Columns: □ Label □ Mining □ Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
C_COW_ALPHA	Target	Binary	No		No	.	.
ID	ID	Nominal	No		No	.	.
V1__Wave	Input	Interval	No		No	.	.
V10__Feeling_	Input	Interval	No		No	.	.
V100__Hard_w	Input	Interval	No		No	.	.
V101__Wealth	Input	Interval	No		No	.	.
V102__How_m	Input	Interval	No		No	.	.
V104__Trust_	Input	Interval	No		No	.	.
V105__Trust_	Input	Interval	No		No	.	.
V106__Trust_	Input	Interval	No		No	.	.
V107__Trust_	Input	Interval	No		No	.	.

Use the File import node to import the excel file into SAS EM.

- Configure the “C_COW_ALPHA__Country_code_CoW_al” role as Target and the level as Binary
- Reject “V3__Original_respondent_number” as the ID. And configure “ID” role as ID
- Leave everything as default

StatExplore

Use the StatExplore node to find the variable worth.
Based on the variable worth, I selected Top 15 variables as it gives a nice separation from the 6th-20th variable.

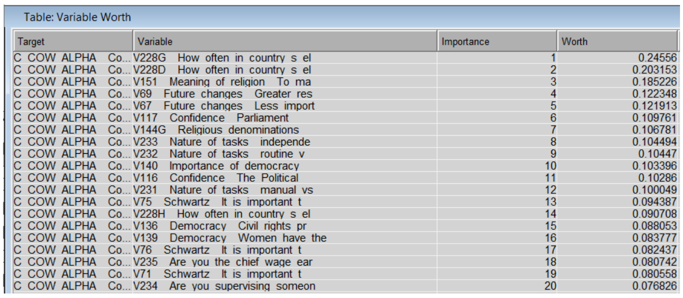
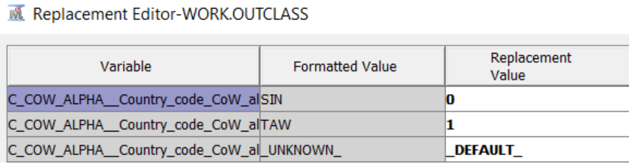


Table: Variable Worth

Target	Variable	Importance	Worth
C COW ALPHA Co...	V228G How often in country s el	1	0.24556
C COW ALPHA Co...	V228D How often in country s el	2	0.203153
C COW ALPHA Co...	V151 Meaning of religion To ma	3	0.185226
C COW ALPHA Co...	V69 Future changes Greater res	4	0.122348
C COW ALPHA Co...	V67 Future changes Less import	5	0.121913
C COW ALPHA Co...	V117 Confidence Parliament	6	0.109761
C COW ALPHA Co...	V144G Religious denominations	7	0.106781
C COW ALPHA Co...	V233 Nature of tasks independe	8	0.104494
C COW ALPHA Co...	V232 Nature of tasks routine v	9	0.10447
C COW ALPHA Co...	V140 Importance of democracy	10	0.103396
C COW ALPHA Co...	V116 Confidence The Political	11	0.10286
C COW ALPHA Co...	V231 Nature of tasks manual vs	12	0.100049
C COW ALPHA Co...	V75 Schwartz It is important t	13	0.094387
C COW ALPHA Co...	V228H How often in country s el	14	0.090708
C COW ALPHA Co...	V136 Democracy Civil rights pr	15	0.088053
C COW ALPHA Co...	V139 Democracy Women have the	16	0.083777
C COW ALPHA Co...	V76 Schwartz It is important t	17	0.082437
C COW ALPHA Co...	V235 Are you the chief wage ear	18	0.080742
C COW ALPHA Co...	V71 Schwartz It is important t	19	0.080558
C COW ALPHA Co...	V234 Are you supervising someon	20	0.076826

Replacement - Replacement node



Replacement Editor-WORK.OUTCLASS

Variable	Formatted Value	Replacement Value
C_COW_ALPHA__Country_code_CoW_al	SIN	0
C_COW_ALPHA__Country_code_CoW_al	TAW	1
C_COW_ALPHA__Country_code_CoW_al	_UNKNOWN_	_DEFAULT_

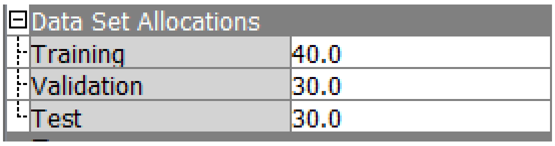
The replacement node is used to ensure that the target is in binary form.

Use the replace node and configure

- replace SIN to 0

- replace TAW to 1

Partitioning Data - Data Partition Node



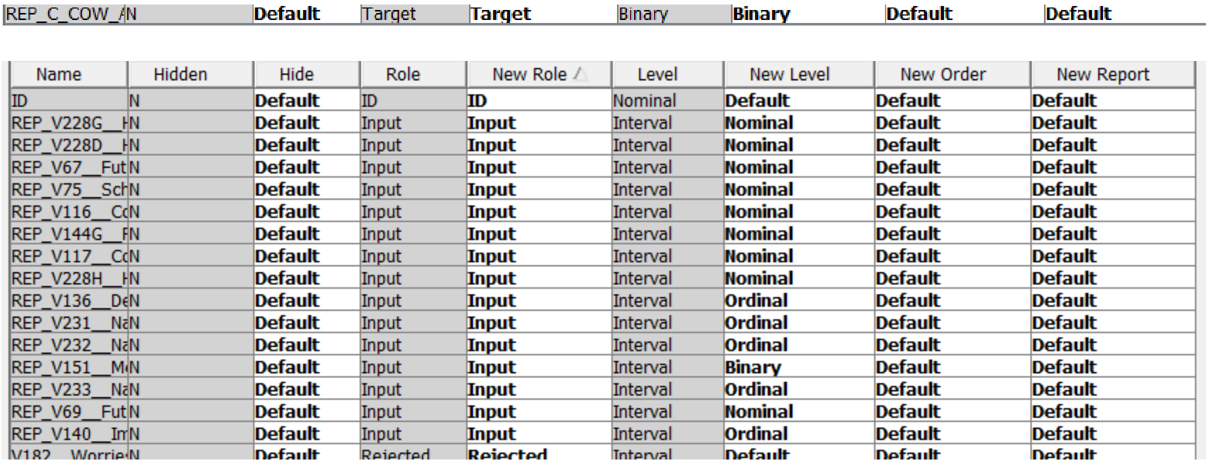
Data Set Allocations	
Training	40.0
Validation	30.0
Test	30.0

Use the Data partition node and allocate 40% to Training, 30% to Validation and 30% to Test.
Training is set at a higher percentage because it helps to make the predictive models more robust and stable. Hence training is more important since it ensures that the model has a higher accuracy.

Metadata - Metadata Node

Since I have used the replacement node to replace the target column, the metadata node is used to reject columns that are not needed. I have only kept the top 15 variables by looking at the variable worth in the statnode. This metadata node will be used throughout the models to ensure fairness.

- Set the replaced “REP_C_COW_ALPHA_Country_code_CO” new role as Target.
- Reject all the other columns except for the top 15 variables
- Assigned the datatype of the top 15 variables under the new Level.



REP_C_COW_A	N	Default	Target	Target	Binary	Binary	Default	Default

Name	Hidden	Hide	Role	New Role	Level	New Level	New Order	New Report
ID	N	Default	ID	ID	Nominal	Default	Default	Default
REP_V228G__H	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V228D__H	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V67__Fut	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V75__Sch	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V116__Co	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V144G__F	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V117__Co	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V228H__H	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V136__De	N	Default	Input	Input	Interval	Ordinal	Default	Default
REP_V231__Na	N	Default	Input	Input	Interval	Ordinal	Default	Default
REP_V232__Na	N	Default	Input	Input	Interval	Ordinal	Default	Default
REP_V151__M	N	Default	Input	Input	Interval	Binary	Default	Default
REP_V233__Na	N	Default	Input	Input	Interval	Ordinal	Default	Default
REP_V69__Fut	N	Default	Input	Input	Interval	Nominal	Default	Default
REP_V140__Im	N	Default	Input	Input	Interval	Ordinal	Default	Default
V182__Worrie	N	Default	Rejected	Rejected	Interval	Default	Default	Default

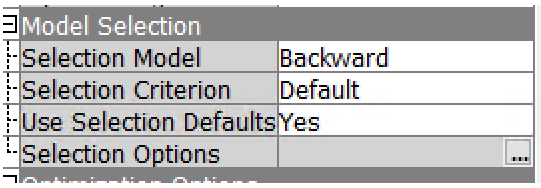
Models Used

Logistic Regression(default)

I used Logistic regression because the task is a classification model with 2 outputs (SIN OR TAW) .

