Georgia State University | J. MACK ROBINSON COLLEGE OF BUSINESS

Term Project

CIS8685 Big Data Analytics, Spring 18

- Shagun Garg
- Ashish Devrani
- Supreet Kaur
- Vrushali Shrikant Tarawade

Project Description

Context

The UK government amassed traffic data from 2000 and 2016, recording over 1.6 million accidents in the process and making this one of the most comprehensive traffic data sets out there. It's a huge picture of a country undergoing change.

Note that all the contained accident data comes from police reports, so this data does not include minor incidents.

Accidents data is only a part of the data recorded by the UK Government: accidents_2012_to_2014.csv. The total time is 2012 through 2014.

Business Problem

Road Accidents are a big reason why people do not feel safe on the roads in UK. The number of accidents increase each year and with that we have an increasing number of casualties as well. From the data set available at our disposal, we are trying to find the major factors which act as a catalyst towards a road accident so that in future necessary steps can be taken to act on these factors.

Acknowledgements

The license for this dataset is the Open Government License used by all data on data.gov.uk (here). The raw datasets are available from the UK Department of Transport website here.

Goals

From this dataset, we are trying to find the major contributors towards a road accident and how they affect the number of casualties in a road accident. Below are the questions that we hope to answer by developing an analytical model towards predicting the number of casualties in a road accident:

1. Have the Light Conditions at the time of the accident played a role in the increase/decrease in the number of casualties?
2. Did a specific day in the week have a high number of casualties?
3. Was the road classified as a high speed limit zone?
4. Can we predict accident rates over time? What might improve accident rates?

Data Exploration and Preprocessing

Data Exploration

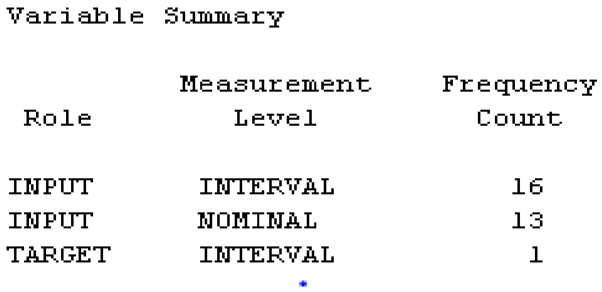
The data obtained from the Kaggle dataset was explored to look for available data, its consistency, data types.
From the available data, a business problem was designed as explained above and in line with the same the predictors and target variables were set.

Output Variable: Since our aim is to determine the factors and conditions affecting the Casualties, we have selected "**Number of Casualties"** as our Target Variable.
Diagram for the same is Available under the Appendix as Figure 1.

Understanding the Data

Let us understand which Variable could affect the output of the Target Variable. For this, we are first using the **"Variable Selection"** Node.

Variable Selection: The Variable Selection node assists in reducing the number of inputs by setting the status of the input variables that are not related to the target as Rejected and variables are not used as model inputs by a successor modeling node. The Variable summary of our model:



```
Variable Summary

           Measurement    Frequency
 Role         Level         Count

INPUT       INTERVAL          16
INPUT       NOMINAL           13
TARGET      INTERVAL           1
```

Please refer to Figures 2 and 3 under the Appendix for a screenshot of the Variable Summary and Variable selection.

From the diagram, we see that the variables which are of least importance to the target variable are rejected and the reason of rejection is also stated in the output. In our model, only 5 variables are selected.
This decision was made post selecting logically the most impactful variables and setting them as input for our model.
The only variable selected which could be found only post the incident was 'Did Police Reach the Scene'. This is because this proved to be an important predictor and though its value is unknown in our case, an approximate value can be decided while feeding data based on the location and the data available with the police force and its office locations while making the actual predictions.

The data was also explored using data exploration techniques to find relations between variables. The visualizations are available a report generated. This was done in both R and SAS and a report of R is attached in the appendix as Element 4

Model

Models Used

In our case, the variable Number_of_Casualties is a categorized variable with 34 different categories of results. On further exploration of data we found that around 90% of the values for this variable were 1. This kind of data set was not ideal to be fitted into a model as a categorized variable, hence, we modified the variable into a binary variable considering two values using the replacement node:

a) 1 Casualty
b) Multiple Casualties

This transformation resulted in the data getting a good fit in for our models and resulted in less error percentage.

To address the business problem of predicting the number of casualties, we have used three models:

Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

The variables used as part of this model are:

1. Did_Police_Officer_Attend_Scene
2. REP_Accident_Severity
3. REP_Number_of_Casualties
4. REP_Number_of_Vehicles
5. REP_Speed_Limit
6. Special_Conditions_at_Site
7. Weather Conditions
8. Dummy Variables of Carriageaway_Hazards
9. Dummy Variables of Light_Conditions
10. Dummy Variables of Road_Surface_Conditions
11. Dummy Variables of Road_Type
12. Dummy Variables of Urban_or_Rural_Area

A snapshot of the regression variables is available under the Appendix as Figure 5 and the results of the model are available as Figure 6.

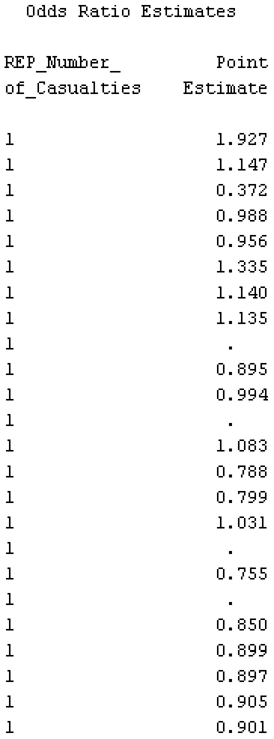
The summary of the variables is given below:

Variable Summary

	Measurement	Frequency
Role	**Level**	**Count**
INPUT	BINARY	16
INPUT	INTERVAL	3
INPUT	NOMINAL	2
REJECTED	INTERVAL	14
REJECTED	NOMINAL	9
TARGET	NOMINAL	1

Predicted and Decision Variables:

Type	**Label**	**Measurement Level**	**Frequency Count**
TARGET	Replacement: Number_of_Casualties	BINARY	16
PREDICTED	Predicted: REP_Number_of_Casualties=1	INTERVAL	3
RESIDUAL	Residual: REP_Number_of_Casualties=1	NOMINAL	2
PREDICTED	Predicted: REP_Number_of_Casualties=0	INTERVAL	14
RESIDUAL	Residual: REP_Number_of_Casualties=0	NOMINAL	9
FROM	From: REP_Number_of_Casualties	NOMINAL	1
INTO	Into: REP_Number_of_Casualties		



```
  Odds Ratio Estimates

REP_Number_          Point
of_Casualties     Estimate

1                    1.927
1                    1.147
1                    0.372
1                    0.988
1                    0.956
1                    1.335
1                    1.140
1                    1.135
1                     .
1                    0.895
1                    0.994
1                     .
1                    1.083
1                    0.788
1                    0.799
1                    1.031
1                     .
1                    0.755
1                     .
1                    0.850
1                    0.899
1                    0.897
1                    0.905
1                    0.901
```

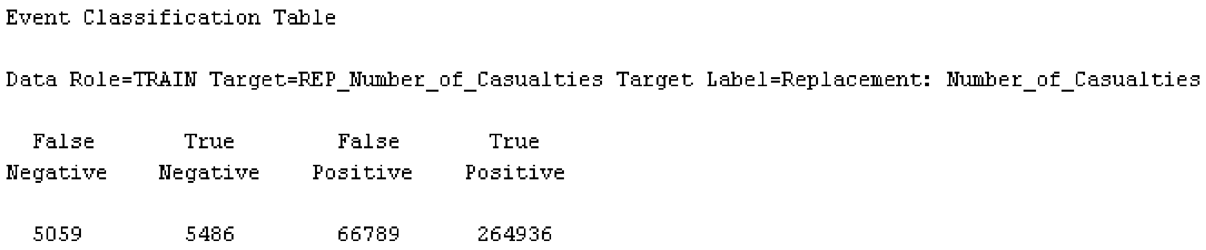
Model Specification is available in the appendix as figure 7.

The column of estimates (coefficients or parameter estimates, from here on labeled coefficients) provides the values for intercept and the variables for this equation. Expressed in terms of the variables used in this example, the regression equation is

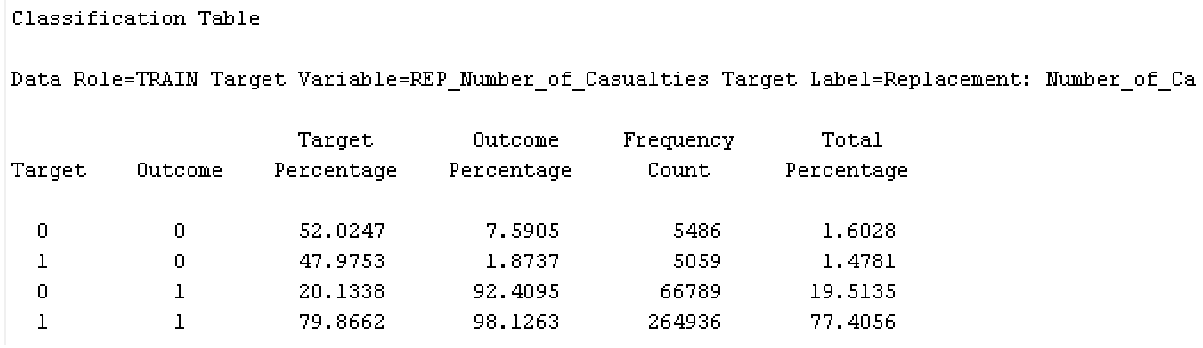
REP_Number_of_Casualties(Logit) = -3.4129 + [0.3281]* Did_Police_Officer_Attend_Scene + [0.1374]* REP_Accident_Severity + [-0.9878]* REP_Number_of_Vehicles + [-0.0124]* REP_Speed_limit + [-0.0224]* Special_Conditions_at_Site + [0.1444]* TI_Light_Conditions1 + [0.0653]* TI_Light_Conditions2 + [0.0632]* TI_Light_Conditions3 + [-0.0557]* TI_Road_Surface_Conditions1 + [-0.00283]* TI_Road_Surface_Conditions2 + [0.0397]* TI_Road_Type1 + [-0.1192]* TI_Road_Type2 + -0.1125]* TI_Road_Type3 + 0.0154]* TI_Road_Type4 + [-0.1403]* TI_Urban_or_Rural_Area1 + [-0.0659]* Weather_Conditions1 + [-0.00910]* Weather_Conditions2

+ [-0.0119]* Weather_Conditions3 + [-0.00293]* Weather_Conditions4 + [-0.00729]* Weather_Conditions5

From the coefficient estimates we can say that for every unit increase in REP_Accident_Severity, the odds of REP_Number_of_Casualties increase by 0.1374 units.



```
Event Classification Table

Data Role=TRAIN Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

  False      True       False      True
Negative   Negative   Positive   Positive

  5059       5486       66789      264936
```



```
Classification Table

Data Role=TRAIN Target Variable=REP_Number_of_Casualties Target Label=Replacement: Number_of_Ca

                      Target      Outcome     Frequency     Total
Target   Outcome    Percentage   Percentage     Count     Percentage

  0         0         52.0247      7.5905        5486       1.6028
  1         0         47.9753      1.8737        5059       1.4781
  0         1         20.1338     92.4095       66789      19.5135
  1         1         79.8662     98.1263      264936      77.4056
```

High %age of True Positives and True Negatives indicates that the model is well fitted and good to make the predictions based on our business problem.

