PREDICTIVE ANALYSIS OF ACCIDENT SEVERITY

CONTENTS

TABLE OF FIGURES

INTRODUCTION

Accidents can be described as unintentional incidents that result in injury, damage to property or death. In the United Kingdom, the Department for Transport (DFT) provides a legal definition of a road traffic accident in the “Road Traffic Act” 1988. According to the Act, a road traffic accident is "an occurrence arising out of the use of a motor vehicle on a road or other public place which causes death or injury to any person." It can occur due to certain factors like human error, malfunctioning vehicles, or even environmental conditions such as weather conditions or road conditions.

The severity of an accident varies depending on the type of accident, the number of people involved and the extent of damage. In our prediction, the severity of accident is classified into three:

- Fatal: These are the most serious accidents, which can result in death, permanent disability, and permanent property destruction.
- Serious: They are not as catastrophic as fatal accidents but results in severe injuries like broken bones, damage to properties.
- Slight: These accidents involve minor damage to property, such as bumper dents. They might usually involve minor injuries like cuts, bruises, and sprains.

BUSINESS CONTEXT

One of the commercial challenges that insurance firms have in terms of accident severity is effectively forecasting the level of severity of an accident. This is significant because it allows insurance firms to charge clients appropriate premiums based on their accident severity. Inaccurate forecasting can result in undercharging, resulting in losses for the insurance company. Overcharging clients, on the other hand, may result in losing the customers to the competitors. As a result, building a credible model for predicting accident severity can represent a big revenue opportunity for an insurance firm.

BUSINESS OBJECTIVE

The accurate prediction of accident severity considering input variables like location, vehicle type, weather conditions, and so on will provide valuable insight for the insurance company to make informed decisions on claim settlement and pricing decisions. This will also help the company to reduce costs and increase revenue.

This analysis will be carried out on a classification model because the target variable "accident severity" is categorical. We will be using some machine learning algorithm to learn and evaluate the likely severity of an accident at a particular time.

FEATURE SELECTION

Feature selection is the process of selecting a subset of relevant features or variables from a larger set of features in order to build a predictive model. The features for this model have been selected in the group assignment, there were selected based on industry knowledge and weight of correlation of these variables to the target variable.

Name	Role	Level
accident_severity	**Target**	**Ordinal**
age_of_driver	**Input**	**Interval**
number_of_casualties	**Input**	**Interval**
number_of_vehicles	**Input**	**Interval**
pedestrian_crossing_human_contro	**Input**	**Nominal**
road_surface_conditions	**Input**	**Nominal**
road_type	**Input**	**Nominal**
sex_of_driver	**Input**	**Nominal**
speed_limit	**Input**	**Interval**
weather_conditions	**Input**	**Nominal**

Figure 1: Variable Selection

BASELINE METHOD

A baseline method refers to a simple, intuitive approach that serves as a point of comparison for more complex machine learning models. The baseline method is usually used as a benchmark to assess the performance of the more sophisticated models. For this analysis, the baseline that will be utilized is the 'Decision Tree'. We will use the default setting of the Decision Tree model as a baseline and the performance of this method will be measured with the misclassification rate because the target variable is categorical.

Baseline statistics

Statistics Label	Train	Validation
Sum of Frequencies	7224	1810
Misclassification Rate	0.504014	0.504972
Maximum Absolute Error	0.964882	0.964882
Sum of Squared Errors	4002.651	1007.667
Average Squared Error	0.184692	0.185574
Root Average Squared Error	0.429758	0.430783
Divisor for ASE	21672	5430

Figure 2: Baseline model statistics

The accuracy of this model is measured as '1- misclassification rate'

Training accuracy: 1 – 0.504014 = 0.495986
Validation accuracy: 1 – 0.504972 = 0.495028

The training and testing set are approximately 50% accurate, this will be the benchmark for developing the more sophisticated models for this prediction.

HYPERPARAMETER TUNING

Hyperparameter tuning is the process of selecting the optimal values for the hyper-parameters of a predictive model. Hyperparameters are parameters that are set before the model is trained and can significantly impact the performance and accuracy of the model. Examples of hyperparameters include the number of hidden layers, learning rate, regularization strength, and number of trees in a random forest model. Sarma K. (2017).

We will consider tuning the parameters in the decision tree model to further improve our baseline model and then look at other machine learning algorithms too, the dataset has been preprocessed in the group aspect of the assignment.

DECISION TREE

The decision tree algorithm learns a tree-like structure from the input data, where each internal node represents a test on one of the input features, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value. The splitting criterion is chosen to maximize the information gain or the reduction in impurity at each node. The hyperparameters to be tuned in this model includes maximum tree depth, splitting rule, minimum observation per leaf and pruning method amongst others.

<u>Decision Tree 1</u>

This is an adjustment of the baseline model, the following hyperparameters were tuned:

- Maximum depth – 8
- Perform cross validation – No
- Assessment – Misclassification rate

Property	Value
Use Input Once	No
Maximum Branch	2
Maximum Depth	8
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1

Figure 3: Decision Tree 1 hyperparameter

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.497785	0.501105
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.967367	1
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3978.423	1010.094
accident_severity	accident_severity	_ASE_	Average Squared Error	0.183574	0.186021
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.428456	0.431302
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 4: Decision Tree 1 statistics

The model made about less than 1% improvement from the baseline.

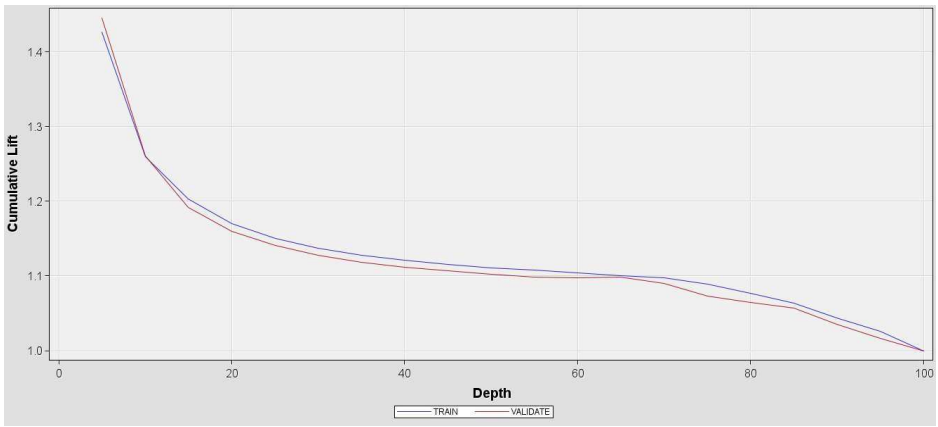


Figure 5: Decision Tree 1 cumulative lift

The cumulative lift pattern may suggest that the model is overfitting to a particular subset of the population, or that there are specific features or characteristics that are more predictive of the positive cases than others.

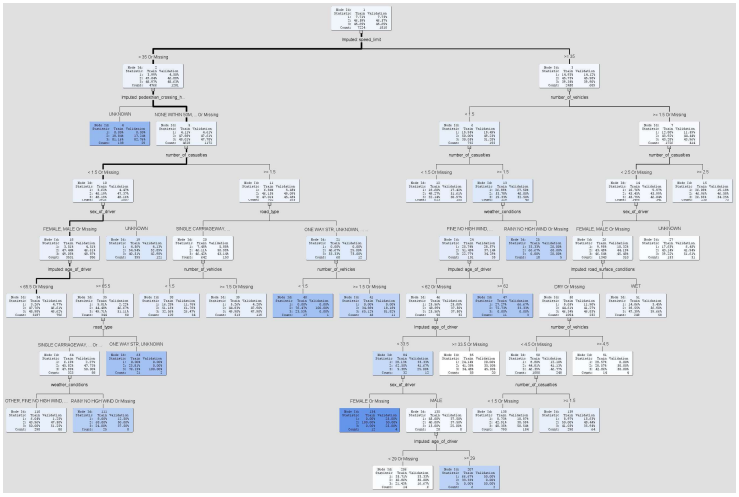


Figure 6: Decision Tree 1

The root node depicts that there is 46% chance for the accident to be serious or slight as the two categories have almost same weight. The fatal category has a low prediction of about 7% in the decision tree. Node 134 may suggest that there is a strong relationship between the sex of the driver and the severity of accidents in certain situations. We will carry out additional analysis to understand the underlying reasons this kind of pattern.

Decision Tree 2

To address the class imbalance that could result to bias on the prediction of our model. The under-sampling technique is used to adjust the training dataset, it involves reducing the number of samples in the majority class to balance the class distribution.

The disadvantage of this technique is that the minority class has a small number of instances, thus it could lead to loss of information that may be important for the learning process. It can also lead to overfitting of the model to the training data.

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	accident_severity	TARGET	2	3366	46.5947
TRAIN	accident_severity	TARGET	3	3301	45.6949
TRAIN	accident_severity	TARGET	1	557	7.7104

Figure 7: Target variable distribution

This manipulation is done to improve the performance of the decision tree. The options selected is level based as this will randomly select a subset of the ‘slight’ and ‘serious’ instances to match the number of instances in the ‘fatal’ class.

Stratified	
Criterion	Level Based
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Rarest Level
Level Proportion	100.0
Sample Proportion	30.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	No
Class Targets	Yes
Status	
Create Time	3/29/23 2:57 PM
Run ID	94e432c0-b15c-a24c-b918-2d3d584163
Last Error	
Last Status	Complete

Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
accident_severity	1	1	557	30.0108	accident_severity
accident_severity	2	2	656	35.3448	accident_severity
accident_severity	3	3	643	34.6444	accident_severity

Figure 8: Oversampling technique inputs

There is a slight decline in the performance of the model, as the accuracy of the training and validation set is about 48%, this test proves that the model is not absolutely biased towards the majority class in its predictions, we can continue the hyperparameter tuning with our initial datasets. This result could also be because of the low population of the sample, it is not diverse enough and the margin of error could be high. The subtree assessment plot shows the high misclassification rate of the model.

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	1483	373
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.525287	0.538874
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.907869	0.907869
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	906.2164	236.3999
accident_severity	accident_severity	_ASE_	Average Squared Error	0.20369	0.21126
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.45132	0.45963
accident_severity	accident_severity	_DIV_	Divisor for ASE	4449	1119
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	2966	.

Figure 9: Decision Tree 2 statistics

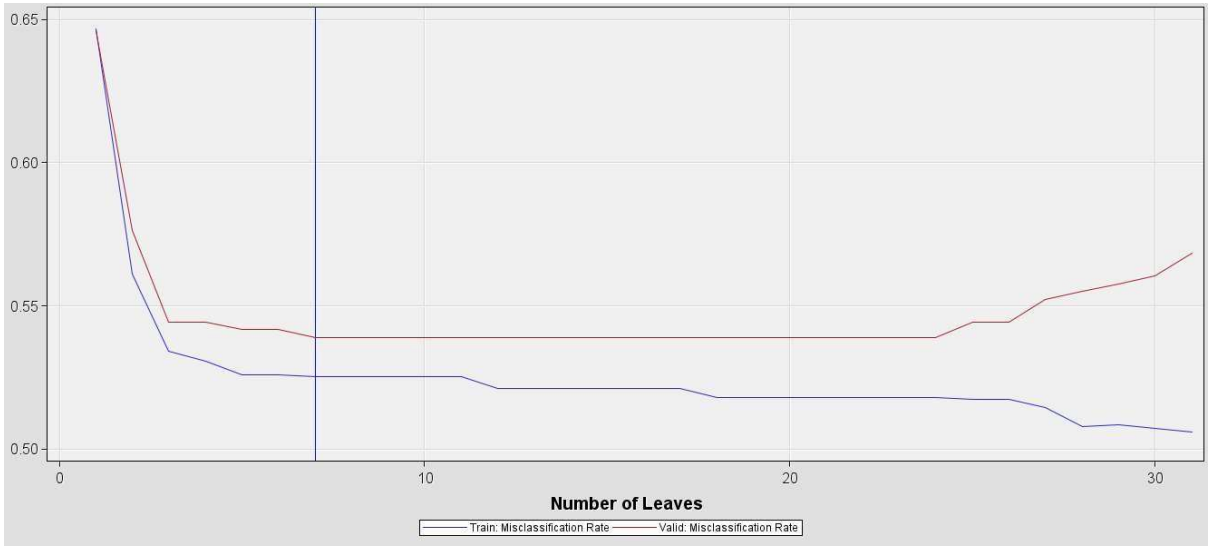


Figure 10: Decision Tree 2 misclassification rate

The decision tree shows a more balanced prediction, as in the root node there is a 30% chance that the accident could be fatal.