This Regression node does not need to be configured and will run with its original configurations and properties. Running it as the default settings gives us an idea what the result will be like without any tuning.

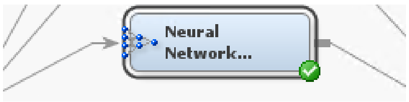
Logistic Regression (Backwards)



Model Selection	
Selection Model	Backward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	...

This Regression node, I have tuned the parameters and selected “Backwards” as the selection Model. The purpose of tuning the parameter is for improvement purposes.

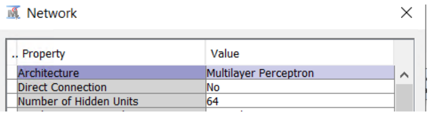
Neural Network (Default)



Neural network is able to do binary classification hence I chose to use this model.
This Neural Network node does not need to be configured and will run with its original configurations and properties. Running this node at its default settings will give us an idea what the result will be like without any tuning.

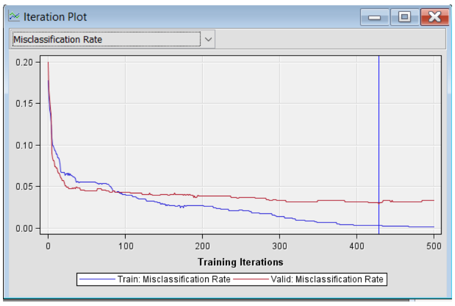
Neural Network - ANN w Settings

For this Neural Network node, I have tuned the parameters and selected the Network and set the Number of Units to the max value 64.The reason why I set it as the max is because the more hidden layers, the better the accuracy. As for the maximum of Iteration I initially kept the default value of 50 and tested. As the results were not ideal, I went to set the maximum iteration as 500. Based on the Iteration Plot, the training iteration did not go up to 500 as I have set. As the model has achieved 0 misclassification rate midway through the process. Using the line as a guideline I have decided to leave 500 as the maximum iteration.



Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	64

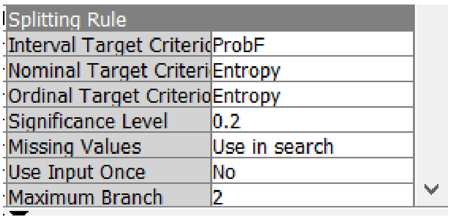


Decision Tree (Default)

Since the Decision tree can be used to predict a classification model, I have chosen to use this model to perform Task C. This Decision tree node does not need to be configured and will run with its original configurations and properties. Running this node at its default settings will give us an idea what the result will be like without any tuning.

Decision Tree (Automatic) (Entropy)

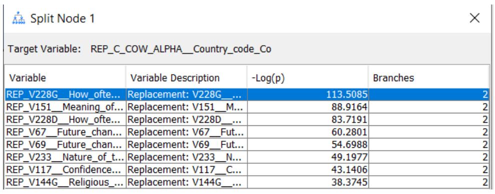
For this Decision Tree (Entropy) node I have tuned the Nominal Target Criteria to Entropy. It is a measure of randomness and it can help to control the way the decision tree is split.



Splitting Rule	
Interval Target Criteric	ProbF
Nominal Target Criteri	Entropy
Ordinal Target Criterio	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2

Decision Tree (Interactive)

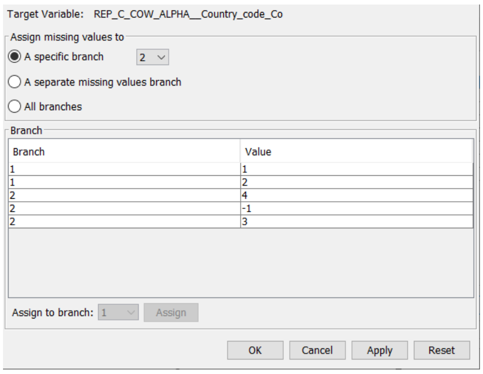
For this node, growing a tree interactively provides us with finer control. Using an interactive tree it allows me to split the branches manually by looking at the logs' worth. By selecting the highest log worth, I have split them into my first branch. By using the edit rule, I am able to assign which values go to which branch. Since for V228G it is nicely split I left it as the default.



Split Node 1

Target Variable: REP_C_COW_ALPHA__Country_code_Co

Variable	Variable Description	-Log(p)	Branches
REP_V228G__How_ofte...	Replacement: V228G__...	113.5085	2
REP_V151__Meaning_of...	Replacement: V151__M...	88.9164	2
REP_V228D__How_ofte...	Replacement: V228D__...	83.7191	2
REP_V67__Future_chan...	Replacement: V67__Fut...	60.2801	2
REP_V69__Future_chan...	Replacement: V69__Fut...	54.6988	2
REP_V233__Nature_of_t...	Replacement: V233__N...	49.1977	2
REP_V117__Confidence...	Replacement: V117__C...	43.1406	2
REP_V144G__Religious_...	Replacement: V144G__...	38.3745	2



Target Variable: REP_C_COW_ALPHA__Country_code_Co

Assign missing values to

● A specific branch 2

○ A separate missing values branch

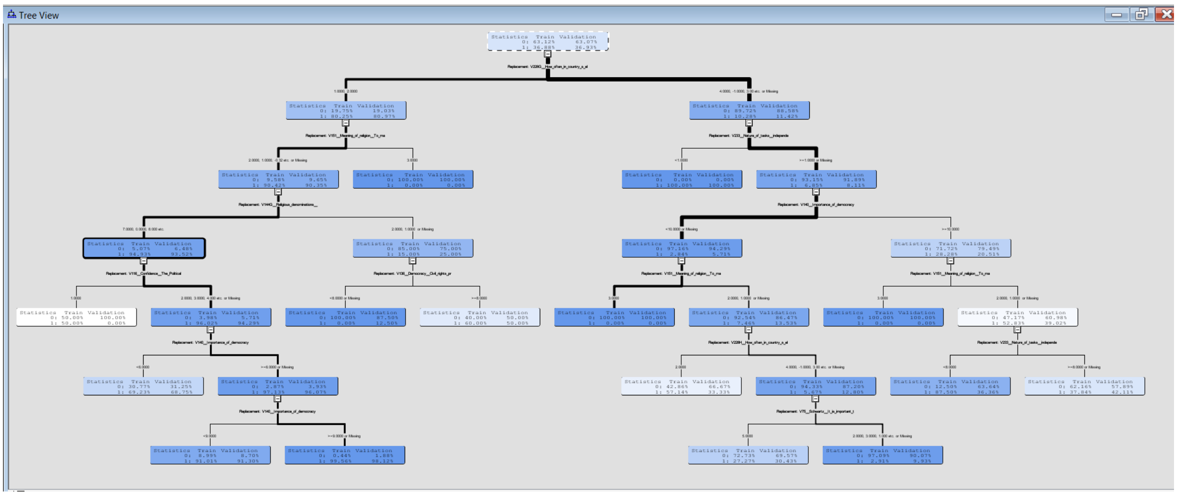
○ All branches

Branch

Branch	Value
1	1
1	2
2	4
2	-1
2	3

Assign to branch: 1

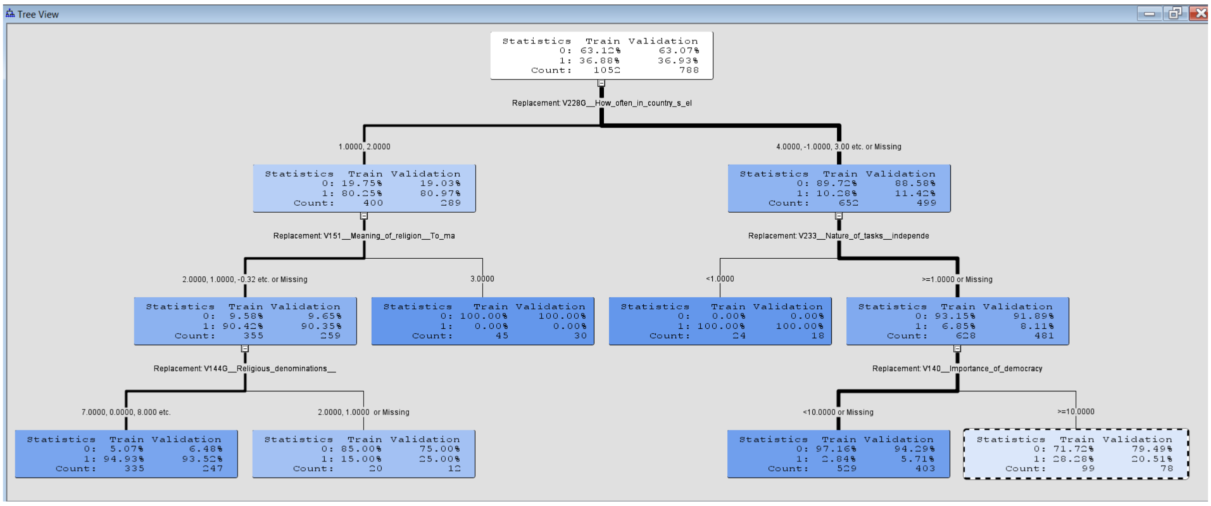
Afterwards, to grow the tree semi-automatically I selected the root node and selected Action and Train. This will grow the tree automatically, while preserving the changes that I have made.