A Fit Statistics of the model is available in the Appendix as Figure 8.

Low %age of Average Squared Error and RMSE also vindicate the fact that the model is well-fitted and good to use.

The motive in applying Regression model first was to remove unnecessary variables from the input for other models. Stepwise selection was used in the regression model and post 4 iterations the independent variables were finalized. The predictors discarded in this process were:

a. Weather conditions
b. Special at sight conditions
c. carriageway hazards

MODEL 2: DECISION TREE:

Decision tree is mostly used in classification problems. It works for categorical as well as continuous target and predictors. In this technique, we split the data into two or more homogeneous sets based on most significant differentiating factor from the input variables.

To maintain uniformity, we have used the same input variables as those used for logistic regression.

The results obtained are summarized below in three steps:

1. The rules used by the decision tree

2. The fit statistics that shows the summary of performance of the model
3. The lift ratio and other output parameters.

RULES:

The rules for the decision tree from the tree diagram can be summarized as follows:
NOTE:Here Number_of_Casualties=1 means that the Casualties in the accident are greater than 1and Number_of_Casualties=0 represents the case where casualties =1

if Number_of_Vehicles < 1.5 then Number_of_Casualties=1

if Number_of_Vehicles < 2.5 AND Replacement: Number_of_Vehicles >= 1.5 or MISSING then Number_of_Casualties=1

if Speed_limit < 35 or MISSING AND Number_of_Vehicles >= 2.5 then number_of_Casualties=1

if Speed_limit >= 35 AND Number_of_Vehicles >= 2.5 AND Accident_Severity < 2.5 AND Did_Police_Officer_Attend_Scene_ IS ONE OF: YES or MISSING then Number_of_Casualties=0

if Speed_limit >= 35 AND Number_of_Vehicles >= 2.5 AND Accident_Severity < 2.5 AND Did_Police_Officer_Attend_Scene_ IS ONE OF: NO
then Number_of_Casualties=0

if Speed_limit >= 35 AND Number_of_Vehicles >= 2.5 AND Accident_Severity >= 2.5 or MISSING AND Did_Police_Officer_Attend_Scene_ IS ONE OF: NO
then Number_of_Casualties=1

if Road_Type: Roundabout IS ONE OF: 0 or MISSING AND Replacement: Speed_limit >= 35 AND Replacement: Number_of_Vehicles >= 2.5 AND Replacement: Accident_Severity >= 2.5 or MISSING AND Did_Police_Officer_Attend_Scene_ IS ONE OF: YES or MISSING then Number_of_Casualties=0

if Road_Type:Roundabout IS ONE OF: 1 AND Replacement: Speed_limit >= 35 AND Replacement: Number_of_Vehicles >= 2.5 AND Replacement: Accident_Severity >= 2.5 or MISSING AND Did_Police_Officer_Attend_Scene_ IS ONE OF: YES or MISSING
then Number_of_Casualties=1

Please Refer to Appendix Figure 9 for the Tree Diagram.

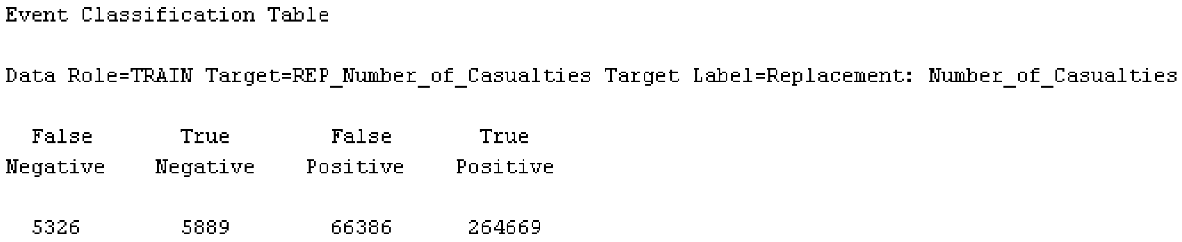
FIT STATISTICS:
The ASE (Average Squared Error) for this model is as below:

	Train	Validation
Average Squared Error	0.15783547705221687	0.15816763530485944

This is comparable to that of the Logistic Regression model and the low error shows that the model has fit the data well. This also means that the model is good.

OTHER PARAMETERS:

Event Classification:



```
Event Classification Table

Data Role=TRAIN Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

  False      True      False      True
Negative   Negative   Positive   Positive

  5326       5889      66386     264669
```

From the above event classification table it can be inferred that the high ratio of true positives and True Negatives to that of the total count shows that the model is a good fit and has very few misclassified records.

The sensitivity and accuracy for this model are very high:

Sensitivity = (264669/(264669+5326))*100 = 98%

Accuracy = (True positive + true negative) / (total) = 0.79

Lift Ratio:

The Figure for the Lift Ratio is available in the appendix as figure 10.

The lift is good till a depth of around 38. Most of the records get classified till this depth. Post that the difference in the vote percentage to the distinct output values is close indicating that some records are not distinguished distinctly.

MODEL 3: NEURAL NETWORK

Neural networks are a class of parametric models that accommodate a wider variety of nonlinear relationships between a set of predictors and a target variable. The most common neural network model is the Multilayer Perceptron (MLP) which is known as a **supervised network** because it requires a desired output to learn. Our output variable is **“Number of Casualties”.**

Neural Network Nodes: The Neural Network node trains a specific neural network configuration; this node is best used when you know a lot about the structure of the model that you want to define.

Please refer to Figure 11 in Appendix for the variables used.

- In the **Network Configuration**, change the **Number of Hidden Unit**s to 20. This example trains a multilayer perceptron neural network with 20 units on the hidden layer.
- In the **Optimization**, for Preliminary Training keep Enable as **“No”**.

The snapshots for these are available as figures 12 and 13 in the appendix.

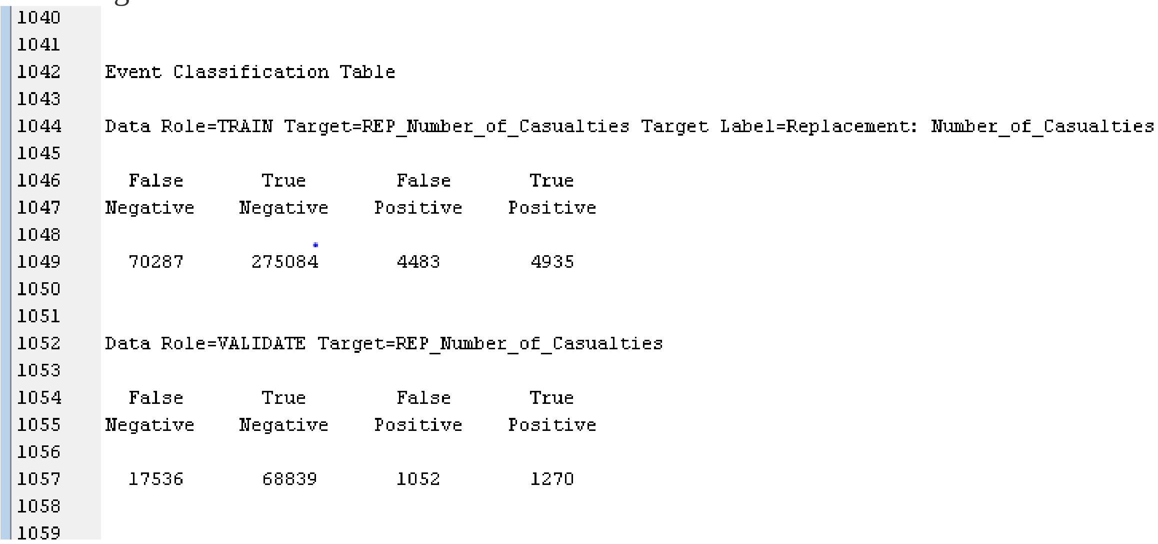
Post implementation of this model the results obtained have been captured and a snapshot of the same has been attached in the appendix as Figure 14.

Given Below is the event classification table from the result. We see that for the Training Data, we get the following results:

True Negative: 275084

True Positive: 4935

False Positive: 70287
False Negative: 4483



```
Event Classification Table

Data Role=TRAIN Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  70287      275084       4483        4935


Data Role=VALIDATE Target=REP_Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  17536       68839       1052        1270
```

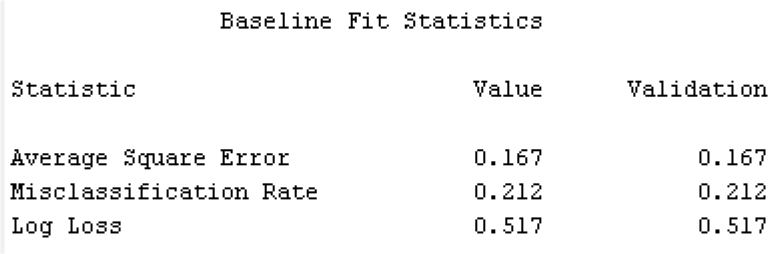
Accuracy = (True positive + true negative) / (total) =280019/354789= 0.7892
Sensitivity: (4935/ (4935+4483)) = 4935/9418=0.5239

Based on the Accuracy and Sensitivity values, we can say that this model is highly accurate, but not as sensitive.