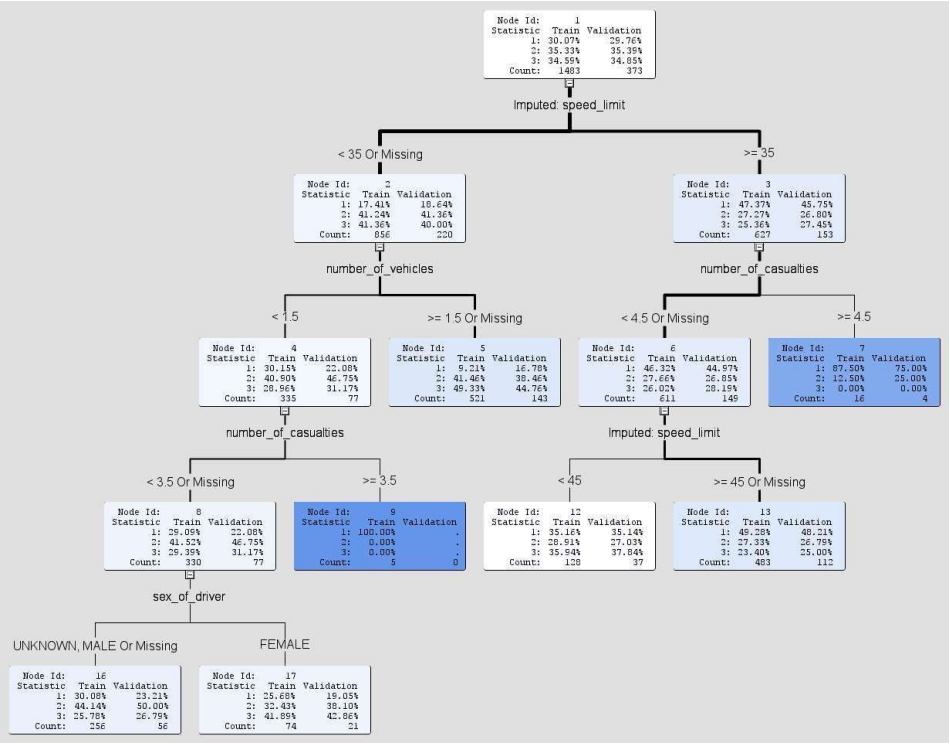


Figure 11: Decision Tree 2

Decision Tree 3

To achieve a better performance the following hyperparameters were tuned using the initial dataset. The stratify sample method was used to reduce bias in the target variable.

- Significance level – 5%
- Maximum depth – 10
- Sample method – stratify
- Number of rules – 4
- Split size – 2

Property	Value
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	Yes
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	User
Sample Method	Stratify
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	

Figure 12: Decision Tree 3 Hyperparameter

The performance of the model improved from the previous one, with an accuracy of about 52% on the training and validation set. This is a little better than the baseline which indicates that the model is learning.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.481174	0.483425
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.991803	1
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3908.39	1017.582
accident_severity	accident_severity	_ASE_	Average Squared Error	0.180343	0.1874
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.424668	0.432897
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 13: Decision Tree 3 statistics

The variable importance shows that the speed limit is the most important predictor for accident severity, the age of the driver is also an important variable.

Variable Importance

Label	Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
I...	IMP_speed_limit	3	1.0000	0.8164	0.8164
I...	IMP_age_of_driver	32	0.9994	0.6125	0.6129
n...	number_of_vehicles	12	0.9912	1.0000	1.0089
I...	IMP_pedestrian_crossing_human_co	1	0.8179	0.7331	0.8963
n...	number_of_casualties	6	0.6990	0.4788	0.6850
w...	weather_conditions	8	0.5437	0.2323	0.4272
s...	sex_of_driver	3	0.5189	0.2397	0.4619
ro...	road_type	9	0.4636	0.4493	0.9692
I...	IMP_road_surface_conditions	3	0.2711	0.1558	0.5744

Figure 14: Decision Tree 3 variable importance

The below shows the variation in the prediction in the validation dataset which indicates that there is a marginal increase in the performance of the tree.

Figure 15: Decision Tree 3 leaf statistics

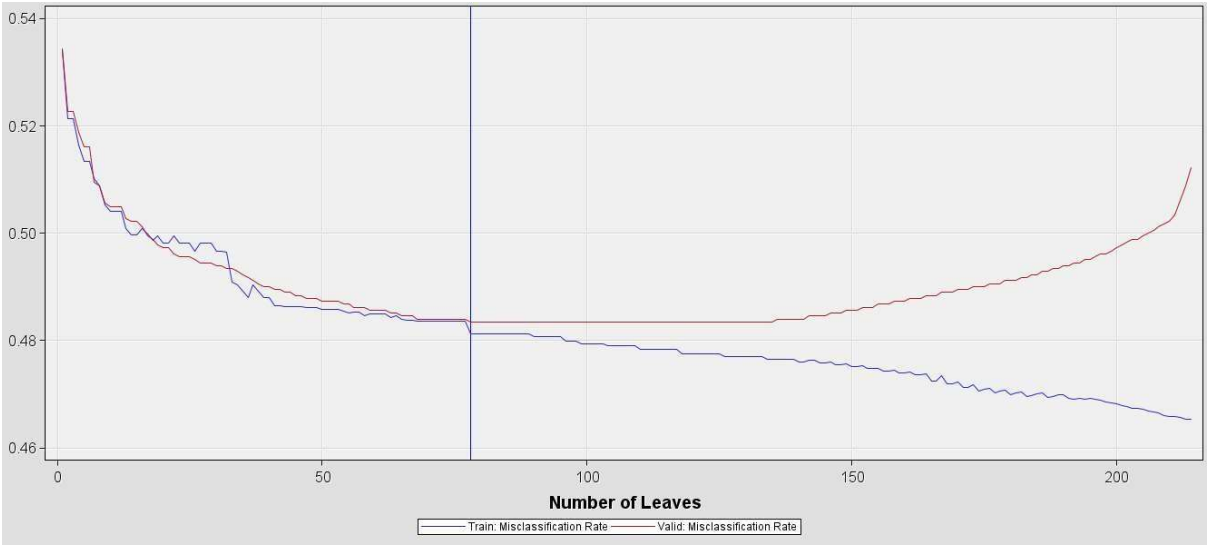


Figure 16: Decision Tree 3 misclassification rate

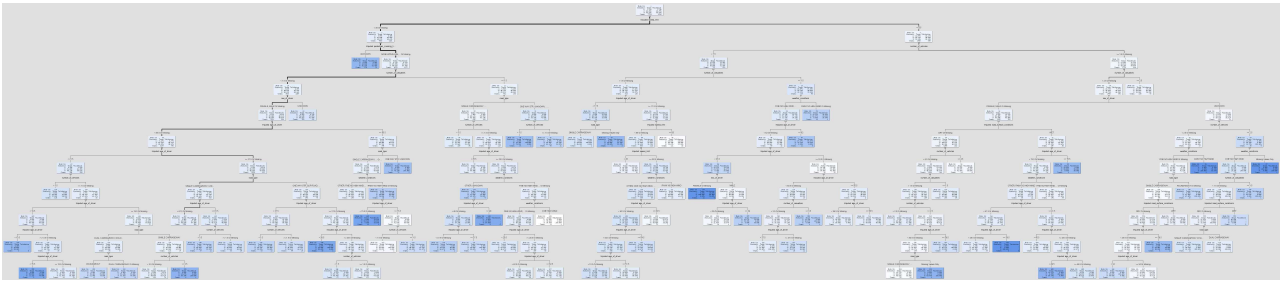


Figure 17: Decision Tree 3

The large number of nodes in the decision tree, is as a result of the number of features selected, and this could indicate a better performance of the model or on the other hand could mean overfitting of the data.

Decision Tree 4

To further improve the model, the following hyperparameters will be tuned.

- Significance level – 5%
- Maximum depth – 5
- Sample method – stratify
- Leaf size - 8
- Split size – 2
- Cross validation – yes

Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.05
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	5
Minimum Categorical Size	5
Node	
Leaf Size	8
Number of Rules	5
Number of Surrogate Rules	0
Split Size	2
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	Yes
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	

Figure 18: Decision Tree 4 Hyperparameter

There is no improvement in the performance of this model, with about 50% accuracy. Although the number of nodes in the decision tree has reduces. It could be that the model is underfitting the data by not capturing all the necessary relationships in the dataset.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.504014	0.504972
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.964882	0.964882
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	4002.651	1007.667
accident_severity	accident_severity	_ASE_	Average Squared Error	0.184692	0.185574
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.429758	0.430783
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 19: Decision Tree 4 statistics