Tree View
Statistics Train Validation
0: 63.12% 63.07%
1: 36.88% 36.93%
Replacement: V228G__How_often_in_country_s_el
1.0000, 2.0000
Statistics Train Validation
0: 19.75% 19.03%
1: 80.25% 80.97%
Replacement: V151__Meaning_of_religion__To_me
2.0000, 1.0000, -0.32 etc. or Missing
3.0000
Replacement: V144G__Religious_denominations__
Replacement: V116__Confidence__The_Political
Replacement: V136__Democracy__Civil_rights_pr
Replacement: V140__Importance_of_democracy

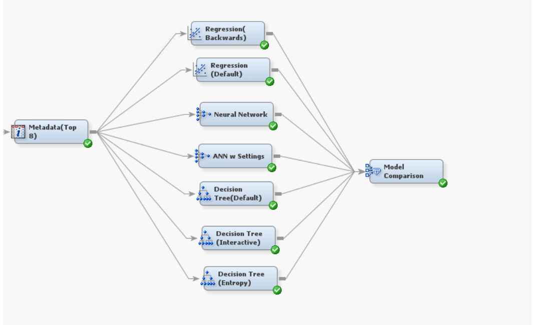
(Before Pruning)

Lastly I decided to prune from a 5-level tree to a 3-level tree for better understanding.
A 3-level decision tree is a good level as it is small but still able to provide us with a good amount of information.


Tree View
Statistics Train Validation
0: 63.12% 63.07%
1: 36.88% 36.93%
Count: 1052 788
Replacement: V228G__How_often_in_country_s_el
1.0000, 2.0000
Statistics Train Validation
0: 19.75% 19.03%
1: 80.25% 80.97%
Count: 400 289
4.0000, -1.0000, 3.00 etc. or Missing
Statistics Train Validation
0: 89.72% 88.58%
1: 10.28% 11.42%
Count: 652 499
Replacement: V151__Meaning_of_religion__To_ma
Replacement: V233__Nature_of_tasks__independe
2.0000, 1.0000, -0.32 etc. or Missing
Statistics Train Validation
0: 9.58% 9.65%
1: 90.42% 90.35%
Count: 355 259
3.0000
Statistics Train Validation
0: 100.00% 100.00%
1: 0.00% 0.00%
Count: 45 30
<1.0000
Statistics Train Validation
0: 0.00% 0.00%
1: 100.00% 100.00%
Count: 24 18
>=1.0000 or Missing
Statistics Train Validation
0: 93.15% 91.89%
1: 6.85% 8.11%
Count: 628 481
Replacement: V144G__Religious_denominations__
Replacement: V140__Importance_of_democracy
7.0000, 0.0000, 8.000 etc.
Statistics Train Validation
0: 5.07% 6.48%
1: 94.93% 93.52%
Count: 335 247
2.0000, 1.0000 or Missing
Statistics Train Validation
0: 85.00% 75.00%
1: 15.00% 25.00%
Count: 20 12
<10.0000 or Missing
Statistics Train Validation
0: 97.16% 94.29%
1: 2.84% 5.71%
Count: 529 403
>=10.0000
Statistics Train Validation
0: 71.72% 79.49%
1: 28.28% 20.51%
Count: 99 78

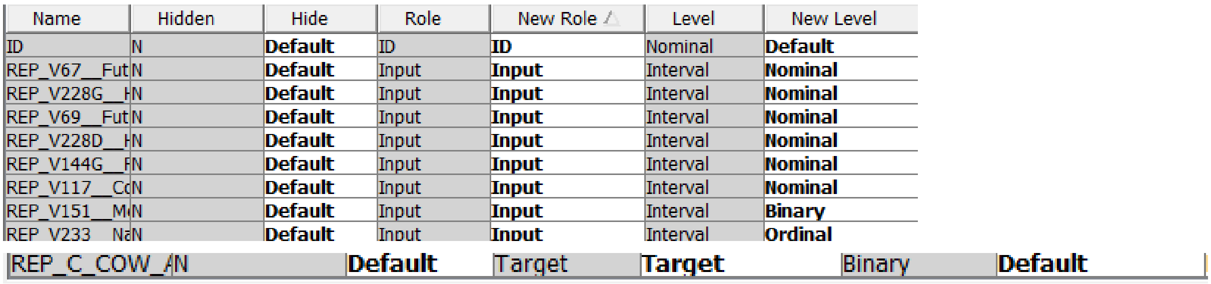
(After Pruning)

Comparison of Number of Variables used



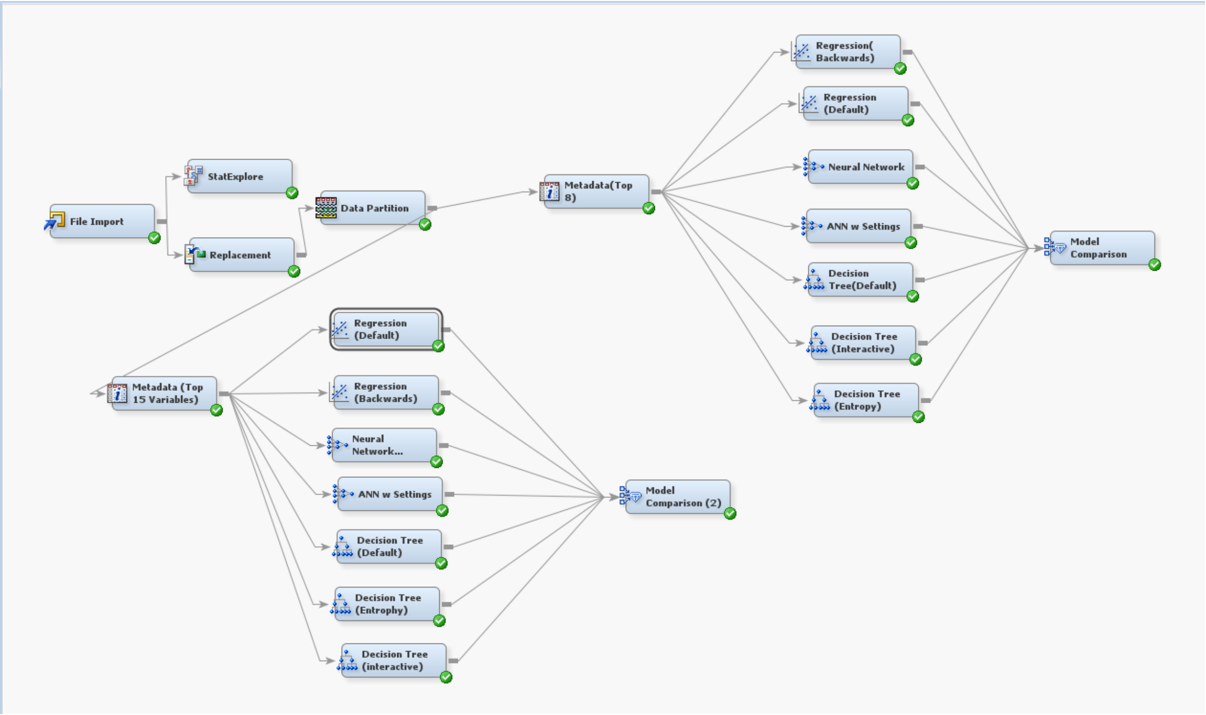
I also did another comparison where I compared the number of variables used.Usually the lesser the number of variables used, the higher the misclassification rate.
Therefore I created and used another metadata node and only input the top 8 variables. To do this comparison I have kept all the settings the same as the Top 15 configuration to ensure fairness. By comparing each the top 15 variable used champion model and top 8 champion model we can then prove if the lesser the number of variables used, the higher the misclassification rate.

After doing a comparison of both Top 15 and Top 8 champion model results, I'll be using the top 15 variables used to validate the misclassification rate.