RANDOM FOREST:

A random forest fits multiple classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
Hence, to get a model with improved accuracy and which is rightly fit, random forest has been used next.
The results from this model are captured and attached as Figure 15 in the appendix.

The fit Statistics for this model can be summarized as:



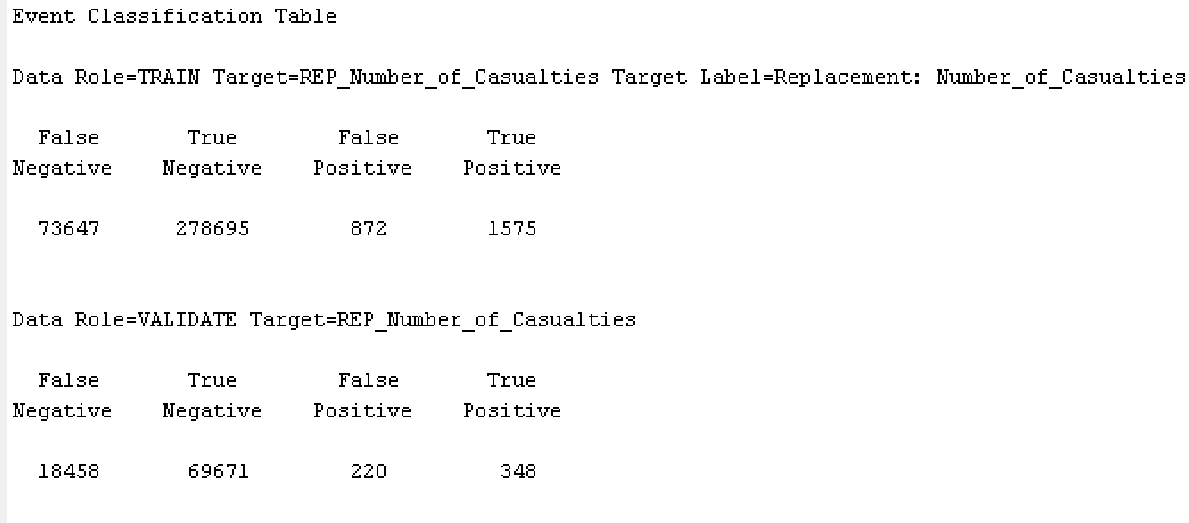
```
               Baseline Fit Statistics

Statistic                         Value      Validation

Average Square Error              0.167           0.167
Misclassification Rate            0.212           0.212
Log Loss                          0.517           0.517
```

It is observed that the value of the ASE and the misclassification rate is quite low.

From the event classification Table Below:
Accuracy = (348 + 69671) / (18458+69671+220+348) = 69671/88697 = 0.7855
Sensitivity = TP/(TP+FP) = 348/(348+220) = 0.6127

The above values reinforce the fact that with Random Forest Classification Accuracy increases and the rate of misclassification decreases.

The Event Classification table for this model is as below:

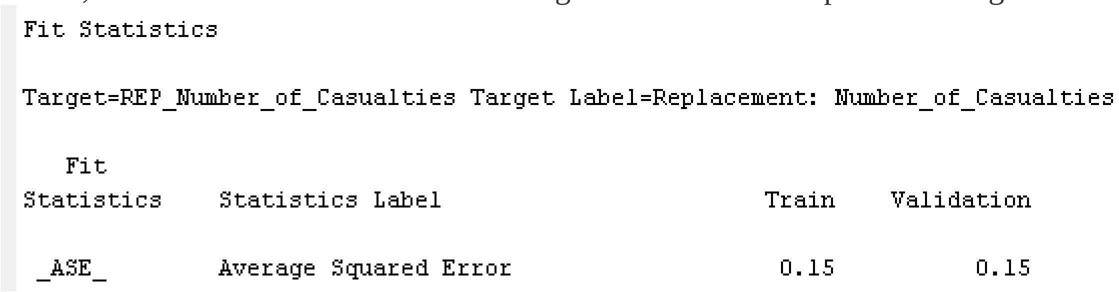


```
Event Classification Table

Data Role=TRAIN Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  73647      278695        872        1575


Data Role=VALIDATE Target=REP_Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  18458       69671        220         348
```

MODEL 5: ENSEMBLE

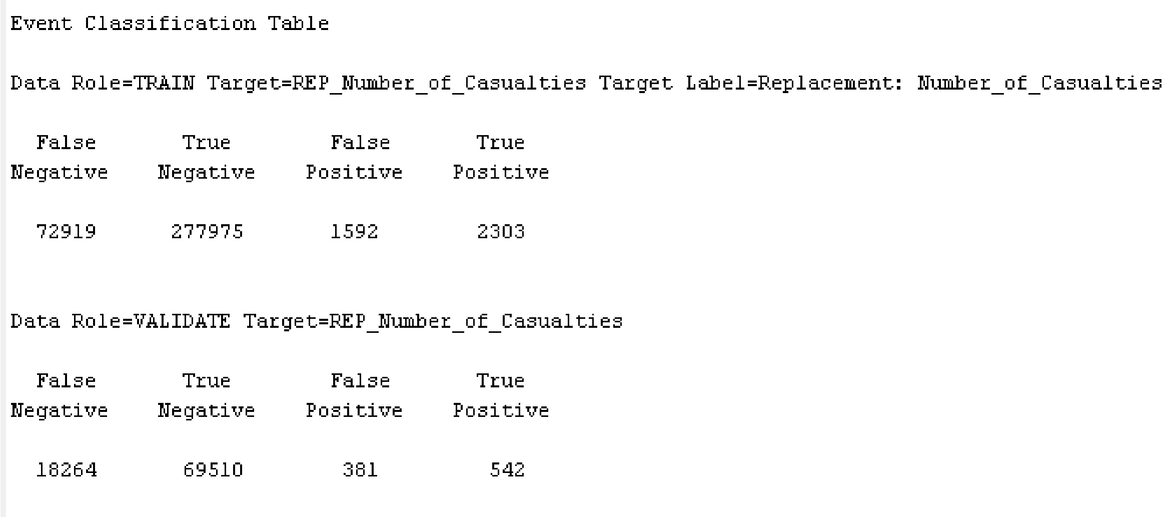
The goal of **ensemble methods** is to combine the predictions of several base estimators built with a given learning algorithm to improve generalizability / robustness over a single estimator.

Hence, the ensemble model is used next to get the best of the implemented algorithms.



```
Fit Statistics

Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

   Fit
Statistics    Statistics Label                        Train    Validation

 _ASE_        Average Squared Error                    0.15          0.15
```

The Fit Statistics for this model gives us the Average Squared Error to be 0.15, which is very low. This is also a better value as compared to the models implemented in previous steps.



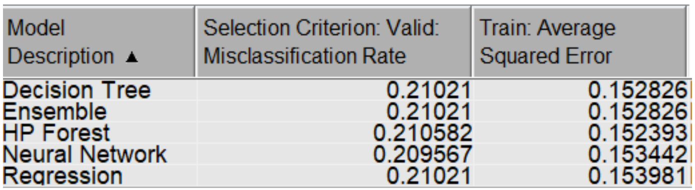
```
Event Classification Table

Data Role=TRAIN Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  72919      277975       1592        2303


Data Role=VALIDATE Target=REP_Number_of_Casualties

  False       True        False       True
Negative    Negative    Positive    Positive

  18264       69510        381         542
```

From the event classification table for this model, the accuracy and sensitivity values for the validation are calculated as:

Accuracy = (542 + 69510) / (18264+69510+381+542) = 70052/88697 = 0.7898
Sensitivity = TP/(TP+FP) = 542/ (542+381) = 542/923 = 0.5872156013001083

The accuracy is almost 79% and sensitivity is almost 59% for the ensemble model. Both these values indicate that this is an improved and a more robust model.

MODEL COMPARISON AND RESULTS SUMMARY:



Model Description ▲	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error
Decision Tree	0.21021	0.152826
Ensemble	0.21021	0.152826
HP Forest	0.210582	0.152393
Neural Network	0.209567	0.153442
Regression	0.21021	0.153981

A glimpse at the model comparison statistics shows that the models that have performed the best are the Decision Tree and the Ensemble Models. The Neural Network has a lower misclassification rate, but a comparatively higher ASE.

Ensemble has an output from other models as its input and hence, is computationally taxing. This implies that to fit our data and scenario, the Decision Tree has the best performance.

All this said, let us look at our business problem. Our goal is to predict whether the number of casualties are greater than 2 in an accident. To analyze this problem, we selected accuracy and sensitivity as the measuring units because, an accurate prediction would help the government avoid those situations with priority.

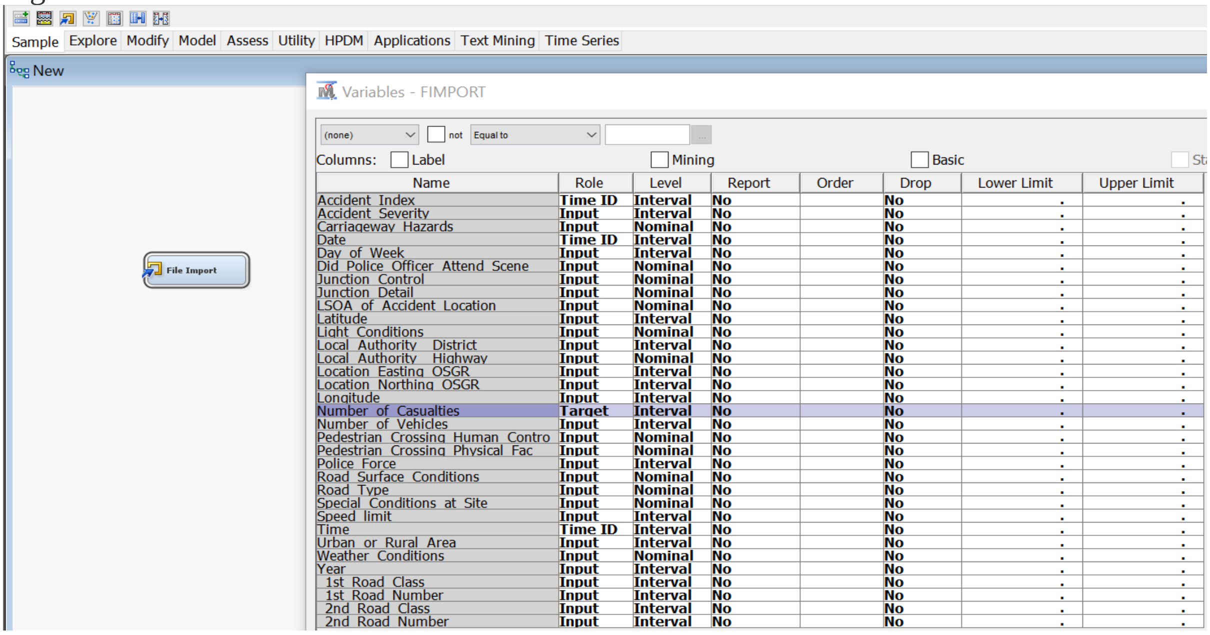
The accuracy of the ensemble model is the best, touching 79% which is equal to that of the decision tree.