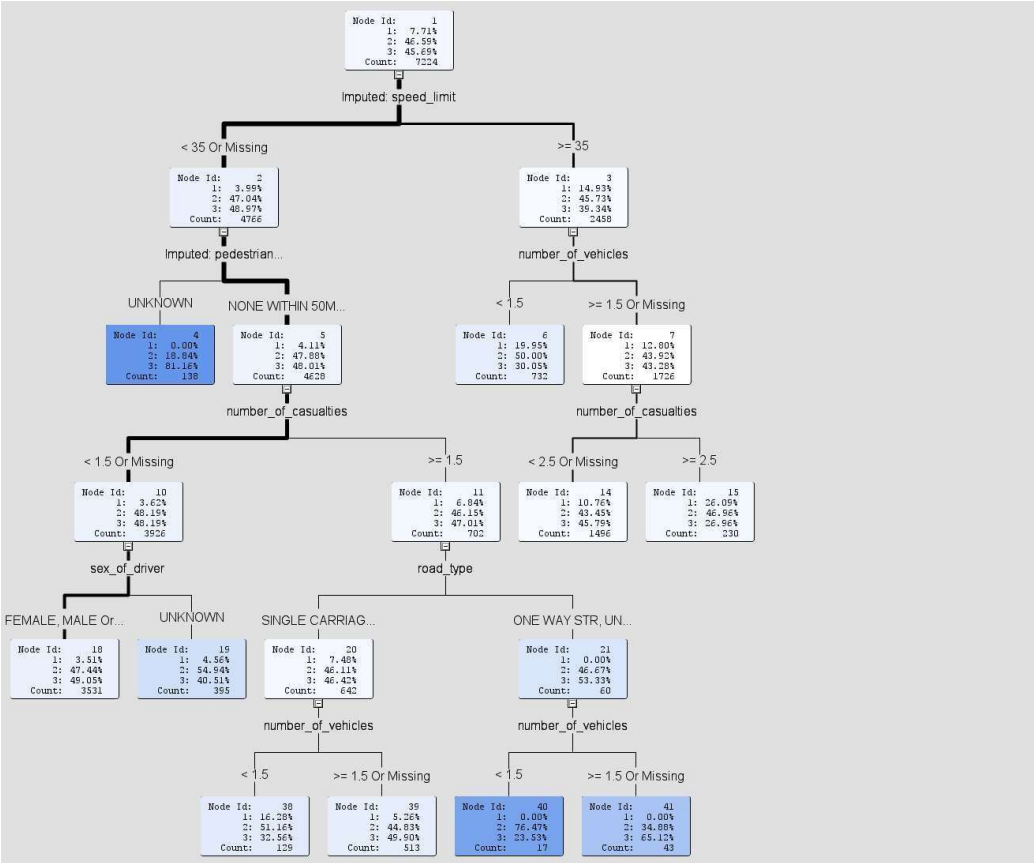


Figure 20: Decision Tree 4

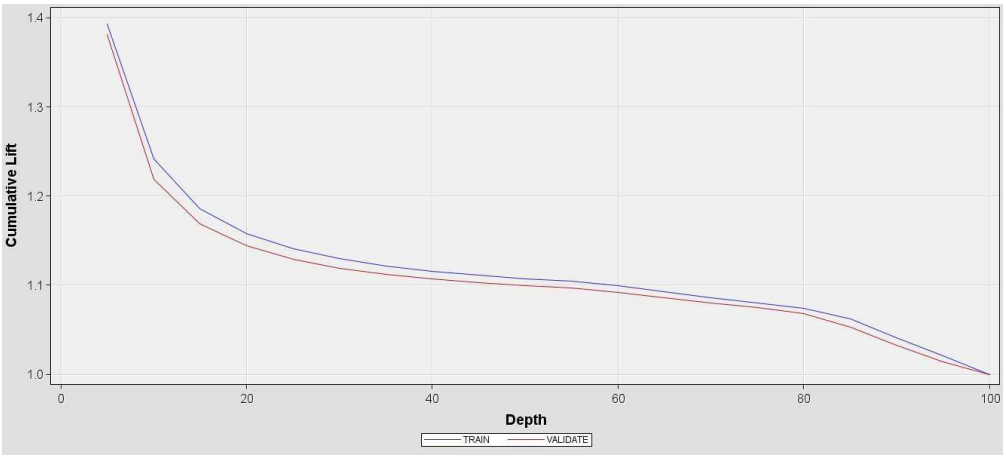


Figure 21: Decision Tree 4 cumulative lift

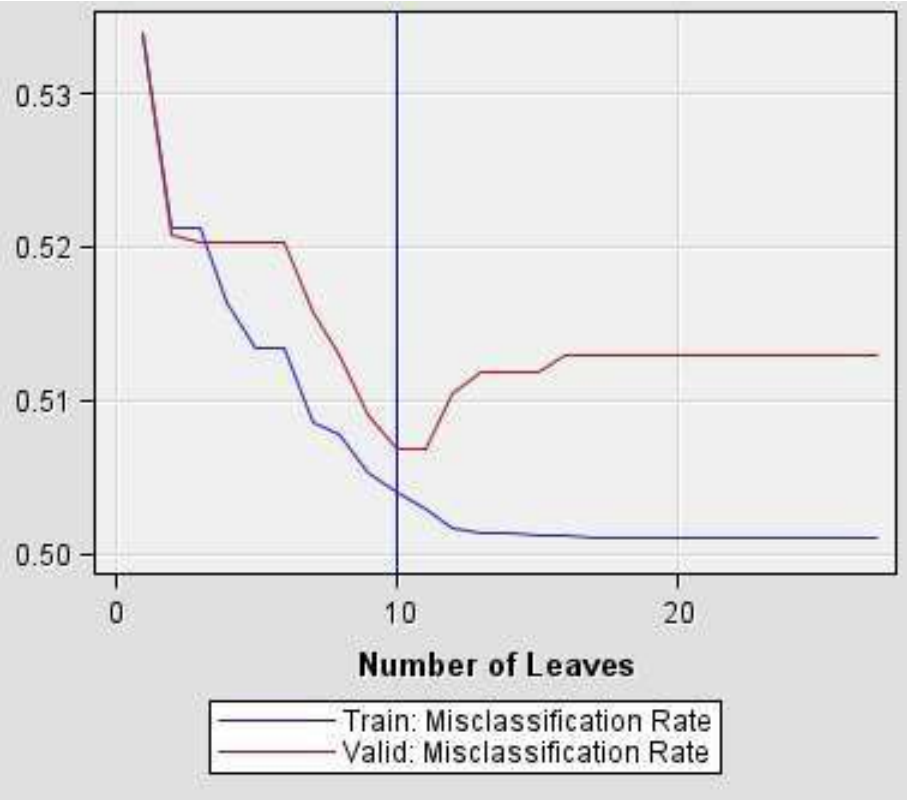


Figure 22: Decision Tree 4 misclassification rate.

The visualization shows the increased misclassification rate of the model.

Decision Tree 5

We will tune the hyperparameters below to improve the performance of the model

- Significance level – 5%
- Maximum depth – 12
- Sample method – default
- Number of rules – 4
- Split size – 3
- Validation – Yes

Property	Value
Maximum Depth	12
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	4
Number of Surrogate Rules	0
Split Size	3
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	Yes
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes

Figure 23: Decision Tree 5 Hyperparameters

There is not much improvement in this model, the misclassification rate reduced below 1%. The number of nodes increased and there is better interpretation of the relationship between variables.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.496539	0.501105
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.964882	0.964882
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3990.04	1008.595
accident_severity	accident_severity	_ASE_	Average Squared Error	0.18411	0.185745
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.429081	0.430981
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 24: Decision Tree 5 statistics

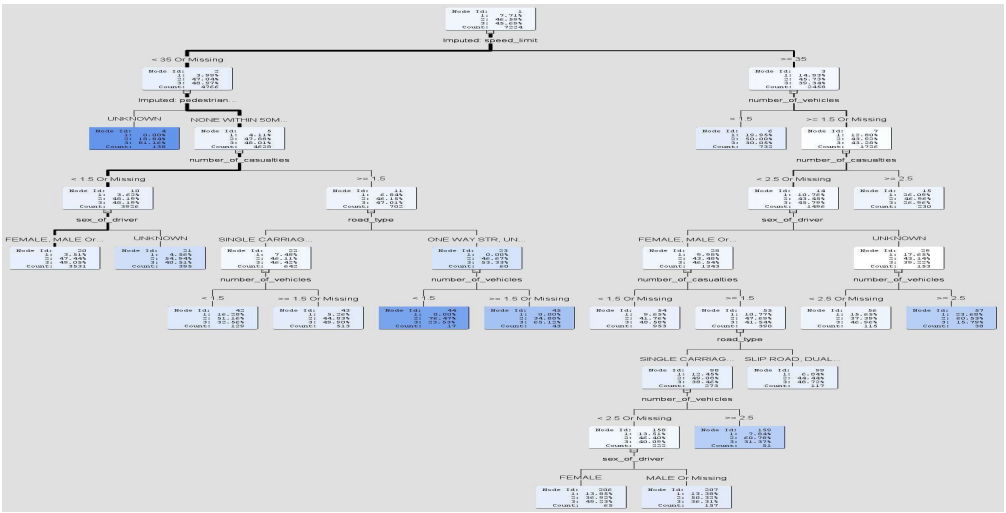


Figure 25: Decision Tree 5

Decision Tree 6

- Significance level – 10% - Maximum depth – 10
- Sample method – default
- Number of rules – 2
- Validation – Yes
- Maximum branch – 4

Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.1
Missing Values	Use in search
Use Input Once	No
Maximum Branch	4
Maximum Depth	10
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	2
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	Yes
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	

Figure 26: Decision Tree 6 Hyperparameter

We tried to simplify the model and increase the branches, the model improved insignificantly, marginally better than the baseline.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.498616	0.503315
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.969697	0.964811
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	4001.401	1014.531
accident_severity	accident_severity	_ASE_	Average Squared Error	0.184635	0.186838
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.429691	0.432248
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 27: Decision Tree 6 statistics

The decision tree has more branches under the root node which indicates that it can capture more complex relationships between variables but may also be more prone to overfitting and less interpretable.

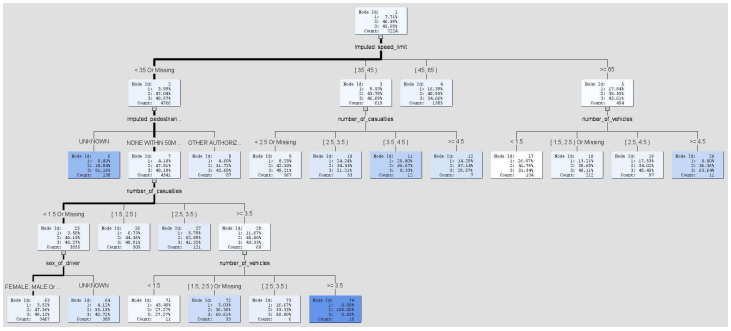
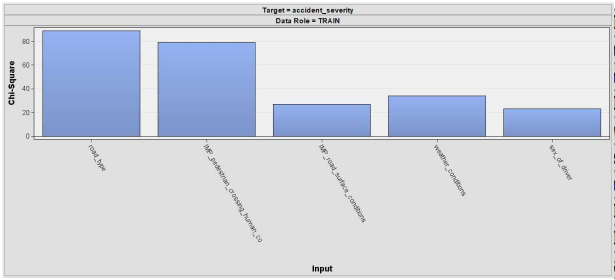


Figure 28: Decision Tree 6

Decision Tree 7

Here we will use the Chi-square statistics to select the top variables and run the model on its default mode, using only the most important predictors to reduce the risk of overfitting to noise in the data.



IMP_age_of_driver	No
number_of_vehicles	No
number_of_casualties	No
IMP_speed_limit	No
weather_conditions	Yes
sex_of_driver	Yes
road_type	Yes
IMP_pedestrian_crossing_hu	Yes
IMP_road_surface_conditions	Yes
accident_severity	Yes

Figure 29: Decision Tree 7 variable ranking and selection

The model's accuracy has a less than 1% improvement from the baseline. It could mean that these top variables may not necessarily be the best for decision tree prediction despite the high correlation with chi-square statistics.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810	.
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.508998	0.516575	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.940356	1	.
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	4054.153	1020.996	.
accident_severity	accident_severity	_ASE_	Average Squared Error	0.187069	0.188029	.
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.432514	0.433623	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430	.
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.	.

Figure 30: Decision Tree 7 statistics

The leaf statistics shows the variation between the predictions of the training sets and validation sets.

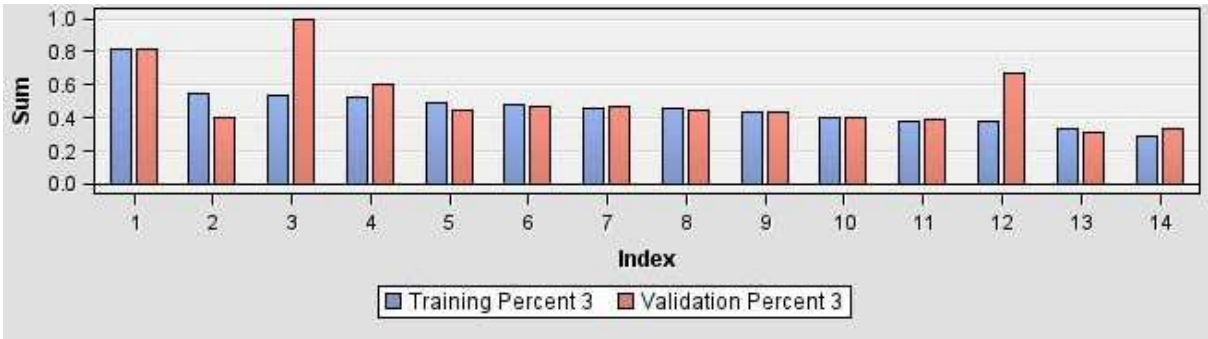


Figure 31: Decision Tree 7 leaf statistics

The table below shows that the pedestrian crossing human condition is the most important variable.