Name	Hidden	Hide	Role	New Role	Level	New Level
ID	N	Default	ID	ID	Nominal	Default
REP_V67__Fut	N	Default	Input	Input	Interval	Nominal
REP_V228G__H	N	Default	Input	Input	Interval	Nominal
REP_V69__Fut	N	Default	Input	Input	Interval	Nominal
REP_V228D__H	N	Default	Input	Input	Interval	Nominal
REP_V144G__F	N	Default	Input	Input	Interval	Nominal
REP_V117__C	N	Default	Input	Input	Interval	Nominal
REP_V151__M	N	Default	Input	Input	Interval	Binary
REP_V233__N	N	Default	Input	Input	Interval	Ordinal
REP_C_COW_A	N	Default	Target	Target	Binary	Default

Final WorkFlow

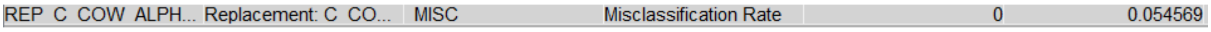


Interpretation of the Results (40 marks, 20%)

<u>Logistic Regression (Default)</u>

Fit statistic

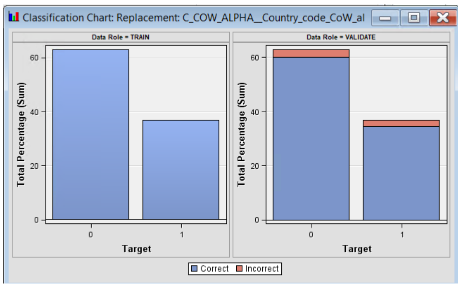
Since we are training the model to predict a binary classification (0 or 1) instead of predicting a continuous value using RMSE or MSE will not make sense. Hence I looked at the Misclassification Rate under validation. The chart below shows us the misclassification rate for the default logistic regression.



REP C COW ALPH...	Replacement: C CO...	MISC	Misclassification Rate	0	0.054569

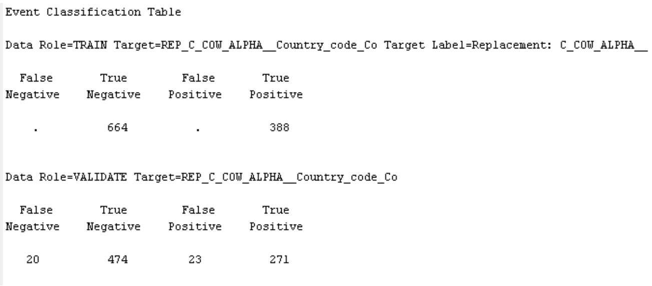
Classification chart

The chart below shows the classification chart for logistic regression. This classification chart gives us an overall view of how the misclassification rate would look like. Under the validate chart, there is a small amount that is red that represents the misclassification. Based on the two classification charts shown below, we can say that the default logistic regression model is quite accurate and is performing well.



Classification table

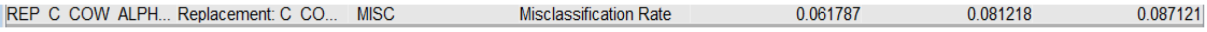
The classification table is also another way where we can access the model's performance. One way to calculate the misclassification rate is looking at the False Negative and False Positive numbers. I will be looking at the validation accuracy and misclassification .
Therefore the misclassification rate for the default logistic regression model is (20+23)/788 = 0.05456(5 s.f)
We can also use the classification table to calculate the accuracy of the model by looking at the True Negative and True positive. Therefore the accuracy rate for this model is (474+271)/788 = 0.94543(5 s.f) around 94.5% accuracy
Based on the result, I can conclude that the default regression model is quite accurate and is performing well.



```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False       True      False      True
Negative   Negative   Positive   Positive

    .         664         .         388

Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False       True      False      True
Negative   Negative   Positive   Positive

   20         474        23         271
```

Logistic Regression (Backwards)

Fit Statistic



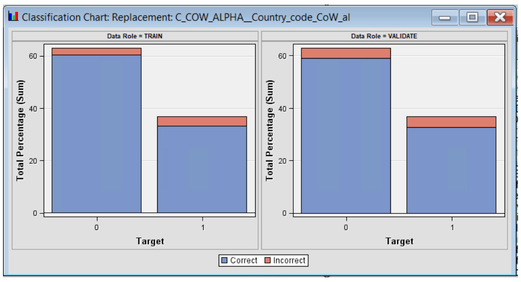
REP_C_COW_ALPH...	Replacement: C_CO...	MISC	Misclassification Rate	0.061787	0.081218	0.087121

Since we are training the model to predict a binary classification (0 or 1) instead of predicting a continuous value using RMSE or MSE will not make sense. Hence I looked at the Misclassification Rate under validation. The chart below shows us the misclassification rate for the default logistic regression.

Compared to the default logistic Regression Model, we can see that this model is not as accurate.

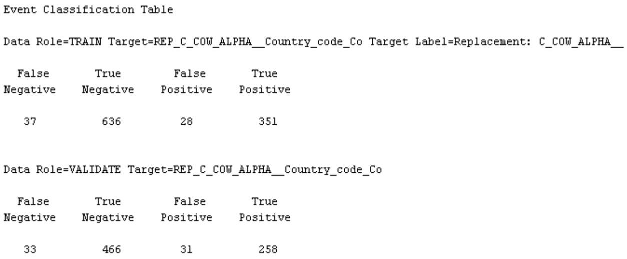
Classification Chart

This Classification chart gives us an overall view of the misclassification. The red portion represents the misclassification. At first glance, both the Train and Validate portion look similar to each other. I had to use the Classification Table to go in depth and look at the difference.



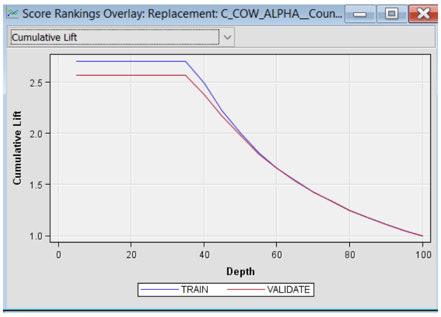
Classification Table

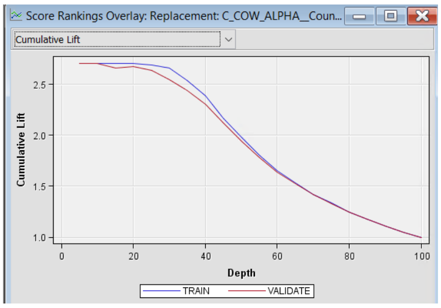
The classification table is also another way where we can access the model's performance. The classification is able to let us know the difference between the train and validate.I will be looking at the validation accuracy and misclassification
One way to calculate the misclassification rate is looking at the False Negative and False Positive numbers. Therefore the misclassification rate for the default logistic regression model is (33+31)/788 = 0.08121(5 s.f)
We can also use the classification table to calculate the accuracy of the model by looking at the True Negative and True positive. Therefore the accuracy rate for this model is (466+258)/788 = 0.91878(5 s.f) around 91.8% accuracy
Based on the result, This also shows that tuning the parameters did not improve the model's performance. I can conclude that the default regression model is not as accurate and is not performing well as compared to the default.



```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False      True      False      True
Negative   Negative   Positive   Positive

   37         636        28         351

Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False      True      False      True
Negative   Negative   Positive   Positive

   33         466        31         258
```

Cumulative Lift(Default VS Backwards)
Cumulative Lift charts show the predictive effectiveness of the model. Based on the chart below, I observed that the lines in the chart for Train and Validate are relatively close to one another. This is an indication that both the models are not overly fitted.

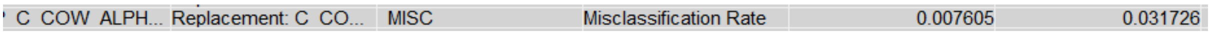




Neural Network (Default)

Fit Statistic

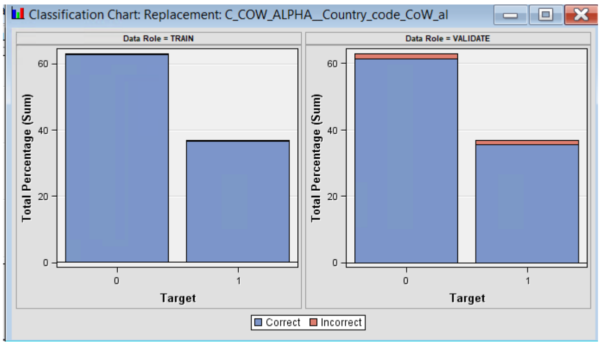
Since we are training the model to predict a binary classification (0 or 1) instead of predicting a continuous value using RMSE or MSE will not make sense. Hence I looked at the Misclassification Rate. The chart below shows us the misclassification rate for the default Neural Network.