As the next measuring factor, sensitivity of the models was considered which was again comparable for decision tree and the ensemble model.

This implies that the model that fits our data the best is the decision tree, which is computationally cheap, has a high performance with increasingly larger data sets, better accuracy and sensitivity and low Average Squared Error and misclassification rate.

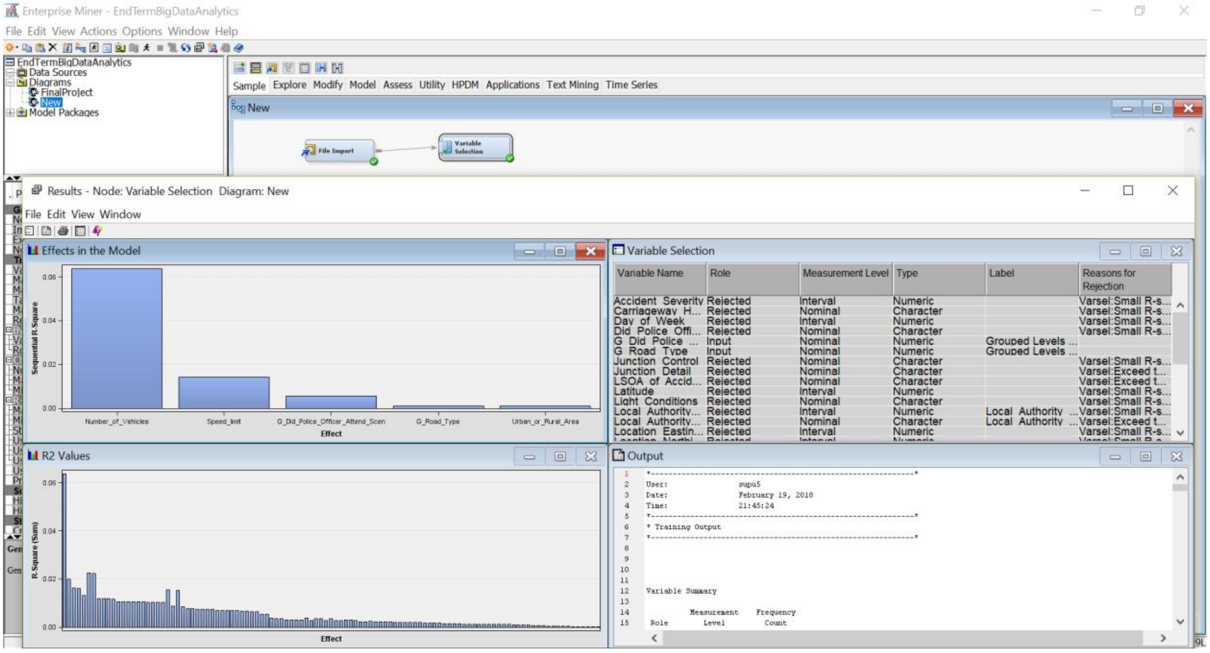
APPENDIX

Figure 1:



Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Accident Index	Time ID	Interval	No		No	.	.
Accident Severity	Input	Interval	No		No	.	.
Carriageway Hazards	Input	Nominal	No		No	.	.
Date	Time ID	Interval	No		No	.	.
Day of Week	Input	Interval	No		No	.	.
Did Police Officer Attend Scene	Input	Nominal	No		No	.	.
Junction Control	Input	Nominal	No		No	.	.
Junction Detail	Input	Nominal	No		No	.	.
LSOA of Accident Location	Input	Nominal	No		No	.	.
Latitude	Input	Interval	No		No	.	.
Light Conditions	Input	Nominal	No		No	.	.
Local Authority District	Input	Interval	No		No	.	.
Local Authority Highway	Input	Nominal	No		No	.	.
Location Easting OSGR	Input	Interval	No		No	.	.
Location Northing OSGR	Input	Interval	No		No	.	.
Longitude	Input	Interval	No		No	.	.
Number of Casualties	Target	Interval	No		No	.	.
Number of Vehicles	Input	Interval	No		No	.	.
Pedestrian Crossing Human Contro	Input	Nominal	No		No	.	.
Pedestrian Crossing Physical Fac	Input	Nominal	No		No	.	.
Police Force	Input	Interval	No		No	.	.
Road Surface Conditions	Input	Nominal	No		No	.	.
Road Type	Input	Nominal	No		No	.	.
Special Conditions at Site	Input	Nominal	No		No	.	.
Speed limit	Input	Interval	No		No	.	.
Time	Time ID	Interval	No		No	.	.
Urban or Rural Area	Input	Interval	No		No	.	.
Weather Conditions	Input	Nominal	No		No	.	.
Year	Input	Interval	No		No	.	.
1st Road Class	Input	Interval	No		No	.	.
1st Road Number	Input	Interval	No		No	.	.
2nd Road Class	Input	Interval	No		No	.	.
2nd Road Number	Input	Interval	No		No	.	.

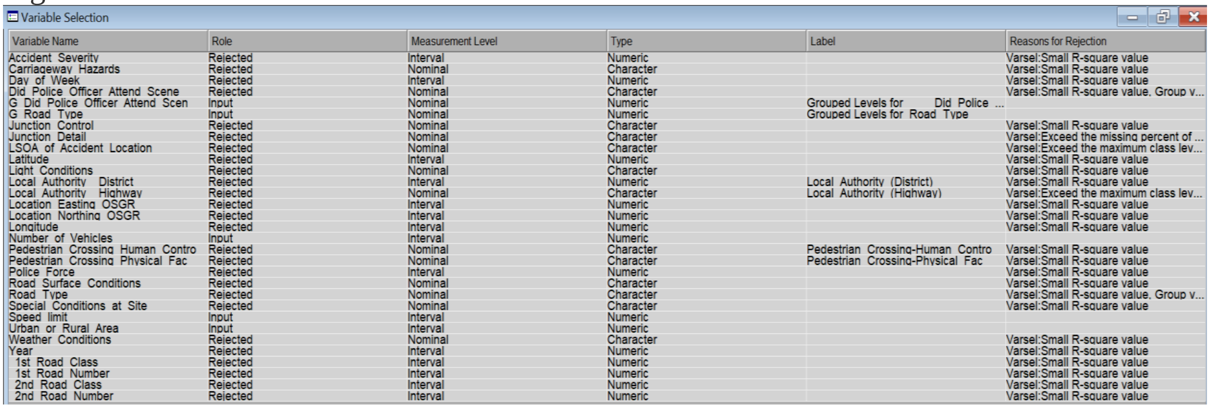
Figure 2: Variable Summary



Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
Accident Severity	Rejected	Interval	Numeric		Varsel:Small R-s...
Carriageway H...	Rejected	Nominal	Character		Varsel:Small R-s...
Day of Week	Rejected	Interval	Numeric		Varsel:Small R-s...
Did Police Offi...	Rejected	Nominal	Character		Varsel:Small R-s...
G Did Police ...	Input	Nominal	Numeric	Grouped Levels ...	
G Road Type	Input	Nominal	Numeric	Grouped Levels ...	
Junction Control	Rejected	Nominal	Character		Varsel:Small R-s...
Junction Detail	Rejected	Nominal	Character		Varsel:Exceed t...
LSOA of Accid...	Rejected	Nominal	Character		Varsel:Exceed t...
Latitude	Rejected	Interval	Numeric		Varsel:Small R-s...
Light Conditions	Rejected	Nominal	Character		Varsel:Small R-s...
Local Authority...	Rejected	Interval	Numeric	Local Authority ...	Varsel:Small R-s...
Local Authority...	Rejected	Nominal	Character	Local Authority ...	Varsel:Exceed t...
Location Eastin...	Rejected	Interval	Numeric		Varsel:Small R-s...

```
*------------------------------------------------------------*
User:                supu5
Date:                February 19, 2018
Time:                21:45:24
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

            Measurement     Frequency
 Role          Level          Count
```