Variable Importance

Variable Name	Label	Number of Splitting Rules	Validation Importance	Importance	Ratio of Validation to Training Importance
IMP_pedestrian_crossing_human_co	Imputed: pedestrian_crossing_human_co...	1	1.0000	1.0000	1.0000
sex_of_driver	sex_of_driver	3	0.4029	0.4520	0.8913
road_type	road_type	6	0.3072	0.4241	0.7243
weather_conditions	weather_conditions	2	0.2953	0.3366	0.8771
IMP_road_surface_conditions	Imputed: road_surface_conditions	1	0.0000	0.3294	0.0000

Figure 32: Decision Tree 7 variable importance

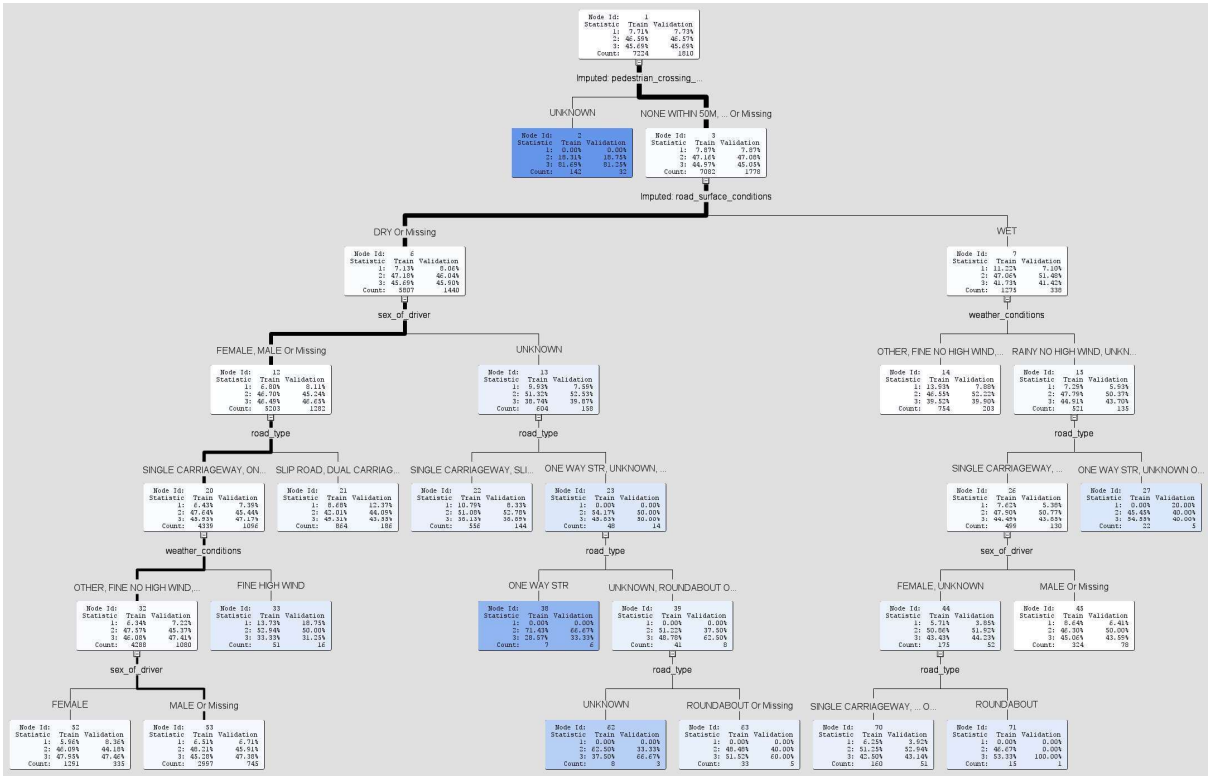


Figure 33: Decision Tree 7

There are more leaves in the three which allows for explanation within variables. The slight and serious severities have a marginal difference in their weights for this model.

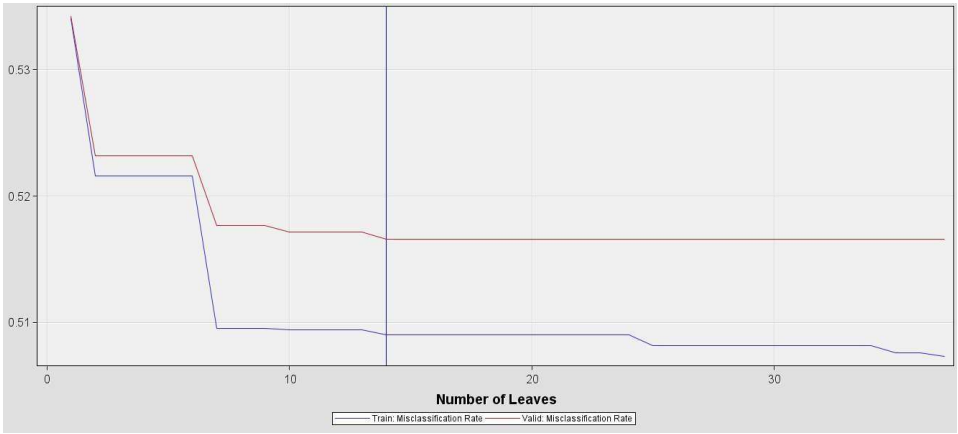


Figure 34: Decision Tree 7 misclassification rate

NEURAL NETWORK

A neural network is a type of machine learning model that is based on the structure and function of the human brain. It is a collection of connected processing nodes, which work together to produce an output given an input. Aggarwal (2019). Neural networks can process large amounts of data quickly and efficiently, making them a practical choice for real-time accident severity prediction.

Neural Network 1

- Number of hidden layers – 3
- Maximum iterations – 50
- Maximum time – 4 hours
- Number of runs – 5

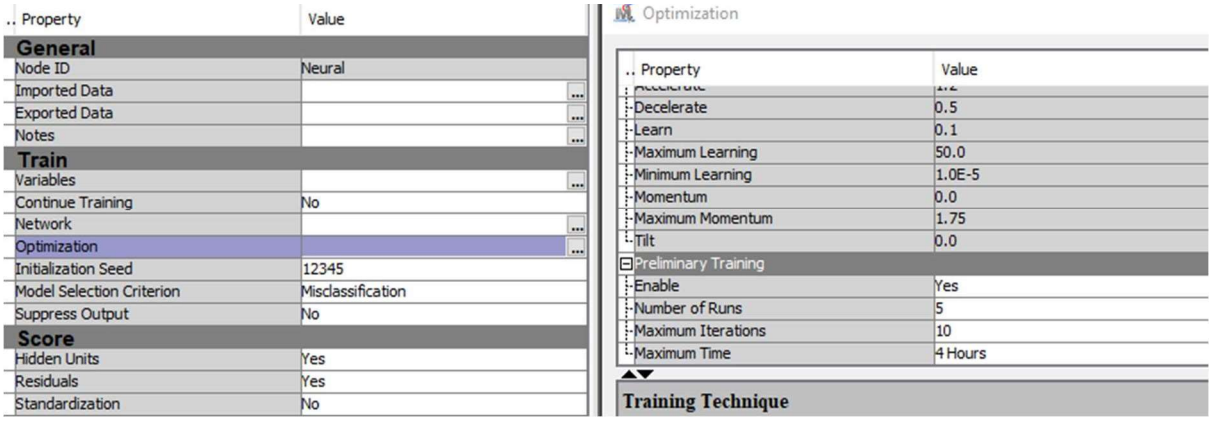


Figure 35: Neural Network 1 hyperparameter

The Misclassification rate indicates a fair improvement of the model, as the accuracy of the validation set has improved to about 55%, which is better than the baseline.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12704.81	.
accident_severity	accident_severity	_AVERR_	Average Error Function	0.58051	0.580826
accident_severity	accident_severity	_ASE_	Average Squared Error	0.182546	0.181561
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14386	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_ERR_	Error Function	12580.81	3153.883
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.184119	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.971536	0.956587
accident_severity	accident_severity	_MSE_	Mean Squared Error	0.183333	0.181561
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.462763	0.445856
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	62	.
accident_severity	accident_severity	_NW_	Number of Estimated Weights	62	.
accident_severity	accident_severity	_WRONG_	Number of Wrong Classifications	3343	807
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.427254	0.426099
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.429091	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.428174	0.426099
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	13174.66	.
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3956.136	985.8746
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 36: Neural network 1 statistics

Neural networks can incorporate disparate sources of data to a single model which allows for a more comprehensive understanding of the factors that contribute to accident severity.

The misclassification rate has shown improvement compared to other models as seen in the graph below.

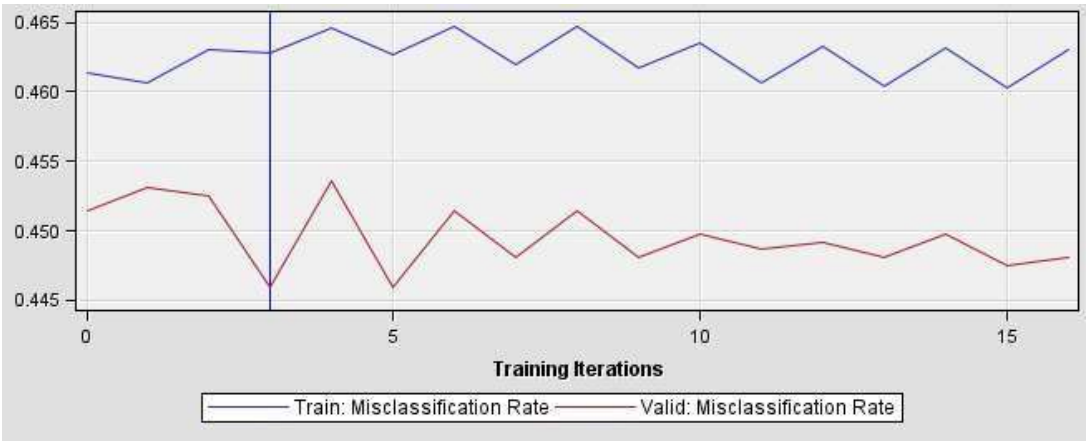


Figure 37: Neural Network 1 misclassification rate

The classification chart below shows the performance of the training set on the validation dataset, it indicates that the model has learned from the dataset to be able to make predictions.

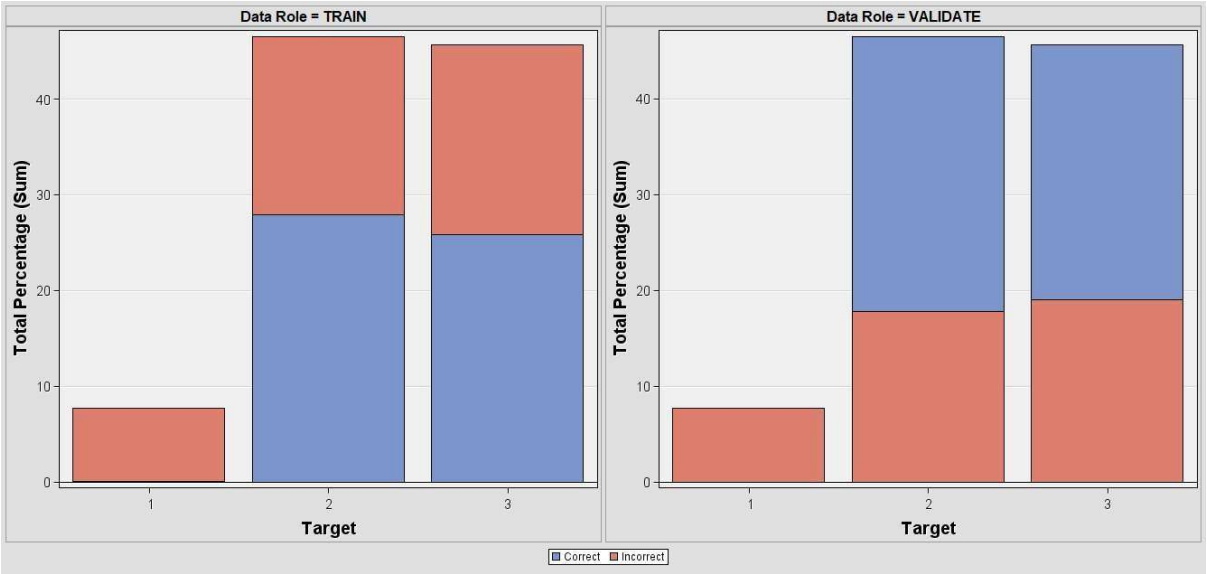
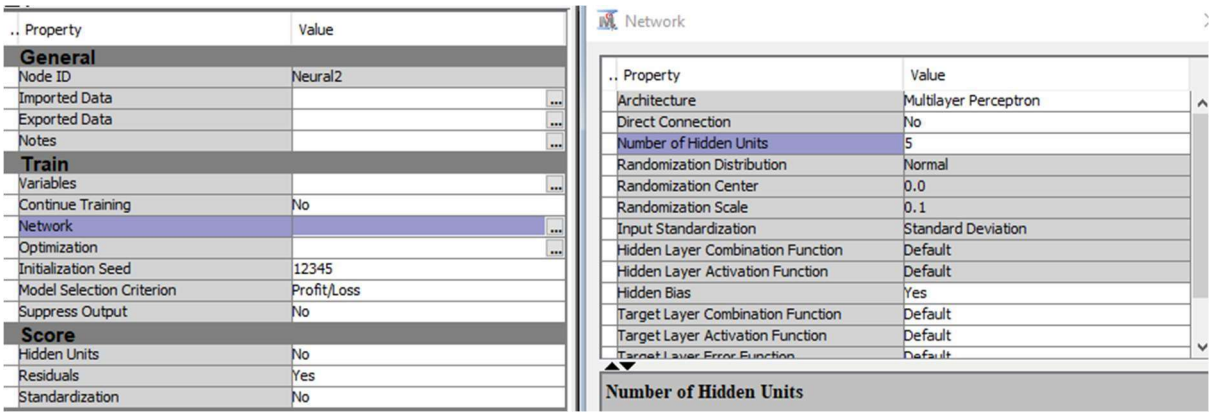