C COW ALPH...	Replacement: C CO...	MISC	Misclassification Rate	0.007605	0.031726

Classification Chart

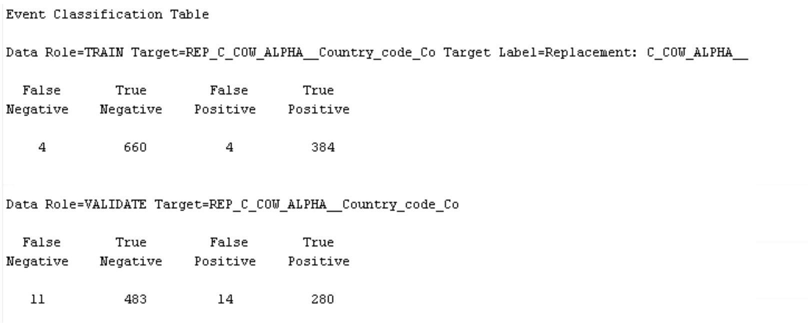
The chart below shows the classification chart for the Neural Network Default . This classification chart gives us an overall view of how the misclassification rate would look like. Under the validate chart, there is a small amount that is red that represents the misclassification. Based on the two classification charts shown below, we can say that the default model is quite accurate and is performing well.



Classification Table

The classification table is also another way where we can access the model's performance. To calculate the misclassification rate is looking at the False Negative and False Positive numbers over Total . I will be looking at the validation accuracy and misclassification Therefore the misclassification rate for the default logistic regression model is (11+14)/788 = 0.03172(5 s.f)
We can also use the classification table to calculate the accuracy of the model by looking at the True Negative and True positive over Total . Therefore the accuracy rate for this model is (483+280)/788 = 0.96827(5 s.f) around 96.8% accuracy.

Based on the result, I can conclude that the default Neural Network is more accurate and is performing well compared to the Logistic Regression (Default)

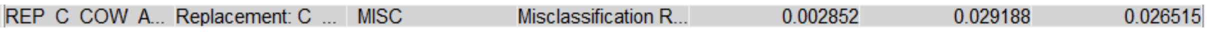


```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False       True       False       True
Negative    Negative    Positive    Positive

    4         660          4          384


Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False       True       False       True
Negative    Negative    Positive    Positive

   11         483         14          280
```

ANN w Settings

Fit Statistic

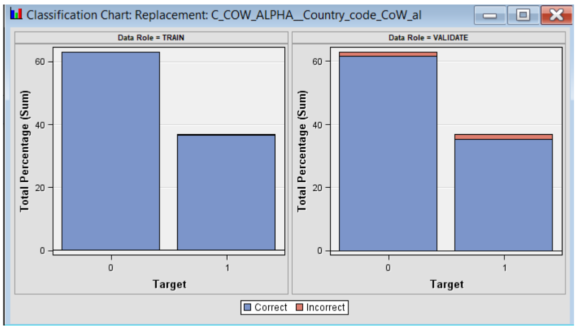
We can see that the misclassification rate is lower compared to the default neural network as this model was tuned. This shows that tuning the parameters has improved the model's performance since the misclassification rate is lower.



REP_C_COW_A...	Replacement: C ...	MISC	Misclassification R...	0.002852	0.029188	0.026515

Classification Chart

The chart below shows the classification chart for the Neural Network Default . This classification chart gives us an overall view of how the misclassification rate would look like. Under the validate chart, there is a small amount that is red that represents the misclassification. Based on the two classification charts shown below, we can say that the tuned model is quite accurate and is performing well.

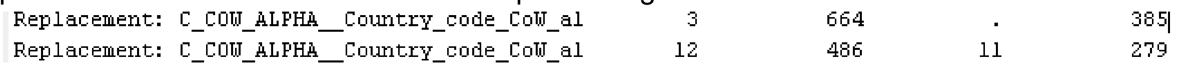


Classification Table

To evaluate the model's performance. I will be assessing the misclassification and accuracy rate. To calculate the misclassification rate is looking at the False Negative and False Positive numbers over Total . I will be looking at the validation accuracy and misclassification Therefore the misclassification rate for the default logistic regression model is (12+11)/788 = 0.02918(5 s.f)

We can also use the classification table to calculate the accuracy of the model by looking at the True Negative and True positive over Total . Therefore the accuracy rate for this model is (486+279)/788 = 0.97081(5 s.f) around 97.0% accuracy was achieved.

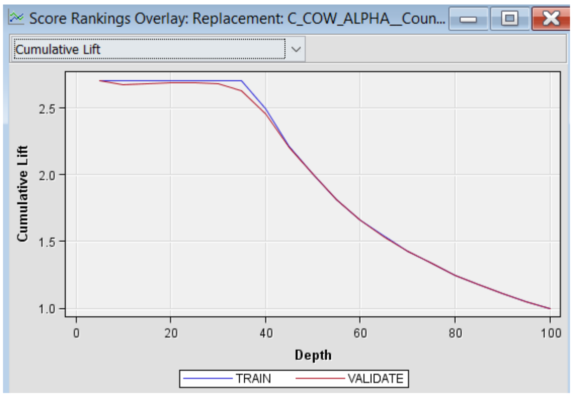
Based on the result, I can conclude that tuning the neural network has improved the model's performance as it more accurate and is performing well.

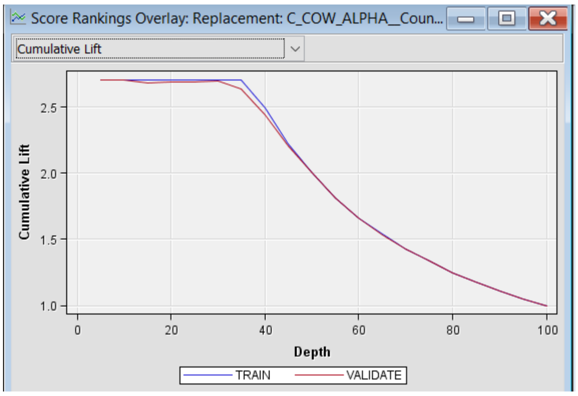


```
Replacement: C_COW_ALPHA__Country_code_CoW_al       3        664         .        385
Replacement: C_COW_ALPHA__Country_code_CoW_al      12        486        11        279
```

Cumulative Lift (Default vs ANN w Settings)

Cumulative Lift charts show the predictive effectiveness of the model. Based on the chart below, I observed that the lines in the chart for Train and Validate are relatively close to one another. This is an indication that both the models are not overly fitted.

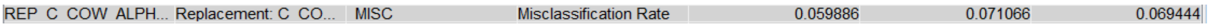




Decision Tree

Fit statistic

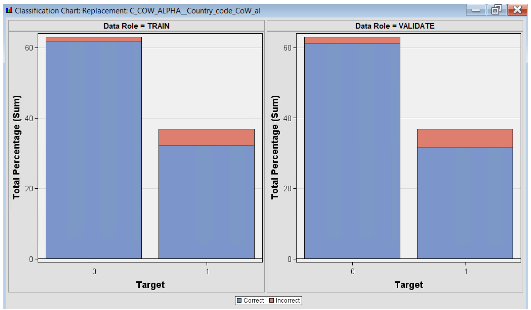
Leaving at its default settings for the decision tree we can see that the misclassification rate is higher compared to the other models. Based on the results we can foresee that using a decision tree may not give us the best performing model.



REP_C_COW_ALPH...	Replacement: C_CO...	_MISC_	Misclassification Rate	0.059886	0.071066	0.069444

Classification Chart

The red portion signifies the incorrect misclassification.Based on this chart we can conclude that this model is still acceptable however it is not as accurate or better than the Logistic regression models and neural network models.



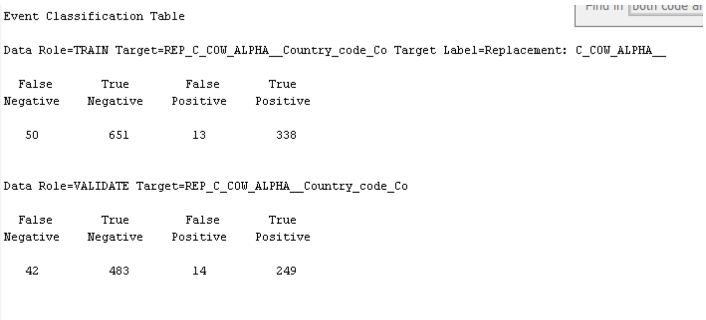
Classification Table

At first glance we notice that the number of false negatives and false positives under validate is higher than the previous two models.