Figure 3: Variable Selection



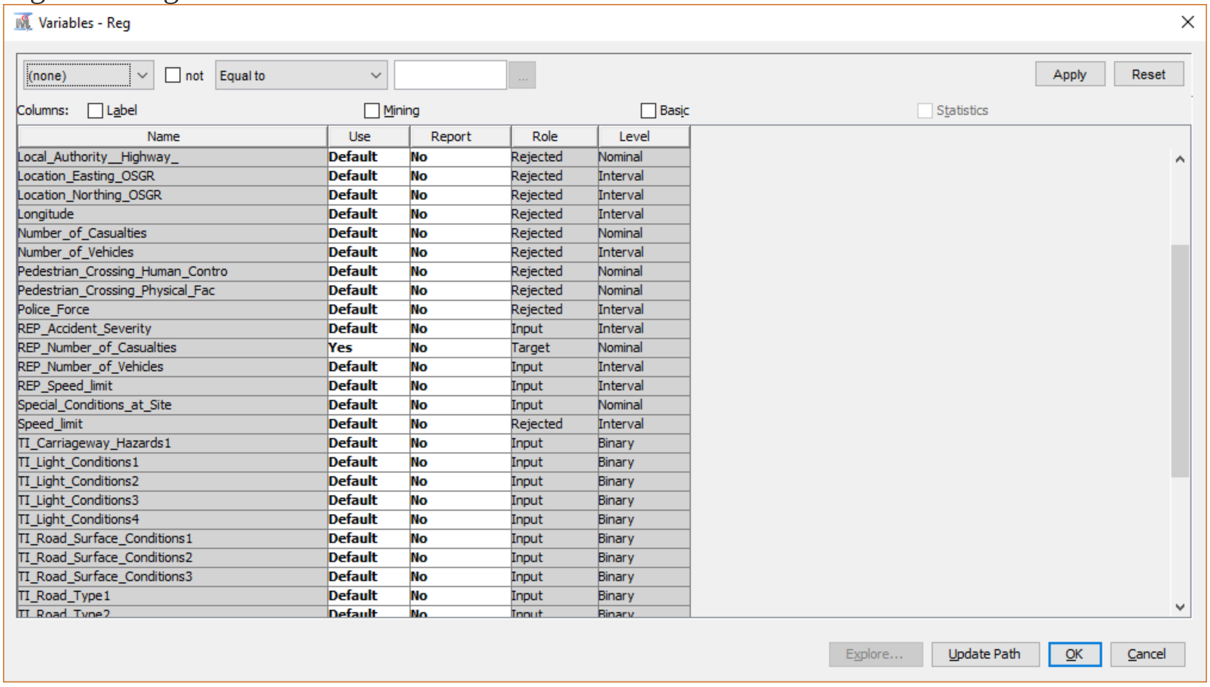
Variable Selection

Variable Name	Role	Measurement Level	Type	Label	Reasons for Rejection
Accident Severity	Rejected	Interval	Numeric		Varsel:Small R-square value
Carriageway Hazards	Rejected	Nominal	Character		Varsel:Small R-square value
Day of Week	Rejected	Interval	Numeric		Varsel:Small R-square value
Did Police Officer Attend Scene	Rejected	Nominal	Character		Varsel:Small R-square value, Group v...
G Did Police Officer Attend Scen	Input	Nominal	Numeric	Grouped Levels for Did Police ...	
G Road Type	Input	Nominal	Numeric	Grouped Levels for Road Type	
Junction Control	Rejected	Nominal	Character		Varsel:Small R-square value
Junction Detail	Rejected	Nominal	Character		Varsel:Exceed the missing percent of ...
LSOA of Accident Location	Rejected	Nominal	Character		Varsel:Exceed the maximum class lev...
Latitude	Rejected	Interval	Numeric		Varsel:Small R-square value
Light Conditions	Rejected	Nominal	Character		Varsel:Small R-square value
Local Authority District	Rejected	Interval	Numeric	Local Authority (District)	Varsel:Small R-square value
Local Authority Highway	Rejected	Nominal	Character	Local Authority (Highway)	Varsel:Exceed the maximum class lev...
Location Easting OSGR	Rejected	Interval	Numeric		Varsel:Small R-square value
Location Northing OSGR	Rejected	Interval	Numeric		Varsel:Small R-square value
Longitude	Rejected	Interval	Numeric		Varsel:Small R-square value
Number of Vehicles	Input	Interval	Numeric		
Pedestrian Crossing Human Contro	Rejected	Nominal	Character	Pedestrian Crossing-Human Contro	Varsel:Small R-square value
Pedestrian Crossing Physical Fac	Rejected	Nominal	Character	Pedestrian Crossing-Physical Fac	Varsel:Small R-square value
Police Force	Rejected	Interval	Numeric		Varsel:Small R-square value
Road Surface Conditions	Rejected	Nominal	Character		Varsel:Small R-square value
Road Type	Rejected	Nominal	Character		Varsel:Small R-square value, Group v...
Special Conditions at Site	Rejected	Nominal	Character		Varsel:Small R-square value
Speed limit	Input	Interval	Numeric		
Urban or Rural Area	Input	Interval	Numeric		
Weather Conditions	Rejected	Nominal	Character		Varsel:Small R-square value
Year	Rejected	Interval	Numeric		Varsel:Small R-square value
1st Road Class	Rejected	Interval	Numeric		Varsel:Small R-square value
1st Road Number	Rejected	Interval	Numeric		Varsel:Small R-square value
2nd Road Class	Rejected	Interval	Numeric		Varsel:Small R-square value
2nd Road Number	Rejected	Interval	Numeric		Varsel:Small R-square value

Element 4:
Data Exploration Report

report.html

Figure 5: Regression Variables



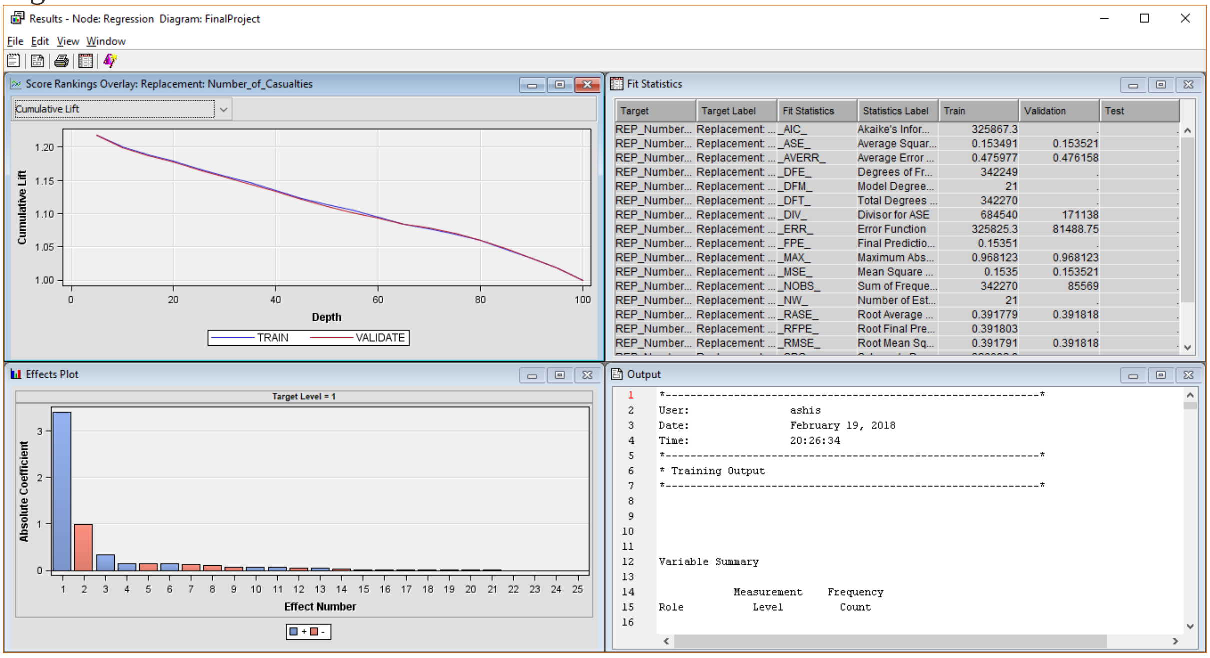
Variables - Reg

(none) ☐ not Equal to

Columns: ☐ Label ☐ Mining ☐ Basic ☐ Statistics

Name	Use	Report	Role	Level
Local_Authority__Highway_	Default	No	Rejected	Nominal
Location_Easting_OSGR	Default	No	Rejected	Interval
Location_Northing_OSGR	Default	No	Rejected	Interval
Longitude	Default	No	Rejected	Interval
Number_of_Casualties	Default	No	Rejected	Nominal
Number_of_Vehicles	Default	No	Rejected	Interval
Pedestrian_Crossing_Human_Contro	Default	No	Rejected	Nominal
Pedestrian_Crossing_Physical_Fac	Default	No	Rejected	Nominal
Police_Force	Default	No	Rejected	Interval
REP_Accident_Severity	Default	No	Input	Interval
REP_Number_of_Casualties	Yes	No	Target	Nominal
REP_Number_of_Vehicles	Default	No	Input	Interval
REP_Speed_limit	Default	No	Input	Interval
Special_Conditions_at_Site	Default	No	Input	Nominal
Speed_limit	Default	No	Rejected	Interval
TI_Carriageway_Hazards1	Default	No	Input	Binary
TI_Light_Conditions1	Default	No	Input	Binary
TI_Light_Conditions2	Default	No	Input	Binary
TI_Light_Conditions3	Default	No	Input	Binary
TI_Light_Conditions4	Default	No	Input	Binary
TI_Road_Surface_Conditions1	Default	No	Input	Binary
TI_Road_Surface_Conditions2	Default	No	Input	Binary
TI_Road_Surface_Conditions3	Default	No	Input	Binary
TI_Road_Type1	Default	No	Input	Binary
TI_Road_Type2	Default	No	Input	Binary

Figure 6:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP_Number...	Replacement: ...	_AIC_	Akaike's Infor...	325867.3	.	.
REP_Number...	Replacement: ...	_ASE_	Average Squar...	0.153491	0.153521	.
REP_Number...	Replacement: ...	_AVERR_	Average Error ...	0.475977	0.476158	.
REP_Number...	Replacement: ...	_DFE_	Degrees of Fr...	342249	.	.
REP_Number...	Replacement: ...	_DFM_	Model Degree...	21	.	.
REP_Number...	Replacement: ...	_DFT_	Total Degrees ...	342270	.	.
REP_Number...	Replacement: ...	_DIV_	Divisor for ASE	684540	171138	.
REP_Number...	Replacement: ...	_ERR_	Error Function	325825.3	81488.75	.
REP_Number...	Replacement: ...	_FPE_	Final Predictio...	0.15351	.	.
REP_Number...	Replacement: ...	_MAX_	Maximum Abs...	0.968123	0.968123	.
REP_Number...	Replacement: ...	_MSE_	Mean Square ...	0.1535	0.153521	.
REP_Number...	Replacement: ...	_NOBS_	Sum of Freque...	342270	85569	.
REP_Number...	Replacement: ...	_NW_	Number of Est...	21	.	.
REP_Number...	Replacement: ...	_RASE_	Root Average ...	0.391779	0.391818	.
REP_Number...	Replacement: ...	_RFPE_	Root Final Pre...	0.391803	.	.
REP_Number...	Replacement: ...	_RMSE_	Root Mean Sq...	0.391791	0.391818	.

```
*------------------------------------------------------------*
User:                    ashis
Date:                    February 19, 2018
Time:                    20:26:34
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement      Frequency
Role          Level           Count
```