Figure 38: Neural Network 1 target variable classification

Neural Network 2

- Number of hidden layers – 5
- Maximum iterations – 30
- Maximum time – 2 hours
- Number of runs – 2



Property	Value
General	
Node ID	Neural2
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Profit/Loss
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No

Network

Property	Value
Architecture	Multilayer Perceptron
Direct Connection	No
Number of Hidden Units	5
Randomization Distribution	Normal
Randomization Center	0.0
Randomization Scale	0.1
Input Standardization	Standard Deviation
Hidden Layer Combination Function	Default
Hidden Layer Activation Function	Default
Hidden Bias	Yes
Target Layer Combination Function	Default
Target Layer Activation Function	Default

Number of Hidden Units

Figure 39: Neural Network 2 hyperparameter

There is no improvement in this model's accuracy as seen in the misclassification rate. It could mean that the hidden layers need to be further tuned.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14311	.
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	137	.
accident_severity	accident_severity	_NW_	Number of Estimated Weights	137	.
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12804.42	.
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	13842.65	.
accident_severity	accident_severity	_ASE_	Average Squared Error	0.18178	0.181372
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.987358	0.965952
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.426357	0.425878
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3939.535	984.8484
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.18526	.
accident_severity	accident_severity	_MSE_	Mean Squared Error	0.18352	0.181372
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.430419	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.428392	0.425878
accident_severity	accident_severity	_AVERR_	Average Error Function	0.578185	0.579844
accident_severity	accident_severity	_ERR_	Error Function	12530.42	3148.555
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.461379	0.457459
accident_severity	accident_severity	_WRONG_	Number of Wrong Classifications	3333	828

Figure 40: Neural Network 2 statistics

The misclassification shows that the pattern of the training and validation sets is not in sync, this could mean that the model has not learned enough for prediction.

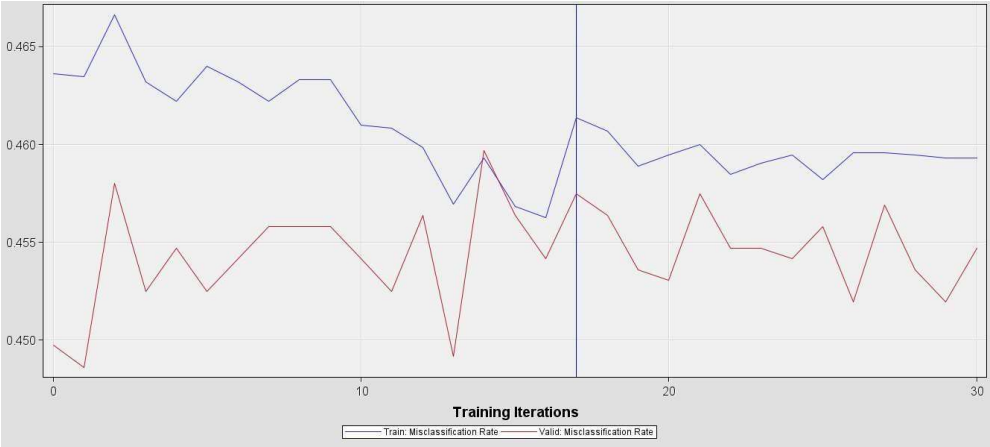


Figure 41: Neural Network 2 misclassification rate

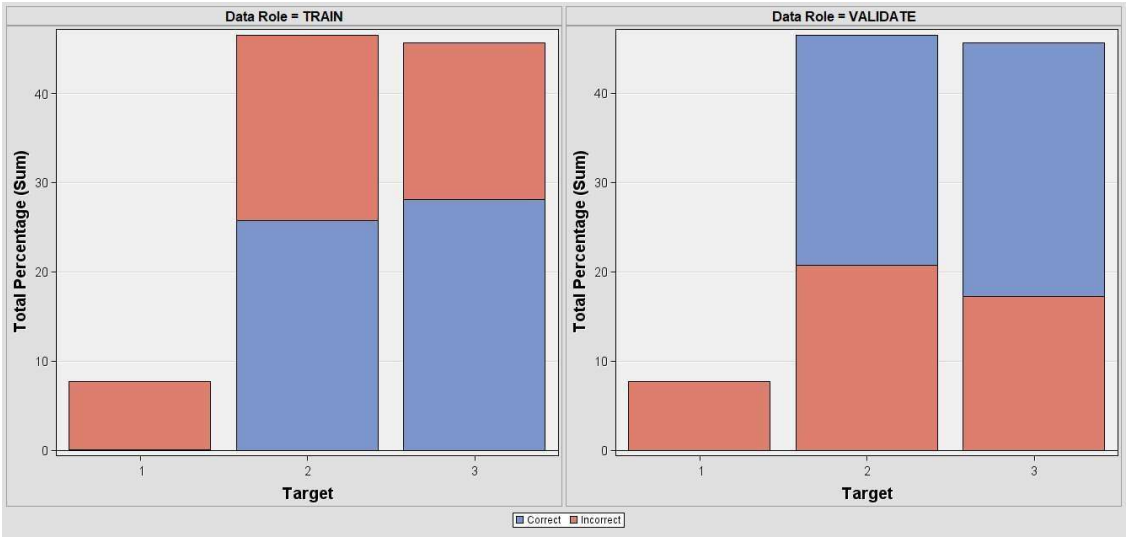
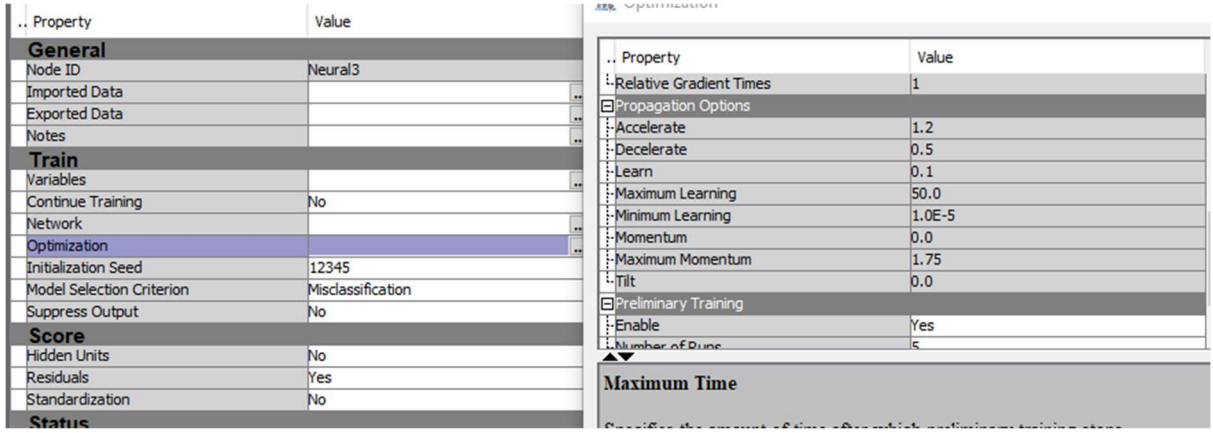


Figure 42: Neural Network 2 target variable classification

Neural Network 3

- Number of hidden layers – 10
- Number of iterations – 100
- Maximum time – 7 hours
- Number of runs – 5
- Model selection criterion – Misclassification



Property	Value
General	
Node ID	Neural3
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Misclassification
Suppress Output	No
Score	
Hidden Units	No
Residuals	Yes
Standardization	No
Status	

Property	Value
Relative Gradient Times	1
Propagation Options	
Accelerate	1.2
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	Yes
Number of Runs	5

Maximum Time

Figure 43: Neural Network 3 hyperparameter

The accuracy of this model on the validation set is about 55%, it is better than the previous model because of the higher value of iterations and hidden layers. The model has greater capacity to learn complex patterns in the data.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label ▲	Train	Validation
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12886.78	.
accident_severity	accident_severity	_AVERR_	Average Error Function	0.573218	0.584301
accident_severity	accident_severity	_ASE_	Average Squared Error	0.180036	0.182482
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14216	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_ERR_	Error Function	12422.78	3172.754
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.185913	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.989464	0.967746
accident_severity	accident_severity	_MSE_	Mean Squared Error	0.182974	0.182482
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.454734	0.441989
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	232	.
accident_severity	accident_severity	_NW_	Number of Estimated Weights	232	.
accident_severity	accident_severity	_WRONG_	Number of Wrong Classifications	3285	800
accident_severity	accident_severity	_RASE_	Root Average Squared Error	0.424307	0.427179
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.431176	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.427755	0.427179
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	14644.95	.
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3901.748	990.8766
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.

Figure 44: Neural Network 3 statistics

Figure 45: Neural Network 3 misclassification rate

The target variable classification below shows that the model has improved, as the correct predictions has increased on the validation data.

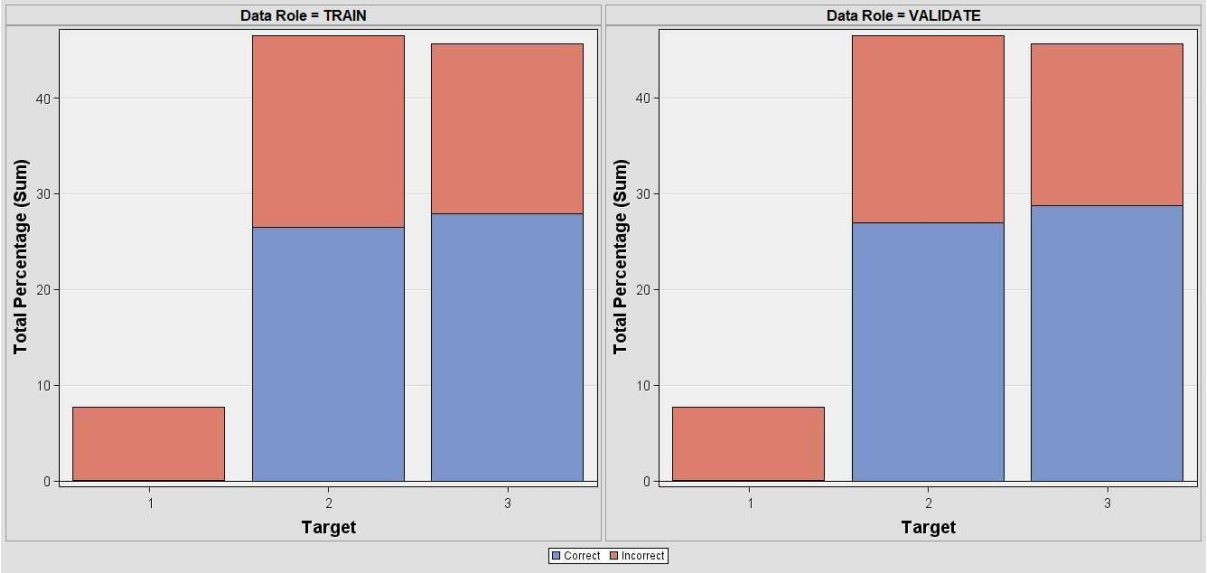


Figure 46: Neural Network 3 target variable classification

LOGISTIC REGRESSION

According to Stevens (2002), ordinal logistic regression can be defined as a statistical method that models the relationship between an ordinal dependent variable and one or more independent variables, while taking into account the ordinal nature of the dependent variable. It models the relationship between the independent variables and the probability of the dependent variable falling into one of the categories.