Miscalculation rate: (42+14)/788 = 0.07106(5s.f)

Accuracy rate : (483+249)/788 = 0.92893(5s.f) around 92.8% accuracy**.**

Overall: we can conclude that the default decision tree accuracy is still acceptable as it did not perform better than the other models.

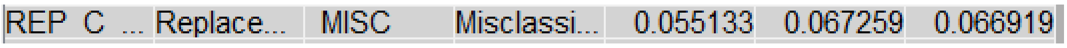


```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False      True      False      True
Negative   Negative   Positive   Positive

   50        651        13         338


Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False      True      False      True
Negative   Negative   Positive   Positive

   42        483        14         249
```

Decision Tree (Entropy)

Fit statistic

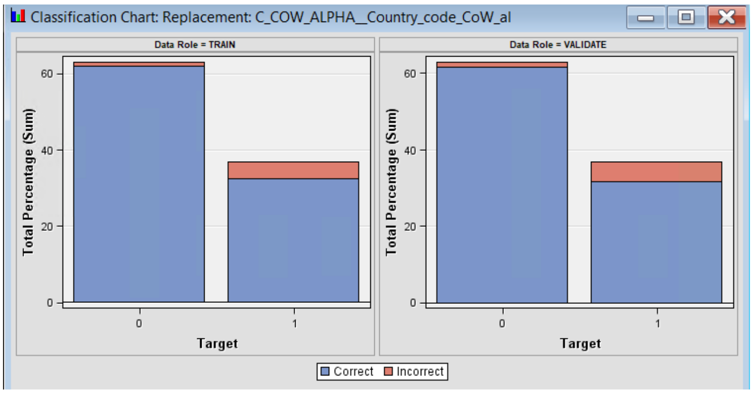
The misclassification rate for the automatic decision tree is slightly lower as compared to the baseline model. This shows that tuning the decision tree to entropy has improved the model’s performance.



REP_C ...	Replace...	MISC	Misclassi...	0.055133	0.067259	0.066919

Classification Chart

The red portion signifies the incorrect misclassification.Based on this chart we can conclude that this model is still performing well however it is not as accurate or better than the Logistic regression models and neural network models.



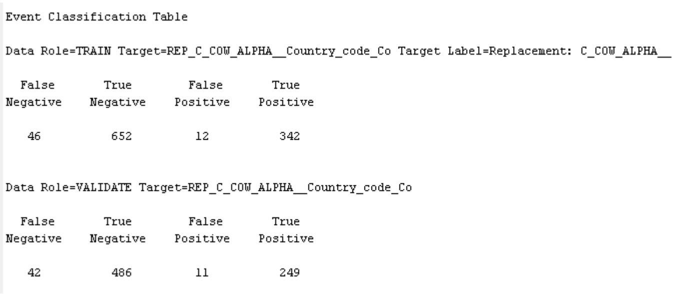
Classification Table

Although we did see an improvement after tuning the parameters, we still can see the number of false negative and false positive is higher as compared to the previous two models(logistic regression and Neural network)

Misclassification rate : (42+11)/788 = 0.06725(5s.f)

Accuracy rate : (486+249)/788=0.93274(5s.f) around 93.2% accuracy.

Therefore, I will not be choosing this model to do my prediction.

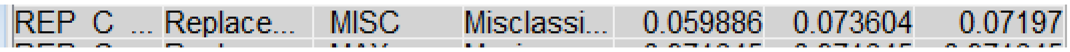


```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False      True      False      True
Negative   Negative   Positive   Positive

   46        652        12         342


Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False      True      False      True
Negative   Negative   Positive   Positive

   42        486        11         249
```

Decision Tree (Interactive)

Fit statistic

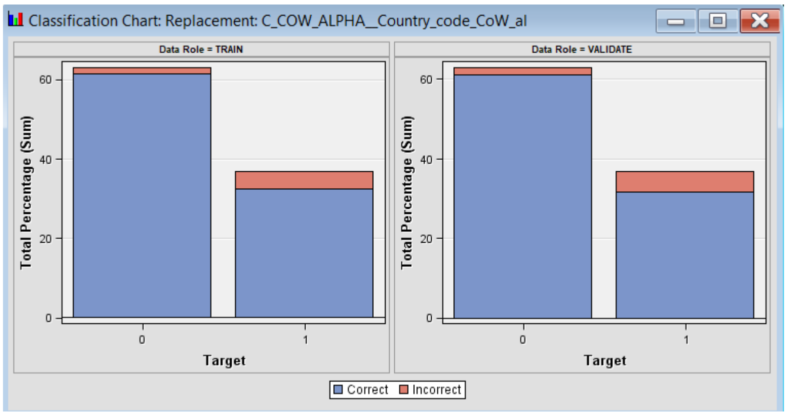
This decision tree is semi-automatic, based on the misclassification rate we can see that this model has performed the worst amongst the decision tree models. Despite pruning to a 3 level decision tree. Hence this model is not an ideal model to use to do the binary classification.



REP_C ...	Replace...	MISC	Misclassi...	0.059886	0.073604	0.07197

Classification Chart

This classification chart gives us an idea of how many were misclassified at the red portion.



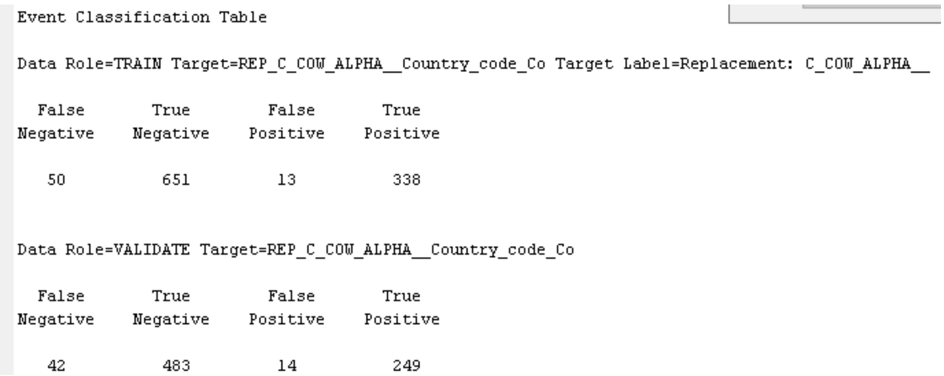
Classification Table

To better understand the model's performance. I have calculated the misclassification rate as well as the accuracy rate.

Misclassification Rate : (42+14)/788 = 0.07106(5s.f)

Accuracy Rate: (483+249)/788 = 0.92893(5s.f) around 92.8% accuracy

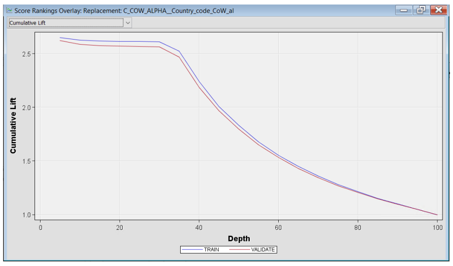
Overall, this is not an ideal model to use for binary classification as it is one of the lowest accuracy rates.

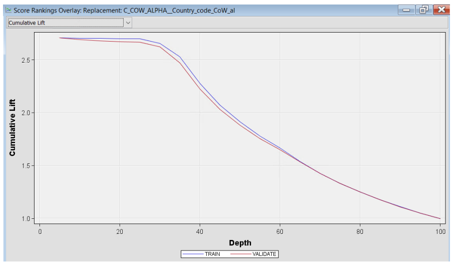


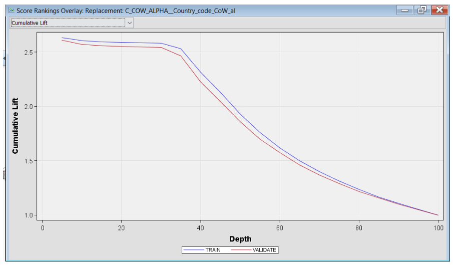
```
Event Classification Table

Data Role=TRAIN Target=REP_C_COW_ALPHA__Country_code_Co Target Label=Replacement: C_COW_ALPHA__

  False       True       False       True
Negative    Negative    Positive    Positive

   50         651         13          338


Data Role=VALIDATE Target=REP_C_COW_ALPHA__Country_code_Co

  False       True       False       True
Negative    Negative    Positive    Positive

   42         483         14          249
```

Cumulative Lift (Decision Tree Default Vs Decision Tree- Entropy Vs Decision Tree - Interactive)

Cumulative Lift charts show the predictive effectiveness of the model. Based on the chart below, I observed that the lines in the chart for Train and Validate are relatively close to one another. This is an indication that both the models are not overly fitted.