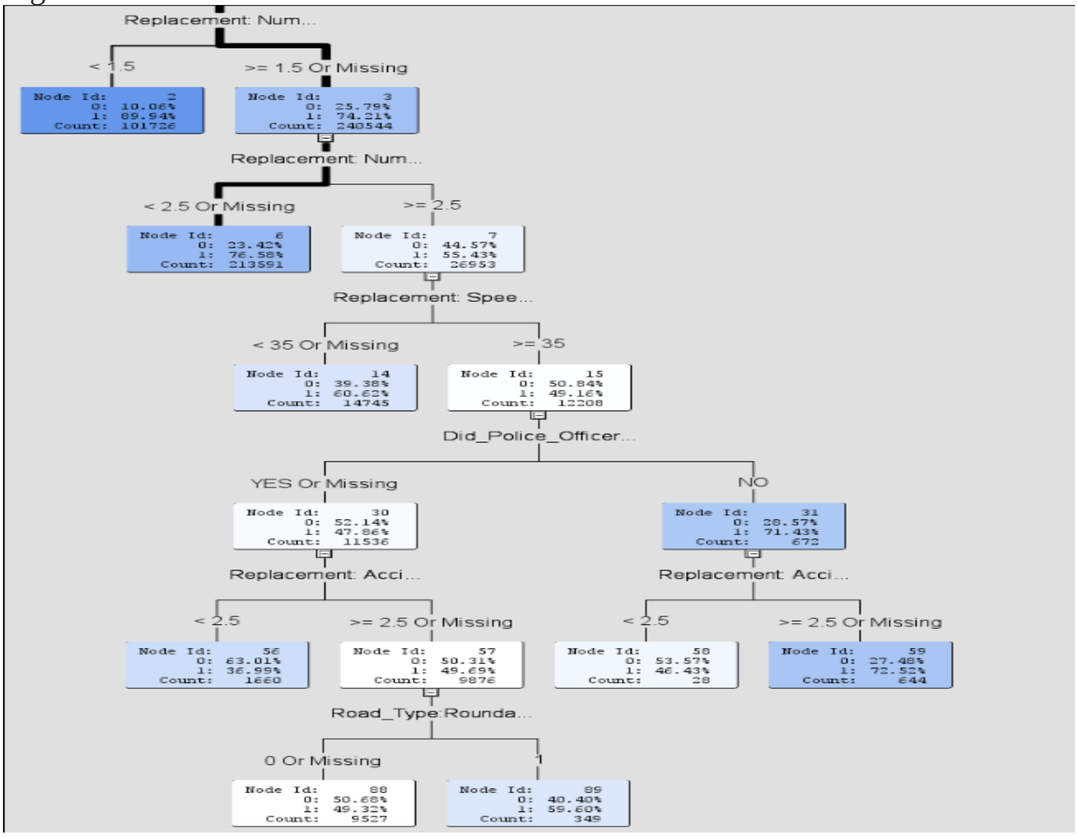
Figure 7: Model Specification

```
                              Analysis of Maximum Likelihood Estimates

                                                                  REP_Number_   Standardized
Parameter                                                         of_Casualties      Estimate Exp(Est)

Intercept                                                         1                            30.352
Did_Police_Officer_Attend_Scene_ No                               1                             1.388
REP_Accident_Severity                                             1                 0.0267      1.147
REP_Number_of_Vehicles                                            1                -0.3120      0.372
REP_Speed_limit                                                   1                -0.0918      0.988
Special_Conditions_at_Site       None                             1                             0.978
TI_Light_Conditions1             0                                1                             1.155
TI_Light_Conditions2             0                                1                             1.068
TI_Light_Conditions3             0                                1                             1.065
TI_Light_Conditions4             0                                1                      .      .
TI_Road_Surface_Conditions1      0                                1                             0.946
TI_Road_Surface_Conditions2      0                                1                             0.997
TI_Road_Surface_Conditions3      0                                1                      .      .
TI_Road_Type1                    0                                1                             1.041
TI_Road_Type2                    0                                1                             0.888
TI_Road_Type3                    0                                1                             0.894
TI_Road_Type4                    0                                1                             1.015
TI_Road_Type5                    0                                1                      .      .
TI_Urban_or_Rural_Area1          0                                1                             0.869
TI_Urban_or_Rural_Area2          0                                1                      .      .
Weather_Conditions               Fine with high winds             1                             0.936
Weather_Conditions               Fine without high winds          1                             0.991
Weather_Conditions               Other                            1                             0.988
Weather_Conditions               Raining with high winds          1                             0.997
Weather_Conditions               Raining without high winds       1                             0.993
```

Figure 8:

```
Fit Statistics

Target=REP_Number_of_Casualties Target Label=Replacement: Number_of_Casualties

   Fit
Statistics    Statistics Label                      Train    Validation

 _AIC_        Akaike's Information Criterion    325867.29          .
 _ASE_        Average Squared Error                  0.15        0.15
 _AVERR_      Average Error Function                 0.48        0.48
 _DFE_        Degrees of Freedom for Error      342249.00          .
 _DFM_        Model Degrees of Freedom              21.00          .
 _DFT_        Total Degrees of Freedom          342270.00          .
 _DIV_        Divisor for ASE                   684540.00   171138.00
 _ERR_        Error Function                    325825.29    81488.75
 _FPE_        Final Prediction Error                 0.15          .
 _MAX_        Maximum Absolute Error                 0.97        0.97
 _MSE_        Mean Square Error                      0.15        0.15
 _NOBS_       Sum of Frequencies                342270.00    85569.00
 _NW_         Number of Estimate Weights            21.00          .
 _RASE_       Root Average Sum of Squares            0.39        0.39
 _RFPE_       Root Final Prediction Error            0.39          .
 _RMSE_       Root Mean Squared Error                0.39        0.39
 _SBC_        Schwarz's Bayesian Criterion      326092.90          .
 _SSE_        Sum of Squared Errors             105070.71    26273.32
 _SUMW_       Sum of Case Weights Times Freq    684540.00   171138.00
 _MISC_       Misclassification Rate                 0.21        0.21
```

Figure 9:



Replacement: Num...
< 1.5
>= 1.5 Or Missing
Node Id: 2 0: 10.06% 1: 89.94% Count: 101726
Node Id: 3 0: 25.79% 1: 74.21% Count: 240544
Replacement: Num...
< 2.5 Or Missing
>= 2.5
Node Id: 6 0: 23.42% 1: 76.58% Count: 213591
Node Id: 7 0: 44.57% 1: 55.43% Count: 26953
Replacement: Spee...
< 35 Or Missing
>= 35
Node Id: 14 0: 39.38% 1: 60.62% Count: 14745
Node Id: 15 0: 50.84% 1: 49.16% Count: 12208
Did_Police_Officer...
YES Or Missing
NO
Node Id: 30 0: 52.14% 1: 47.86% Count: 11536
Node Id: 31 0: 28.57% 1: 71.43% Count: 672
Replacement: Acci...
Replacement: Acci...
< 2.5
>= 2.5 Or Missing
< 2.5
>= 2.5 Or Missing
Node Id: 56 0: 63.01% 1: 36.99% Count: 1660
Node Id: 57 0: 50.31% 1: 49.69% Count: 9876
Node Id: 58 0: 53.57% 1: 46.43% Count: 28
Node Id: 59 0: 27.48% 1: 72.52% Count: 644
Road_Type:Rounda...
0 Or Missing
1
Node Id: 88 0: 50.68% 1: 49.32% Count: 9527
Node Id: 89 0: 40.40% 1: 59.60% Count: 349

Figure 10:

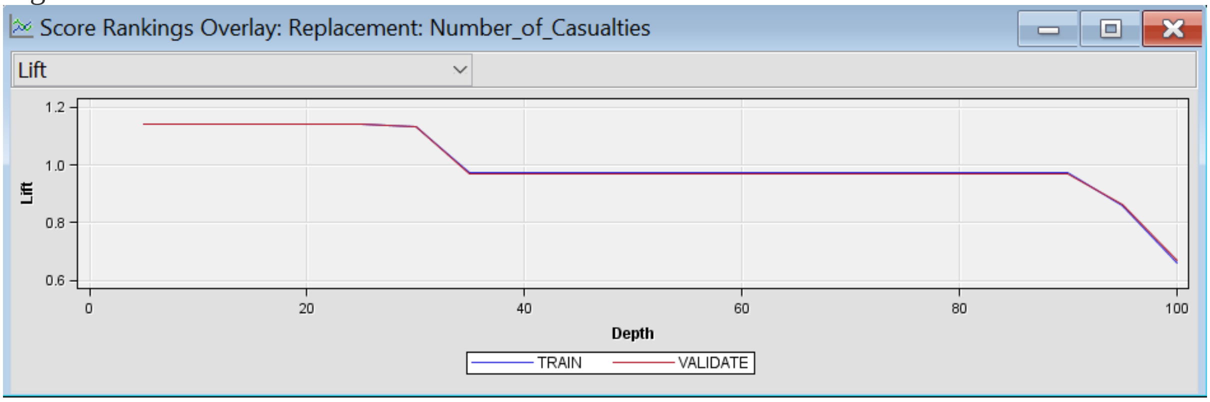
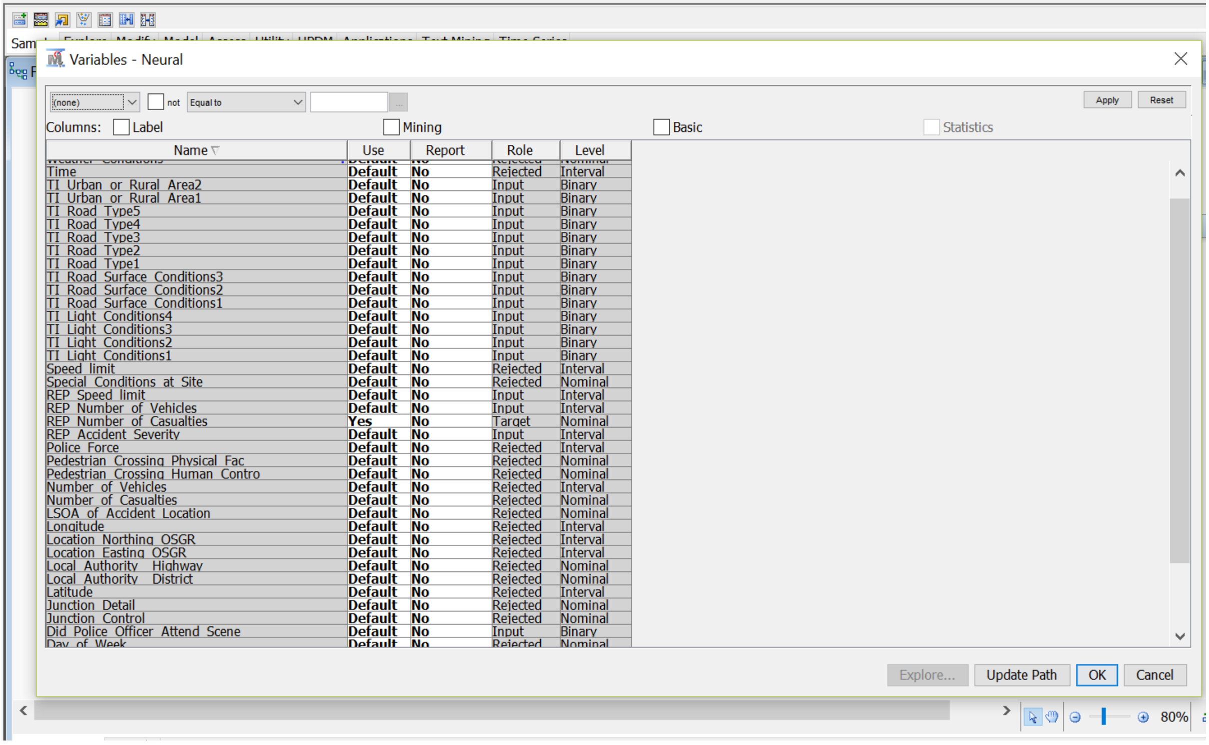
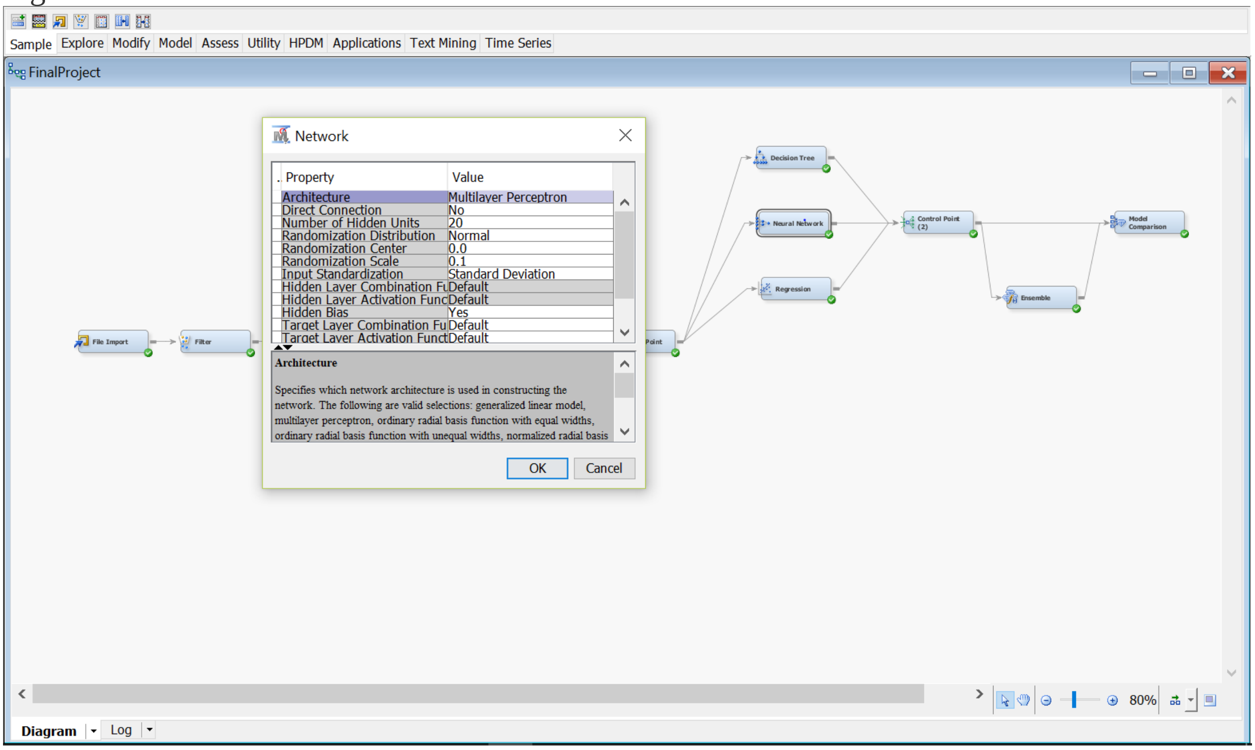


Figure 11:



Name	Use	Report	Role	Level
Weather Conditions	Default	No	Rejected	Nominal
Time	Default	No	Rejected	Interval
TI Urban or Rural Area2	Default	No	Input	Binary
TI Urban or Rural Area1	Default	No	Input	Binary
TI Road Type5	Default	No	Input	Binary
TI Road Type4	Default	No	Input	Binary
TI Road Type3	Default	No	Input	Binary
TI Road Type2	Default	No	Input	Binary
TI Road Type1	Default	No	Input	Binary
TI Road Surface Conditions3	Default	No	Input	Binary
TI Road Surface Conditions2	Default	No	Input	Binary
TI Road Surface Conditions1	Default	No	Input	Binary
TI Light Conditions4	Default	No	Input	Binary
TI Light Conditions3	Default	No	Input	Binary
TI Light Conditions2	Default	No	Input	Binary
TI Light Conditions1	Default	No	Input	Binary
Speed limit	Default	No	Rejected	Interval
Special Conditions at Site	Default	No	Rejected	Nominal
REP Speed limit	Default	No	Input	Interval
REP Number of Vehicles	Default	No	Input	Interval
REP Number of Casualties	Yes	No	Target	Nominal
REP Accident Severity	Default	No	Input	Interval
Police Force	Default	No	Rejected	Interval
Pedestrian Crossing Physical Fac	Default	No	Rejected	Nominal
Pedestrian Crossing Human Contro	Default	No	Rejected	Nominal
Number of Vehicles	Default	No	Rejected	Interval
Number of Casualties	Default	No	Rejected	Nominal
LSOA of Accident Location	Default	No	Rejected	Nominal
Longitude	Default	No	Rejected	Interval
Location Northing OSGR	Default	No	Rejected	Interval
Location Easting OSGR	Default	No	Rejected	Interval
Local Authority Highway	Default	No	Rejected	Nominal
Local Authority District	Default	No	Rejected	Nominal
Latitude	Default	No	Rejected	Interval
Junction Detail	Default	No	Rejected	Nominal
Junction Control	Default	No	Rejected	Nominal
Did Police Officer Attend Scene	Default	No	Input	Binary
Day of Week	Default	No	Rejected	Nominal

Figure 12:



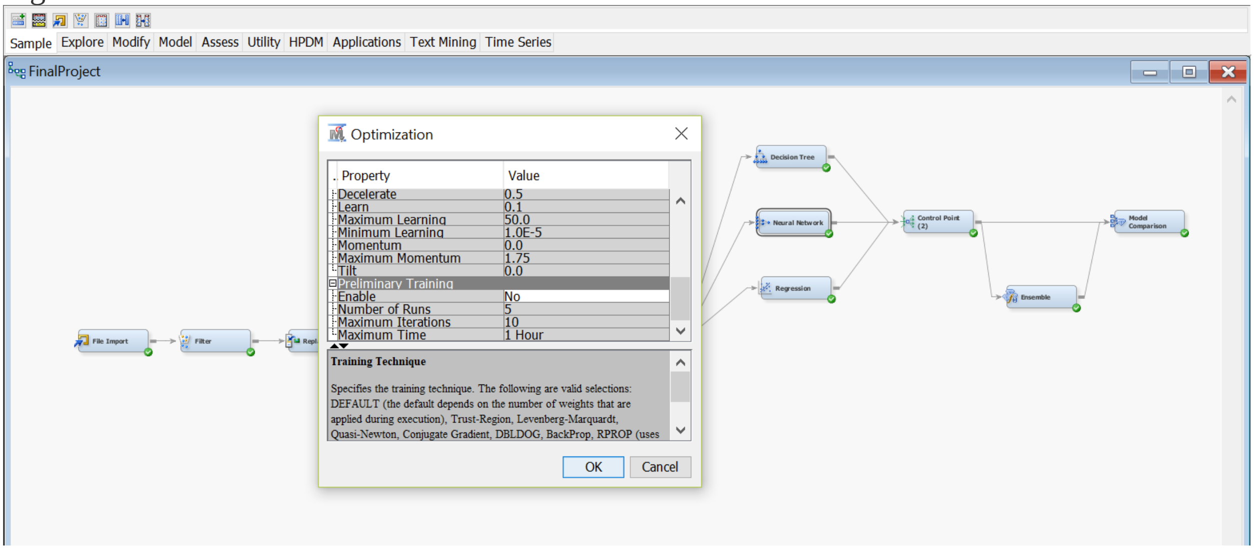
Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	20
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Fu	Default
Hidden Layer Activation Func	Default
Hidden Bias	Yes
Target Layer Combination Fu	Default
Target Layer Activation Funct	Default

Architecture

Specifies which network architecture is used in constructing the network. The following are valid selections: generalized linear model, multilayer perceptron, ordinary radial basis function with equal widths, ordinary radial basis function with unequal widths, normalized radial basis

Figure 13:



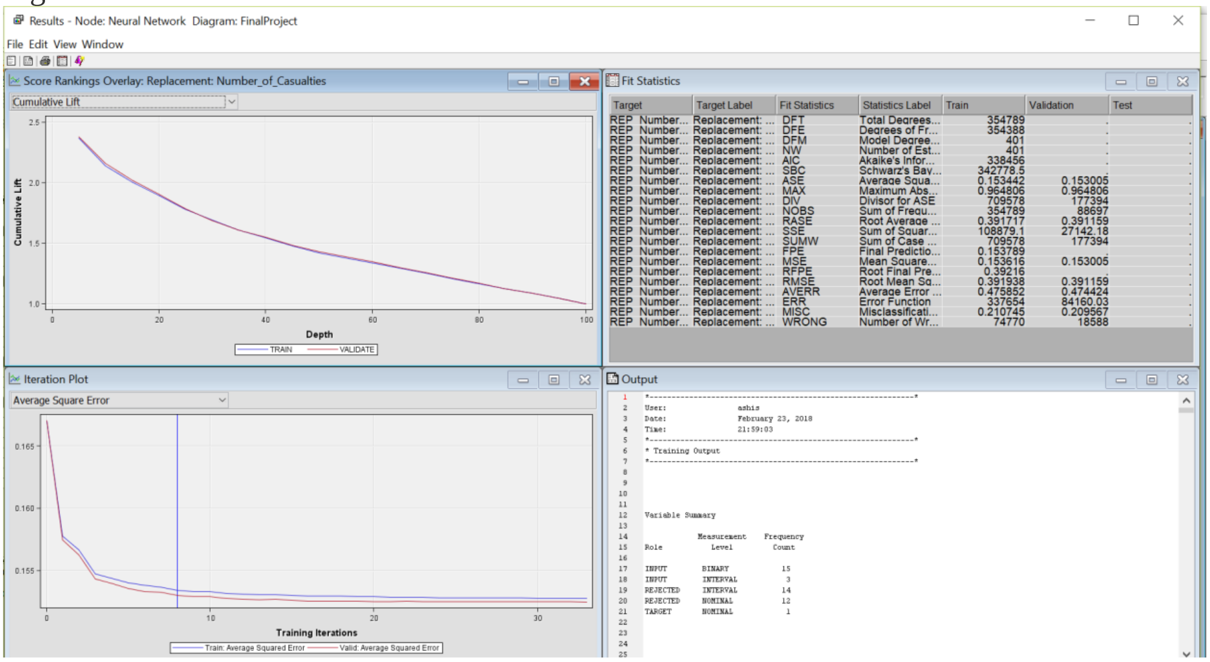
Optimization

Property	Value
Decelerate	0.5
Learn	0.1
Maximum Learning	50.0
Minimum Learning	1.0E-5
Momentum	0.0
Maximum Momentum	1.75
Tilt	0.0
⊟ Preliminary Training	
Enable	No
Number of Runs	5
Maximum Iterations	10
Maximum Time	1 Hour

Training Technique

Specifies the training technique. The following are valid selections: DEFAULT (the default depends on the number of weights that are applied during execution), Trust-Region, Levenberg-Marquardt, Quasi-Newton, Conjugate Gradient, DBLDOG, BackProp, RPROP (uses

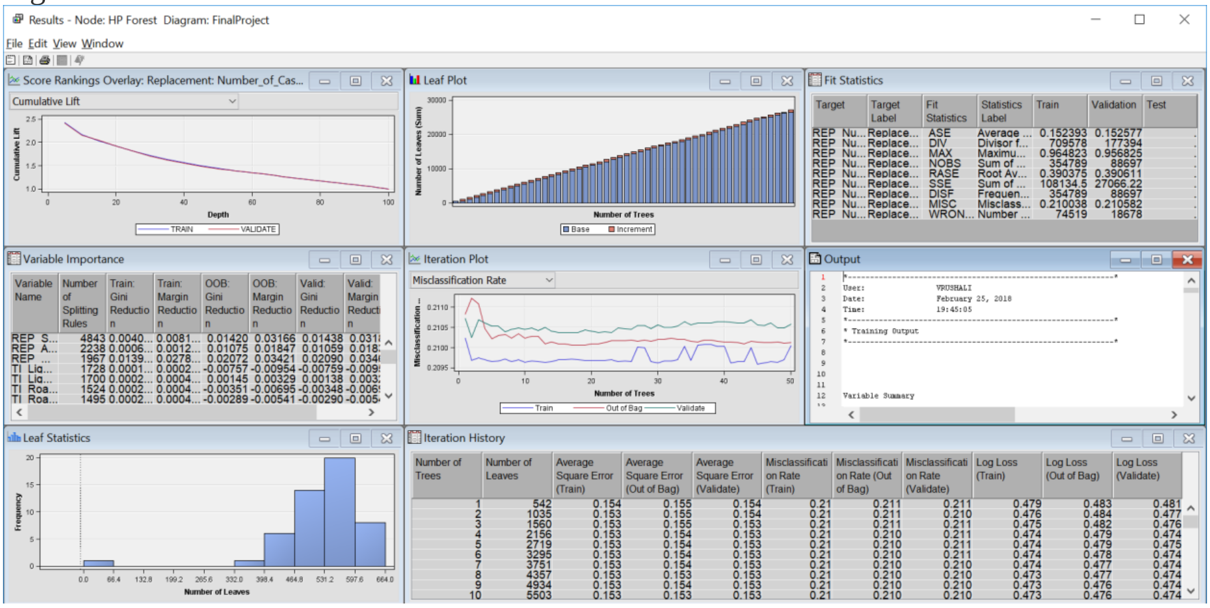
Figure 14:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP Number...	Replacement: ...	DFT	Total Degrees...	354789	.	.
REP Number...	Replacement: ...	DFE	Degrees of Fr...	354388	.	.
REP Number...	Replacement: ...	DFM	Model Degree...	401	.	.
REP Number...	Replacement: ...	NW	Number of Est...	401	.	.
REP Number...	Replacement: ...	AIC	Akaike's Infor...	338456	.	.
REP Number...	Replacement: ...	SBC	Schwarz's Bay...	342778.5	.	.
REP Number...	Replacement: ...	ASE	Average Squa...	0.153442	0.153005	.
REP Number...	Replacement: ...	MAX	Maximum Abs...	0.964806	0.964806	.
REP Number...	Replacement: ...	DIV	Divisor for ASE	709578	177394	.
REP Number...	Replacement: ...	NOBS	Sum of Frequ...	354789	88697	.
REP Number...	Replacement: ...	RASE	Root Average ...	0.391717	0.391159	.
REP Number...	Replacement: ...	SSE	Sum of Squar...	108879.1	27142.18	.
REP Number...	Replacement: ...	SUMW	Sum of Case ...	709578	177394	.
REP Number...	Replacement: ...	FPE	Final Predictio...	0.153789	.	.
REP Number...	Replacement: ...	MSE	Mean Square...	0.153616	0.153005	.
REP Number...	Replacement: ...	RFPE	Root Final Pre...	0.39216	.	.
REP Number...	Replacement: ...	RMSE	Root Mean Sq...	0.391938	0.391159	.
REP Number...	Replacement: ...	AVERR	Average Error ...	0.475852	0.474424	.
REP Number...	Replacement: ...	ERR	Error Function	337654	84160.03	.
REP Number...	Replacement: ...	MISC	Misclassificati...	0.210745	0.209567	.
REP Number...	Replacement: ...	WRONG	Number of Wr...	74770	18588	.

```
*------------------------------------------------------------*
User:                ashis
Date:                February 23, 2018
Time:                21:59:03
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

             Measurement    Frequency
Role            Level         Count

INPUT         BINARY           15
INPUT         INTERVAL          3
REJECTED      INTERVAL         14
REJECTED      NOMINAL          12
TARGET        NOMINAL           1
```

Figure 15:



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
REP Nu...	Replace...	ASE	Average ...	0.152393	0.152577	.
REP Nu...	Replace...	DIV	Divisor f...	709578	177394	.
REP Nu...	Replace...	MAX	Maximu...	0.964823	0.956825	.
REP Nu...	Replace...	NOBS	Sum of ...	354789	88697	.
REP Nu...	Replace...	RASE	Root Av...	0.390375	0.390611	.
REP Nu...	Replace...	SSE	Sum of ...	108134.5	27066.22	.
REP Nu...	Replace...	DISF	Frequen...	354789	88697	.
REP Nu...	Replace...	MISC	Misclass...	0.210038	0.210582	.
REP Nu...	Replace...	WRON...	Number ...	74519	18678	.

Variable Name	Number of Splitting Rules	Train: Gini Reductio n	Train: Margin Reductio n	OOB: Gini Reductio n	OOB: Margin Reductio n	Valid: Gini Reductio n	Valid: Margin Reducti n
REP S...	4843	0.0040...	0.0081...	0.01420	0.03166	0.01438	0.031
REP A...	2238	0.0006...	0.0012...	0.01075	0.01847	0.01059	0.018
REP ...	1967	0.0139...	0.0278...	0.02072	0.03421	0.02090	0.034
TI Liq...	1728	0.0001...	0.0002...	-0.00757	-0.00954	-0.00759	-0.009
TI Liq...	1700	0.0002...	0.0004...	0.00145	0.00329	0.00138	0.003
TI Roa...	1524	0.0002...	0.0004...	-0.00351	-0.00695	-0.00348	-0.006
TI Roa...	1495	0.0002...	0.0004...	-0.00289	-0.00541	-0.00290	-0.005

```
*------------------------------------------------------------*
User:                VRUSHALI
Date:                February 25, 2018
Time:                19:45:05
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary
```

Number of Trees	Number of Leaves	Average Square Error (Train)	Average Square Error (Out of Bag)	Average Square Error (Validate)	Misclassificati on Rate (Train)	Misclassificati on Rate (Out of Bag)	Misclassificati on Rate (Validate)	Log Loss (Train)	Log Loss (Out of Bag)	Log Loss (Validate)
1	542	0.154	0.155	0.154	0.21	0.211	0.211	0.479	0.483	0.481
2	1035	0.153	0.155	0.154	0.21	0.211	0.210	0.476	0.484	0.477
3	1560	0.153	0.155	0.153	0.21	0.211	0.211	0.475	0.482	0.476
4	2156	0.153	0.154	0.153	0.21	0.210	0.211	0.474	0.479	0.474
5	2719	0.153	0.154	0.153	0.21	0.210	0.211	0.474	0.479	0.475
6	3295	0.153	0.154	0.153	0.21	0.210	0.211	0.474	0.478	0.474
7	3751	0.153	0.154	0.153	0.21	0.210	0.210	0.474	0.477	0.474
8	4357	0.153	0.153	0.153	0.21	0.210	0.210	0.473	0.477	0.474
9	4934	0.153	0.154	0.153	0.21	0.210	0.210	0.473	0.476	0.474
10	5503	0.153	0.153	0.153	0.21	0.210	0.210	0.473	0.476	0.474