Logistic Regression 1

Variable transformation

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
IMP_age_of_driver	INPUT	41.1704	15.43207	7224	0	13	39	86	0.603642	-0.15897
number_of_casualties	INPUT	1.301772	0.696716	7224	0	1	1	5	2.803858	8.58294
number_of_vehicles	INPUT	1.792497	0.694508	7224	0	1	2	5	0.980723	2.353302

Figure 47: Continuous variable skewness

- The age of driver dataset has a positive skewness of 0.603642, which indicates that the distribution is slightly right skewed. This means that there are more observations with lower ages than higher ages, and the tail of the distribution extends towards higher ages.
- The number of casualties' dataset has a highly positive skewness of 2.803858, indicating a highly right-skewed distribution. This means that there are very few observations with high values, and most observations have low values.
- The number of vehicles has a positive skewness of 0.980723, indicating a slightly right-skewed distribution. This means that there are more observations with fewer vehicles involved in accidents, and the tail of the distribution extends towards higher numbers of vehicles involved.

Here, we transform the variables by using the Log10 for the continuous variables to normalize the data and reduce the skewness, it will also help capture non-Logistic relationships and improve the fit of the model.

Input Name	Role	Input Level	Name	Level	Formula
IMP_age_of_driver	INPUT	INTERVAL	LG10_IMP_age_of_driver	INTERVAL	log10(IMP_age_of_driver + 1)
number_of_casualties	INPUT	INTERVAL	LG10_number_of_casualties	INTERVAL	log10(number_of_casualties + 1)
number_of_vehicles	INPUT	INTERVAL	LG10_number_of_vehicles	INTERVAL	log10(number_of_vehicles + 1)

Figure 48: Transformed continuous variables

Now, we run the model on the transformed variable dataset using the stepwise selection model

Property	Value
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Stepwise
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	
Output Options	

Figure 49: Logistic Regression 1 hyperparameter

The model has a better performance than the baseline with an accuracy of about 54% for the validation set.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12699.94	.	.
accident_severity	accident_severity	_ASE_	Average Squared Error	0.183905	0.182686	.
accident_severity	accident_severity	_AVERR_	Average Error Function	0.584622	0.584132	.
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14433	.	.
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	15	.	.
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430	.
accident_severity	accident_severity	_ERR_	Error Function	12669.94	3171.838	.
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.184287	.	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.966282	0.970539	.
accident_severity	accident_severity	_MSE_	Mean Square Error	0.184096	0.182686	.
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810	.
accident_severity	accident_severity	_NW_	Number of Estimate Weights	15	.	.
accident_severity	accident_severity	_RASE_	Root Average Sum of Squares	0.428841	0.427417	.
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.429286	.	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.429064	0.427417	.
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	12813.61	.	.
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3985.58	991.9834	.
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430	.
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.472176	0.456906	.

Figure 50: Logistic Regression 1 statistics

The classification of the target variable shows that the model learn and there is a little better performance in the prediction of the validation set.

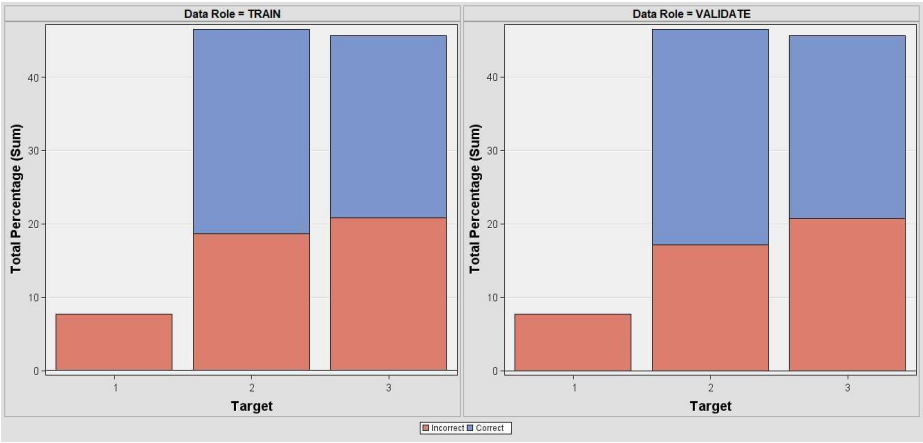


Figure 51: Logistic Regression 1 target variable classification

The misclassification rate has fairly improved as there is uniformity in the training set pattern and the validation set pattern.

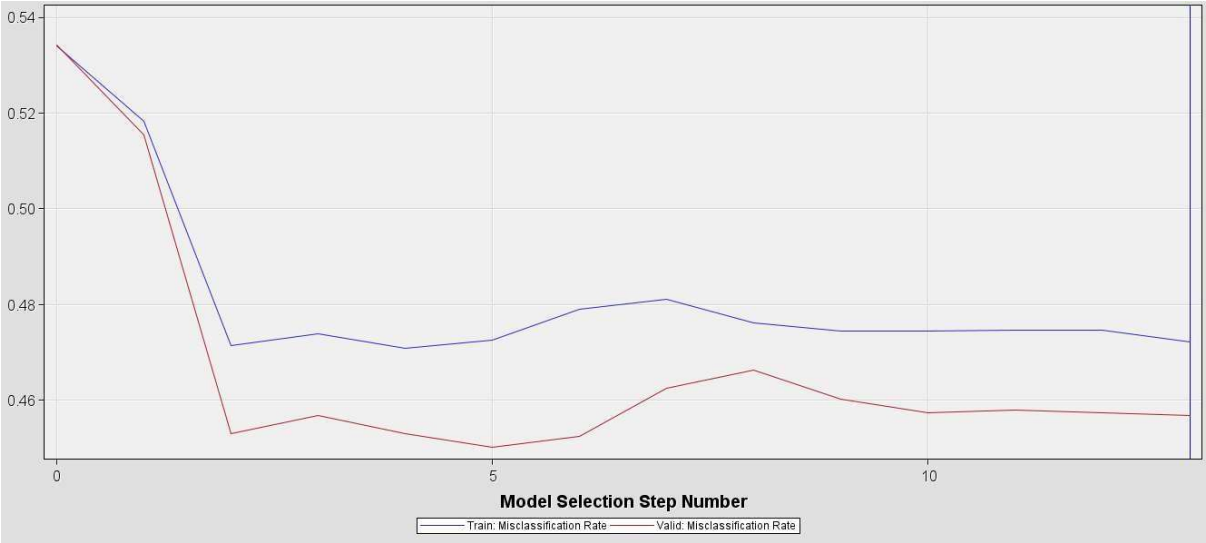


Figure 52: Logistic Regression 1 misclassification rate

Logistic Regression 2

Here, we will use the same parameters as with the Logistic Regression 1 model, the difference will be the selection models this will be tuned backwards, and the criterion will be set to "Validation Misclassification".

Property	Value
User Terms	No
Term Editor	...
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Backward
Selection Criterion	Validation Misclassification
Use Selection Defaults	Yes
Selection Options	...
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0
Max Function Calls	0
Maximum Time	1 Hour
Convergence Criteria	
Uses Defaults	Yes
Options	...

Figure 53: Logistic Regression 2 hyperparameter

There is no difference in the performance of the stepwise selection and the backwards selection in this model.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12699.94	.	.
accident_severity	accident_severity	_ASE_	Average Squared Error	0.183905	0.182686	.
accident_severity	accident_severity	_AVERR_	Average Error Function	0.584622	0.584132	.
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14433	.	.
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	15	.	.
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430	.
accident_severity	accident_severity	_ERR_	Error Function	12669.94	3171.838	.
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.184287	.	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.966282	0.970539	.
accident_severity	accident_severity	_MSE_	Mean Square Error	0.184096	0.182686	.
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810	.
accident_severity	accident_severity	_NW_	Number of Estimate Weights	15	.	.
accident_severity	accident_severity	_RASE_	Root Average Sum of Squares	0.428841	0.427417	.
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.429286	.	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.429064	0.427417	.
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	12813.61	.	.
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3985.58	991.9834	.
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430	.
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.472176	0.456906	.

Figure 54: Logistic Regression 2 statistics

Logistic Regression 3

The original dataset will be used to see the performance of the model without transforming the variables. The forward model selection was used for this prediction.

Property	Value
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	Forward
Selection Criterion	Default
Use Selection Defaults	Yes
Selection Options	
Optimization Options	
Technique	Default
Default Optimization	Yes
Max Iterations	0

Figure 55: Logistic Regression 3 hyperparameter

The accuracy is not so far apart from the accuracy of the logged variable. The prediction is about 54% accurate, which indicates that the dataset is good enough to make predictions without transformation.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12735.37	.
accident_severity	accident_severity	_ASE_	Average Squared Error	0.184236	0.183262
accident_severity	accident_severity	_AVERR_	Average Error Function	0.585519	0.585707
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14425	.
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	23	.
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448	.
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430
accident_severity	accident_severity	_ERR_	Error Function	12689.37	3180.387
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.184823	.
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.967975	0.972351
accident_severity	accident_severity	_MSE_	Mean Square Error	0.184529	0.183262
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810
accident_severity	accident_severity	_NW_	Number of Estimate Weights	23	.
accident_severity	accident_severity	_RASE_	Root Average Sum of Squares	0.429227	0.428092
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.429911	.
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.429569	0.428092
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	12909.68	.
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3992.754	995.1147
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.475083	0.457459

Figure 56: Logistic Regression 3 statistics

The value of correct prediction in both training and validation set is average, there is room for improvement.

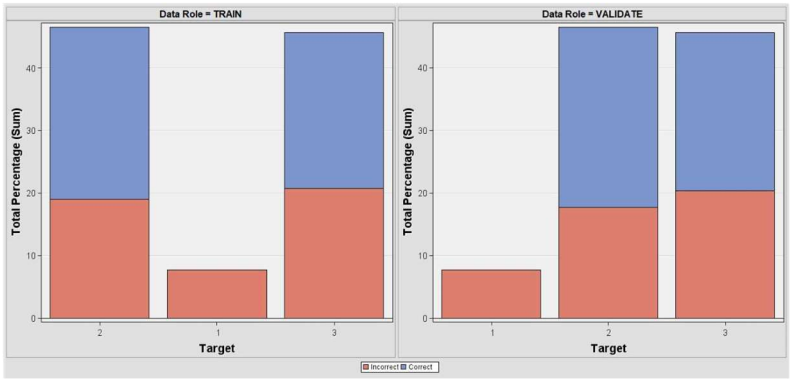


Figure 57: Logistic Regression 3 classification of target variable

There is not much difference in the pattern which indicates that the model has been learned

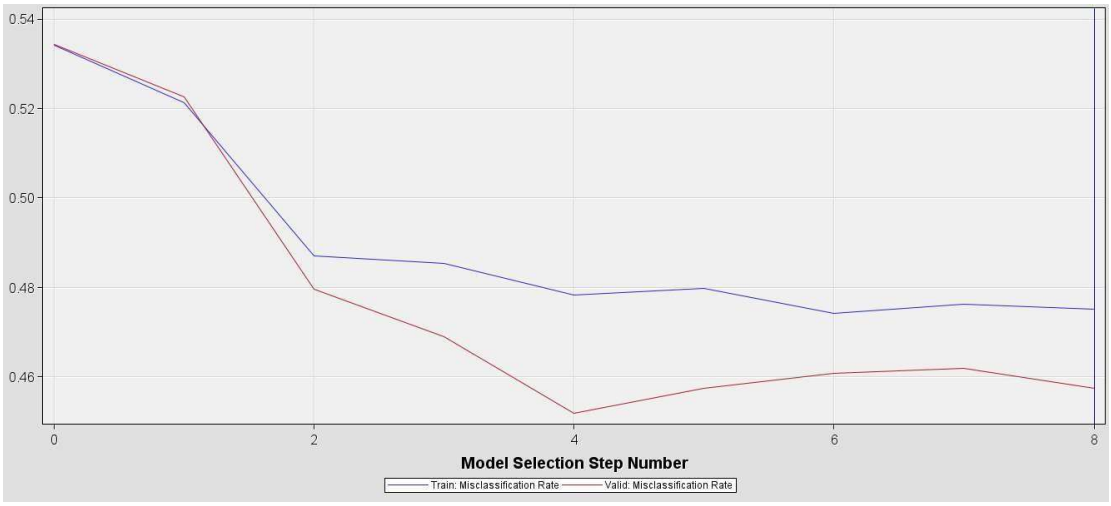


Figure 58: Logistic Regression 3 misclassification rate

Logistic Regression 4

Here, we used the variables that are not transformed as inputs, then selected none as the model selection