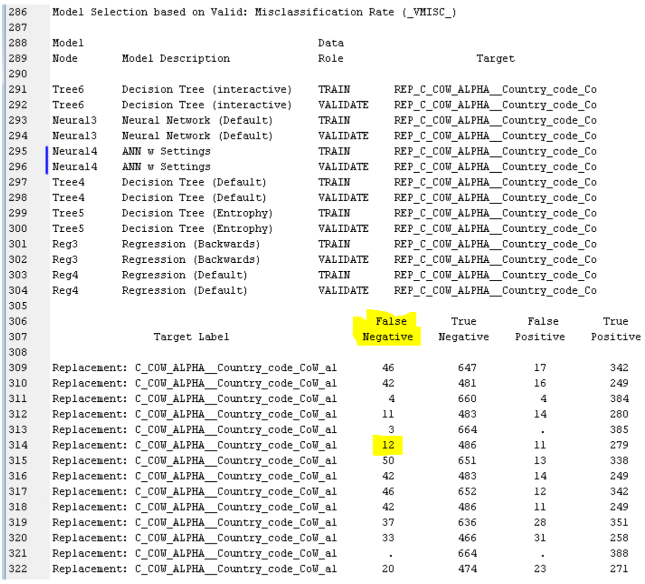




Model Comparison (Top 15 variables used)

Overall the best performing model is ANN w settings with only 0.029188 misclassification rate. And the accuracy rate for the best performing model is 97.0%.
The worst performing model is the backwards regression model with 0.81218 misclassification rate. while the accuracy is 91.8%

```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)


                                                                  Valid:
Selected    Model                                          Misclassification
 Model      Node       Model Description                         Rate

   Y        Neural4    ANN w Settings                          0.029188
            Neural3    Neural Network (Default)                0.031726
            Reg4       Regression (Default)                    0.054569
            Tree5      Decision Tree (Entrophy)                0.067259
            Tree4      Decision Tree (Default)                 0.071066
            Tree6      Decision Tree (interactive)             0.073604
            Reg3       Regression (Backwards)                  0.081218
```



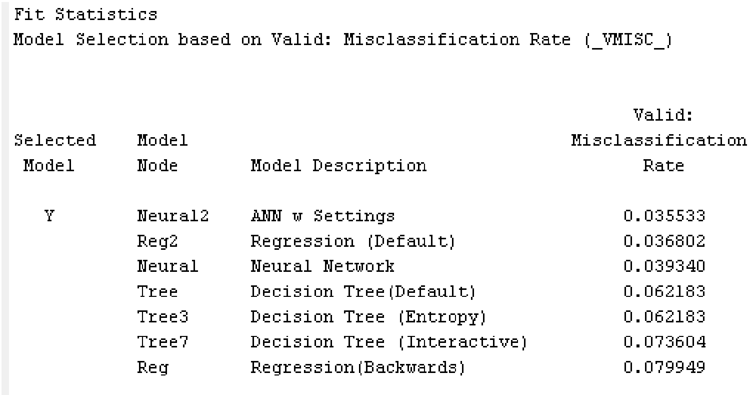
```
Model Selection based on Valid: Misclassification Rate (_VMISC_)

Model                                      Data
Node       Model Description               Role                  Target

Tree6      Decision Tree (interactive)     TRAIN      REP_C_COW_ALPHA__Country_code_Co
Tree6      Decision Tree (interactive)     VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Neural3    Neural Network (Default)        TRAIN      REP_C_COW_ALPHA__Country_code_Co
Neural3    Neural Network (Default)        VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Neural4    ANN w Settings                  TRAIN      REP_C_COW_ALPHA__Country_code_Co
Neural4    ANN w Settings                  VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Tree4      Decision Tree (Default)         TRAIN      REP_C_COW_ALPHA__Country_code_Co
Tree4      Decision Tree (Default)         VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Tree5      Decision Tree (Entrophy)        TRAIN      REP_C_COW_ALPHA__Country_code_Co
Tree5      Decision Tree (Entrophy)        VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Reg3       Regression (Backwards)          TRAIN      REP_C_COW_ALPHA__Country_code_Co
Reg3       Regression (Backwards)          VALIDATE   REP_C_COW_ALPHA__Country_code_Co
Reg4       Regression (Default)            TRAIN      REP_C_COW_ALPHA__Country_code_Co
Reg4       Regression (Default)            VALIDATE   REP_C_COW_ALPHA__Country_code_Co

                                                False       True      False      True
             Target Label                      Negative   Negative   Positive   Positive

Replacement: C_COW_ALPHA__Country_code_CoW_al     46         647        17         342
Replacement: C_COW_ALPHA__Country_code_CoW_al     42         481        16         249
Replacement: C_COW_ALPHA__Country_code_CoW_al      4         660         4         384
Replacement: C_COW_ALPHA__Country_code_CoW_al     11         483        14         280
Replacement: C_COW_ALPHA__Country_code_CoW_al      3         664         .         385
Replacement: C_COW_ALPHA__Country_code_CoW_al     12         486        11         279
Replacement: C_COW_ALPHA__Country_code_CoW_al     50         651        13         338
Replacement: C_COW_ALPHA__Country_code_CoW_al     42         483        14         249
Replacement: C_COW_ALPHA__Country_code_CoW_al     46         652        12         342
Replacement: C_COW_ALPHA__Country_code_CoW_al     42         486        11         249
Replacement: C_COW_ALPHA__Country_code_CoW_al     37         636        28         351
Replacement: C_COW_ALPHA__Country_code_CoW_al     33         466        31         258
Replacement: C_COW_ALPHA__Country_code_CoW_al      .         664         .         388
Replacement: C_COW_ALPHA__Country_code_CoW_al     20         474        23         271
```

Model Comparison (Top 8 variables)

The best performing model is ANN w Setting with 0.035533 misclassification rate and the accuracy is 96.4%

While the worst performing model is Logistic Regression (backwards) with 0.079949 misclassification and the accuracy is 92.0%

Although the worst performing model has a slightly higher accuracy rate of 0.2% than the Top 15’s worst model it still doesn't mean that we should use the Top 8 variable to do our predictive modelling. As we should be comparing the best accuracy rate to minimise any errors when performing the prediction.



```
Fit Statistics
Model Selection based on Valid: Misclassification Rate (_VMISC_)


                                                      Valid:
Selected   Model                                 Misclassification
 Model     Node       Model Description                 Rate

   Y       Neural2    ANN w Settings                  0.035533
           Reg2       Regression (Default)            0.036802
           Neural     Neural Network                  0.039340
           Tree       Decision Tree(Default)          0.062183
           Tree3      Decision Tree (Entropy)         0.062183
           Tree7      Decision Tree (Interactive)     0.073604
           Reg        Regression(Backwards)           0.079949
```

Comparison between the champion model

Top 15’s Champion Model : ANN w Setting with 0.029188 misclassification (about 0.029%)
Top 15’s Champion Model Accuracy : 97.0%

Top 8’s Champion Model : ANN w setting with 0.035533 misclassification rate. (about 0.035%)
Top 8’ Champion Model Accuracy: 96.4%

Comparing both champion models, we can see that there is a 0.006% difference in misclassification rate as well as a 0.6% accuracy difference.
This proves that the number of variables used does impact the model’s performance and accuracy. The more variables used the lower the misclassification rate.
Despite a small difference, in predictive modelling accuracy is crucial and one of the important factors hence the best performing model is Top 15 ANN w Setting.

Reason for used variables

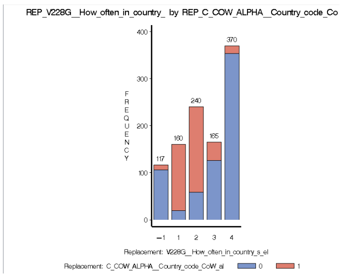
The reason why I have chosen to use those top 15 variables is because they have a high relation to the targeted field(C_COW_ALPHA__Country_code_CoW_al). Most of the selected variables were questions based on how the respondents feel towards their country's political system and how it is governed. An example is V228G, where rich people buy elections and V228D where voters are bribed. Based on the charts we can see that V228G and V228D do not happen to respondents from “SIN” whereas respondents from “TAW” stated that it happens often in their national election. Using such variables as the key predictors, we can predict where the respondents are from.

Recommendations for Policy Makers (20 marks, 10%)

0 is for SIN(blue) , 1 for TAW (red)

1)Rich People buy elections

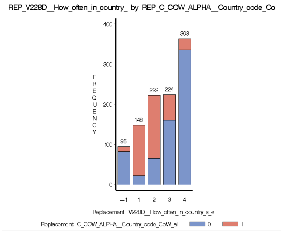
Based on V228G, rich people buy elections in the country. We can see that a significant number of “TAW” respondents say that rich people buy elections very often and fairly often. However, according to the respondents from “SIN” this rarely happens and not at all often. Therefore I would propose to create a law to disqualify the election party that uses such tactics. As for the voters who are involved, their votes should be rejected and both parties are to receive punishment from the law such as banning them from participating in any elections in the future.