Property	Value
General	
Node ID	Reg4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Equation	
Main Effects	Yes
Two-Factor Interactions	No
Polynomial Terms	No
Polynomial Degree	2
User Terms	No
Term Editor	
Class Targets	
Regression Type	Logistic Regression
Link Function	Logit
Model Options	
Suppress Intercept	No
Input Coding	Deviation
Model Selection	
Selection Model	None
Selection Criterion	Default

Figure 59: Logistic Regression 4 hyperparameter

The statistics show that there is not much improvement with the model. The validation accuracy is about 54%. However, the model improved from the baseline accuracy.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
accident_severity	accident_severity	_AIC_	Akaike's Information Criterion	12736.47		
accident_severity	accident_severity	_ASE_	Average Squared Error	0.184177	0.183258	
accident_severity	accident_severity	_AVERR_	Average Error Function	0.585477	0.58553	
accident_severity	accident_severity	_DFE_	Degrees of Freedom for Error	14424		
accident_severity	accident_severity	_DFM_	Model Degrees of Freedom	24		
accident_severity	accident_severity	_DFT_	Total Degrees of Freedom	14448		
accident_severity	accident_severity	_DIV_	Divisor for ASE	21672	5430	
accident_severity	accident_severity	_ERR_	Error Function	12688.47	3179.427	
accident_severity	accident_severity	_FPE_	Final Prediction Error	0.18479		
accident_severity	accident_severity	_MAX_	Maximum Absolute Error	0.968147	0.972059	
accident_severity	accident_severity	_MSE_	Mean Square Error	0.184483	0.183258	
accident_severity	accident_severity	_NOBS_	Sum of Frequencies	7224	1810	
accident_severity	accident_severity	_NW_	Number of Estimate Weights	24		
accident_severity	accident_severity	_RASE_	Root Average Sum of Squares	0.429158	0.428086	
accident_severity	accident_severity	_RFPE_	Root Final Prediction Error	0.429872		
accident_severity	accident_severity	_RMSE_	Root Mean Squared Error	0.429515	0.428086	
accident_severity	accident_severity	_SBC_	Schwarz's Bayesian Criterion	12918.35		
accident_severity	accident_severity	_SSE_	Sum of Squared Errors	3991.476	995.0884	
accident_severity	accident_severity	_SUMW_	Sum of Case Weights Times Freq	21672	5430	
accident_severity	accident_severity	_MISC_	Misclassification Rate	0.475637	0.460221	

Figure 60: Logistic Regression 4 statistics

The validation data shows slight improvement in the prediction of accident severity in the classification of the target variable

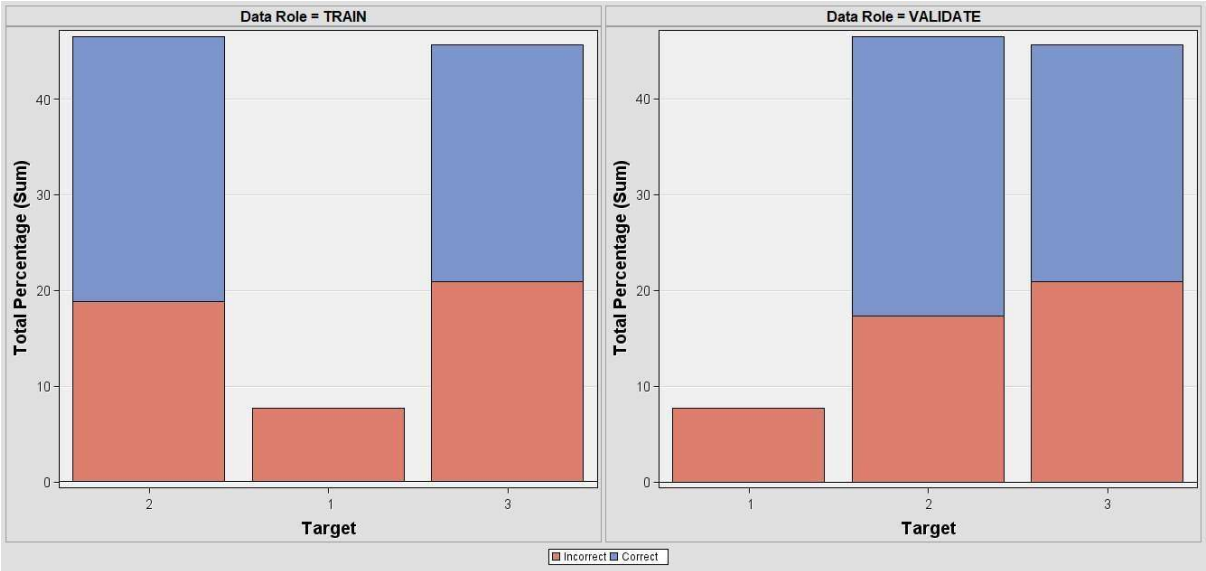


Figure 61: Logistic Regression 4 classification of target variable

MODEL EVALUATION

The performance metric used to assess the performance of each of the model in this analysis, is 'accuracy'. This is calculated with the below equation

$$\text{Accuracy} = 1 - \text{Misclassification}$$

Model Description	Train: Sum of Frequencies	Valid: Sum of Frequencies	Train: Misclassification Rate	Valid: Misclassification Rate	Train: Average Squared Error ▼	Valid: Average Squared Error	Train: Maximum Absolute Error
Undersampling Decision Tree	1483	373	0.525287	0.538874	0.20369	0.21126	0.907869
seventh decision tree	7224	1810	0.508998	0.516575	0.187069	0.188029	0.940356
Fourth decision tree	7224	1810	0.504014	0.504972	0.184692	0.185574	0.964882
Baseline model	7224	1810	0.502076	0.5	0.184499	0.185685	0.964882
fifth decision tree	7224	1810	0.497647	0.502762	0.184244	0.185946	0.964882
Sixth decision tree	7224	1810	0.496955	0.505525	0.184243	0.187208	0.969697
Regression 3	7224	1810	0.475083	0.457459	0.184236	0.183262	0.967975
Regression (4)	7224	1810	0.475637	0.460221	0.184177	0.183258	0.968147
Transformed regression stepwise	7224	1810	0.472176	0.456906	0.183905	0.182686	0.966282
Transformed Regression Backw...	7224	1810	0.472176	0.456906	0.183905	0.182686	0.966282
First decision tree	7224	1810	0.498616	0.498343	0.183473	0.185949	0.967367
First Neural Network	7224	1810	0.458887	0.446961	0.181951	0.181825	0.969367
third neural network	7224	1810	0.458333	0.445304	0.181451	0.182137	0.968146
second neural network	7224	1810	0.45778	0.450829	0.181174	0.182304	0.972061
Third decision tree	7224	1810	0.483527	0.481768	0.18088	0.186727	0.991803

Figure 62: Model evaluation

We assessed the accuracy of each model and then tuned various hyperparameters to achieve the optimum accuracy from the models used.

The decision tree model is a simple yet powerful algorithm that is easy to interpret. However, decision trees are prone to overfitting, and their performance can suffer when dealing with

imbalanced datasets like the target variable. The performance of the decision tree is barely above the benchmark which is not good enough for analysis.

Neural networks, on the other hand, are more complex models that are capable of handling non-Logistic relationships between variables. However, neural networks can be difficult to interpret, and their decisions may not be easily explained. The neural network gave the least misclassification rate in this analysis and is about 5% better than the baseline model.

Logistic regression is a simple computationally efficient model that can be easily interpreted. However, logistic regression assumes a Logistic relationship between the input variables and the output, which may not always be the case. The performance is also barely above the baseline model and is not as good as the neural network.

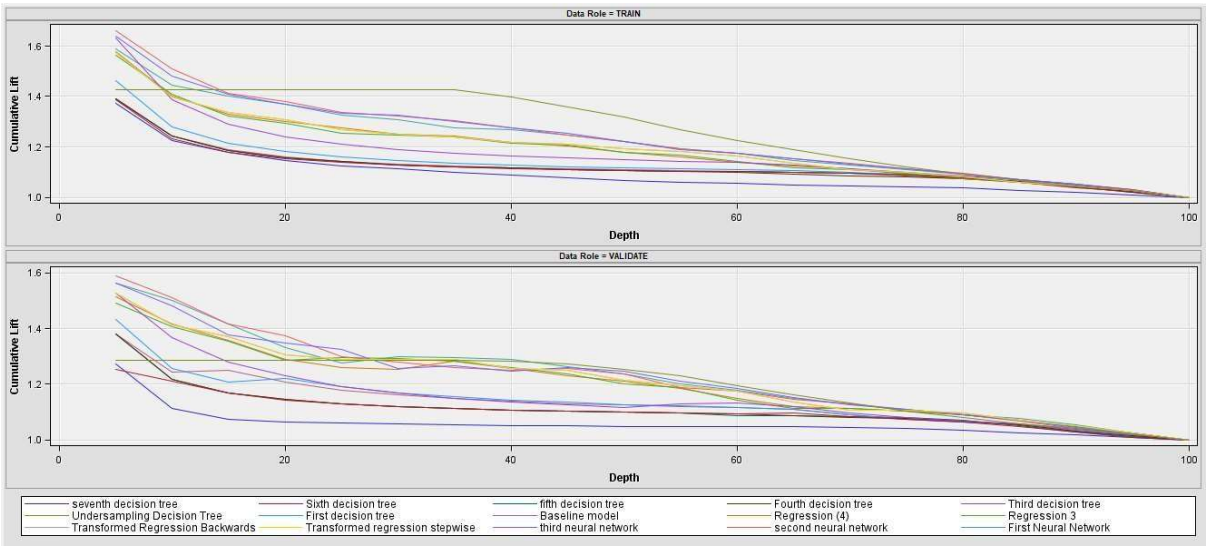


Figure 63: Model Evaluation cumulative lift

CONCLUSION

In conclusion, we have evaluated the performance of decision tree, neural network, and logistic regression models for predicting accident severity in a dataset. We found that the neural network model performed the best in terms of accuracy, although it is important to consider other metrics such as precision, recall, F1-score, and AUC-ROC score.

Some possible future improvements for this study include:

Feature engineering: Adding new features or combining existing ones may improve the performance of the models.

Ensemble methods: Combining multiple models through ensemble methods such as bagging, boosting, or stacking may improve the overall performance of the model.

Model interpretability: Improving the interpretability of the models may provide more insights into the factors that contribute to accident severity. Techniques such as decision rules can be used to explain the models' decisions.

The model and findings of this study can help a motor insurance company improve risk assessment, claims processing, accident prevention, and customer experience, which can ultimately lead to improved profitability and customer satisfaction.

Recommendation

The model is not 100% accurate, hence there is likely to be error in prediction which could give false negative or false positive values that could affect the budget and forecast of the motor insurance company. To mitigate this, we can adjust the algorithm to consider the cost of errors in the loss function and evaluate the model based on business metric like investment return and then ensure it is optimized to provide value to the company.

REFERENCES

Aggarwal, C.C. (2019). Neural networks and deep learning: A textbook. New York: Springer.

Department for Transport (2022). Road Traffic Act 1988. Retrieved from https://www.legislation.gov.uk/ukpga/1988/52/contents

Department for Transport (2022). Road Safety Data. [online] www.data.gov.uk. Available at: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safetydata.

Sarma, K.S. (2017). Predictive Modelling with SAS Enterprise Miner. SAS Institute.

Stevens, J. (2009). Applied multivariate statistics for the social sciences. New York (N.Y.): Routledge, Cop.

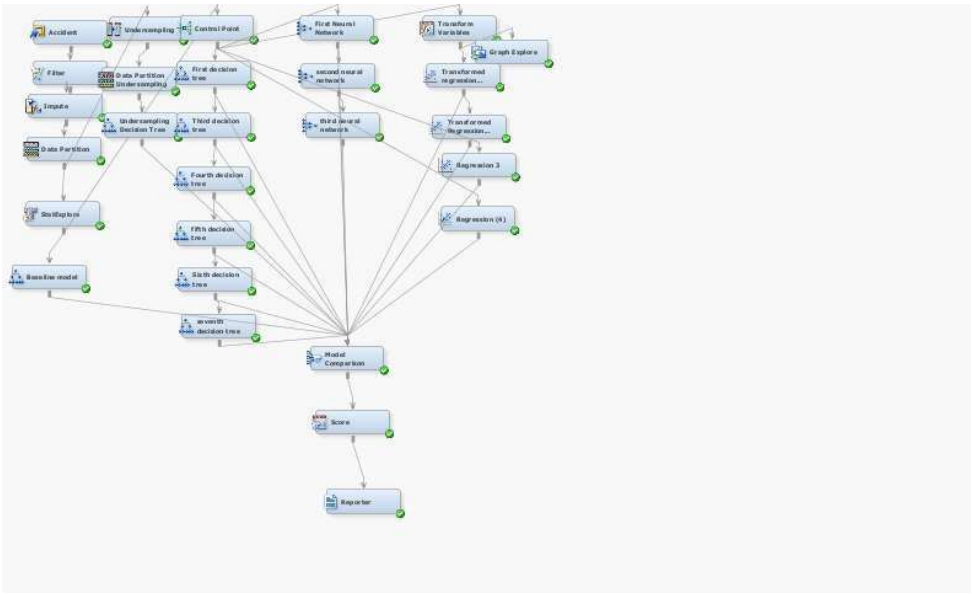
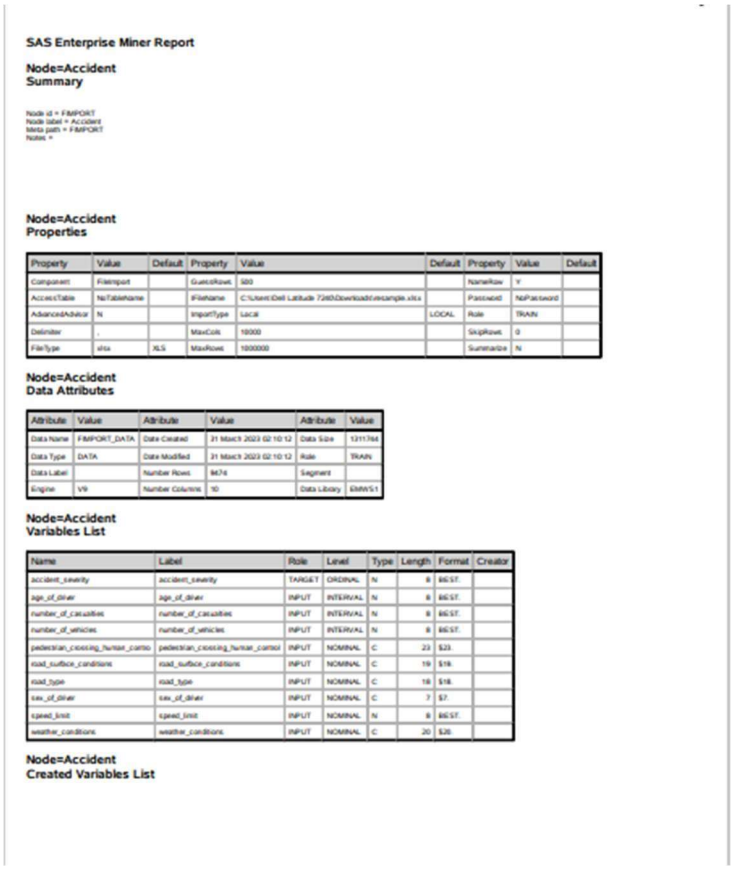


Figure 64: SAS enterprise miner inputs

Below is the output of the report file. It is over 100 pages, so snapshots of few of the pages have been added.



SAS Enterprise Miner Report

Node=Accident
Summary

Node id = FIMPORT
Node label = Accident
Meta path = FIMPORT
Notes =

Node=Accident
Properties

Property	Value	Default	Property	Value	Default	Property	Value	Default
Component	Filelmport		GuessRows	500		NameRow	Y	
AccessTable	NoTableName		FileName	C:\Users\Dell Latitude 7240\Download\rescample.xlsx		Password	NoPassword	
AdvancedAdvice	N		ImportType	Local	LOCAL	Role	TRAIN	
Delimiter			MaxCols	10000		SkipRows	0	
FileType	xlsx	XLS	MaxRows	1000000		Summarize	N	

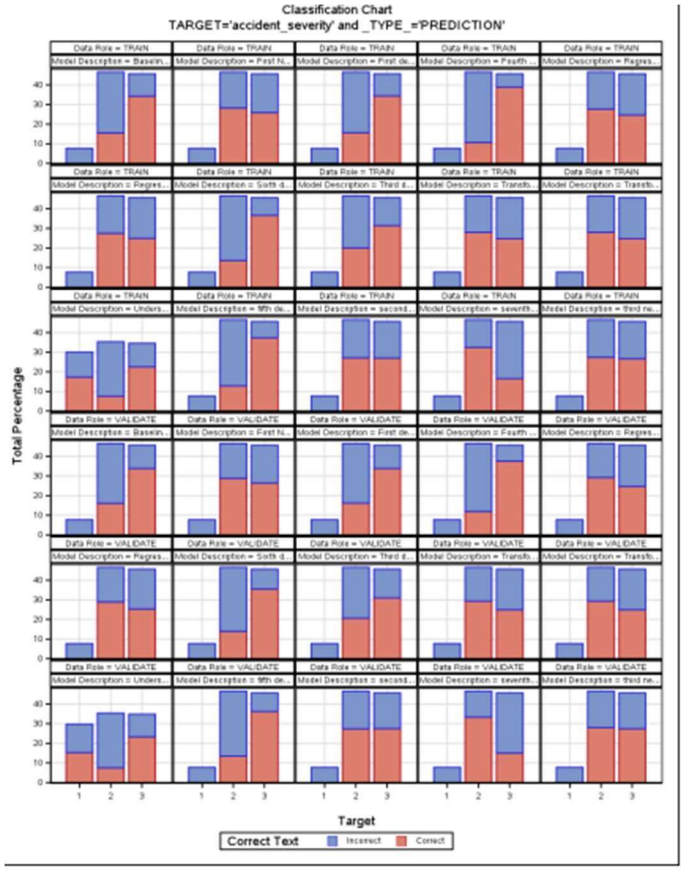
Node=Accident
Data Attributes

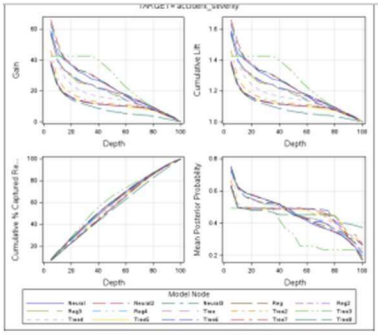
Attribute	Value	Attribute	Value	Attribute	Value
Data Name	FIMPORT_DATA	Date Created	31 March 2023 02:10:12	Data Size	1211764
Data Type	DATA	Date Modified	31 March 2023 02:10:12	Role	TRAIN
Data Label		Number Rows	8674	Segment	
Engine	V9	Number Columns	10	Data Library	EMWS1

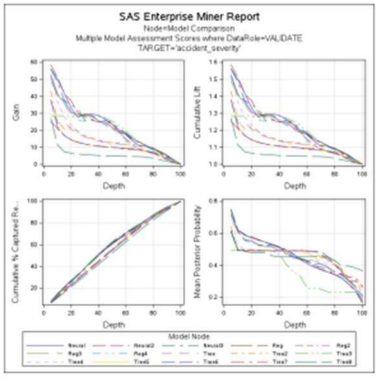
Node=Accident
Variables List

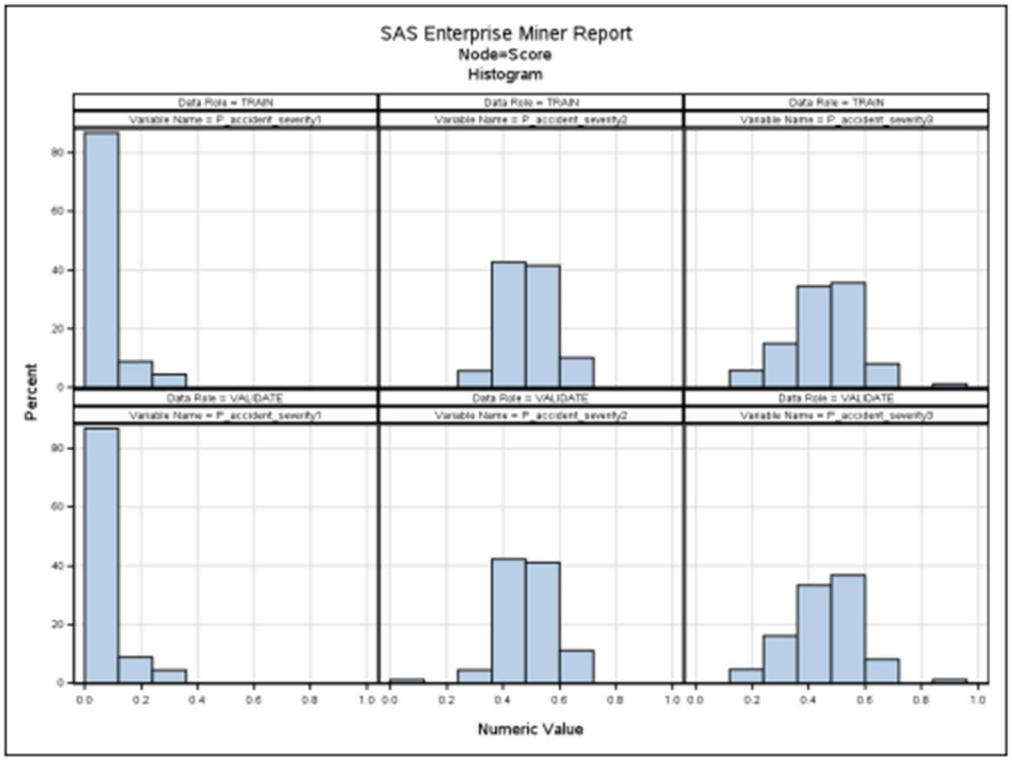
Name	Label	Role	Level	Type	Length	Format	Creator
accident_severity	accident_severity	TARGET	ORDINAL	N	8	BEST.	
age_of_driver	age_of_driver	INPUT	INTERVAL	N	8	BEST.	
number_of_casualties	number_of_casualties	INPUT	INTERVAL	N	8	BEST.	
number_of_vehicles	number_of_vehicles	INPUT	INTERVAL	N	8	BEST.	
pedestrian_crossing_human_contro	pedestrian_crossing_human_control	INPUT	NOMINAL	C	23	$23.	
road_surface_conditions	road_surface_conditions	INPUT	NOMINAL	C	19	$19.	
road_type	road_type	INPUT	NOMINAL	C	18	$18.	
sex_of_driver	sex_of_driver	INPUT	NOMINAL	C	7	$7.	
speed_limit	speed_limit	INPUT	NOMINAL	N	8	BEST.	
weather_conditions	weather_conditions	INPUT	NOMINAL	C	20	$20.	

Node=Accident
Created Variables List


Classification Chart
TARGET='accident_severity' and _TYPE_='PREDICTION'
Total Percentage
Target
Correct Text
Incorrect
Correct


Gain
Cumulative Lift
Depth
Model Node


SAS Enterprise Miner Report
Node=Model Comparison
Multiple Model Assessment Scores where DataRole=VALIDATE
TARGET='accident_severity'
Gain
Cumulative Lift
Depth
Model Node



End of Report