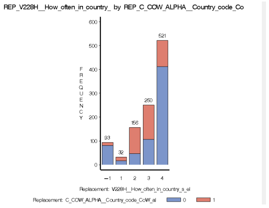


2)Voters are bribed
3)Voters are threatened at the polls

Based on the VH228D chart on Voters are bribed in elections we can see that A large number of respondents from “TAW” stated bribes in elections are very often(1) and fairly often (2). This means that the election system is flawed.
Thus my recommendation is to impose a law to ensure that voting is fair in the elections . The voters who are bribed will be severely punished such as they will not be able to partake in any elections for next 10 years and their rights will be taken away.

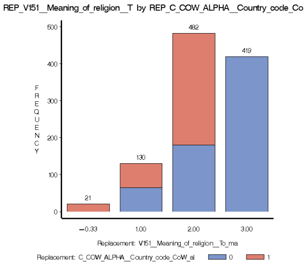
Based on V228H, we can see that voters are threatened at the polling station. As a democratic country, this should not be happening as voters are entitled to their own voting rights and they should not feel threatened. As such my recommendation for the policy makers would be to have civil servants to guard the polling booth to ensure the citizen’s safety.

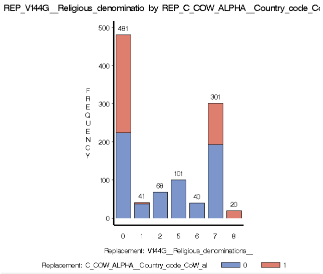




<u>4) V151 Meaning of religion and V144G Religion denomination</u>

Based on the charts below, we can see that Both “SIN” and “TAW” have a few different types of Religion. Hence, my recommendation is to have a policy that promotes religious harmony. There should not be any criticism or decrimination amongst the different religious groups.

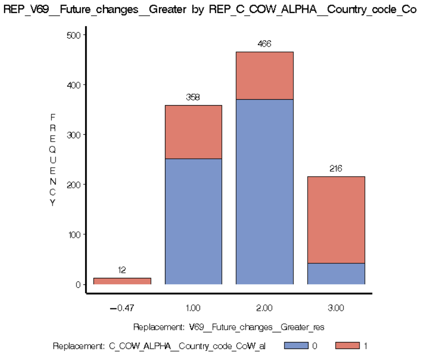




5)V69 is about greater respect for authorities

Based on this chart, We can see that the majority of the respondents feel that giving respect to authorities is good/ don't mind.

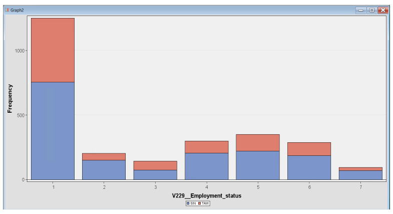
To encourage more people to show their respect to authorities my recommendation would be through education and cultivating the habit of giving respect since young.

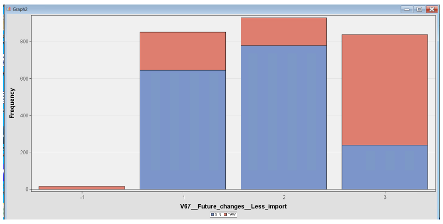


6)In V67, is about their view less importance placed on work in our lives

Based on the chart on the V229, we can see that the majority of the respondents are full time employed working for 30 hours a week or more, especially respondents from "SIN". Assuming that the respondents feel overworked thus the respondent feels that it is good to place less importance on work in their lives.

Hence, I would recommend both countries to have a work life balance. One way of work life balance is to reduce the number of working days or offer flexible working hours. By doing so, this could lead to an increase in productivity and the workers are able to have some time for themselves to do self improvement spending time with their loved ones and ease their stress so they won't feel burnout.





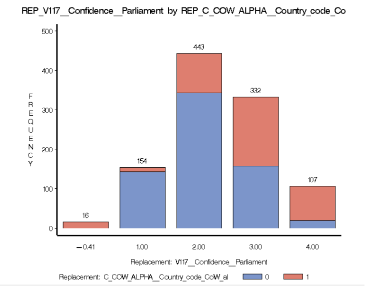
7) V117 is about how confident are the respondents in the parliament.

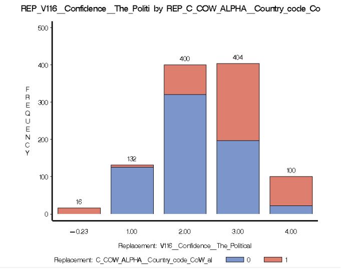
We can see that most of the respondents from "SIN" have a lot of confidence in the parliament.On the other hand respondents from "TAW" do not have much of confidence if not no confidence at all. This shows that the parliament system in the "TAW" country is flawed as a result their citizens do not have confidence in them. Hence, my recommendation to boost their confidence in the parliament, the government has to be more transparent.

8)V116 is about the confidence in the political parties

Similarly, we can see that respondents from "SIN" have confidence in the political system.

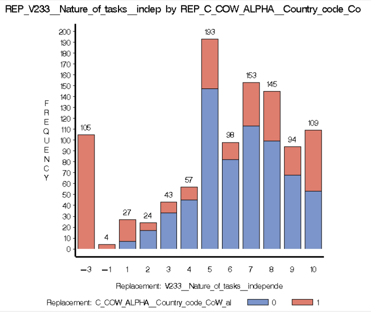
However, respondents from “TAW” think otherwise. Therefore my recommendation to boost their citizen confidence is that the government should focus on economic development, education and cross-strait relation issues.





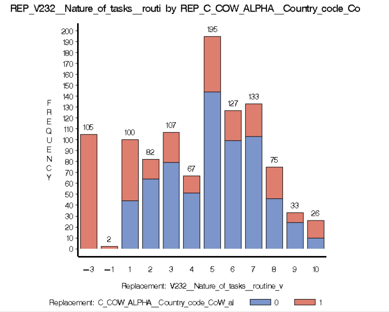
9) Independence

Based on this chart, we can see that the majority of the respondents have ranked their independence relatively high. My recommendation to policy makers to promote independence in the workplace is to clearly explain their roles and expectations. By doing so, workers are less likely to seek feedback and approval for every decision they make. This will also make the workers take ownership of one’s actions.



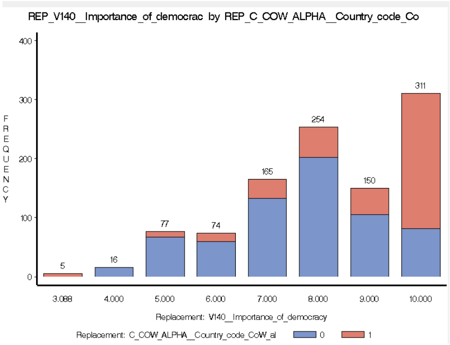
10)routine VS creative task

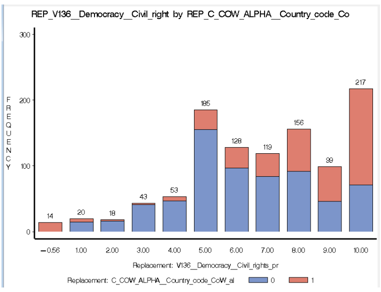
Assuming that people in “TAW” and “SIN” are tasked to follow orders of their superior, workers aren’t given many opportunities where they can speak up their mind. As a result, based on this chart not many people picked option 9 and 10. Hence my recommendation to the policy makers is that superior should highly encourage their workers to voice out their opinion and thoughts to their project. This way, it will not only allow the department to bond better as a team but also improve the project. By doing so , the workers will then have more opportunities to display their creative side.



11)Importance for people to live in a country governed democratically
12) Civil rights protect people from state oppression

Based on Chart V140, we can see that the respondents value democracy as they have ranked the importance of democracy relatively high as most of them choose value 6 onwards. Being in a democractic country, it gives/encourages people to speak their mind. In Singapore, there is a speaker's corner where citizens can express themselves as long as they comply with the terms and conditions and adhere to the restrictions.
Likewise in V136, people also view civil rights as one of the characteristics that is essential to a democractic country.
Hence my recommendation to the policy makers is to have a day where citizens are able to exercise their freedom of speech at the speaker's corner place.This is provided that the applicant must have the license to proceed and there will be rules to follow to prevent any disputes/protest.





References (if applicable)

https://www.brookings.edu/articles/taiwans-democracy-and-the-china-challen

***** END OF ASSIGNMENT *****