Advanced Analytics with SAS Enterprise Miner (EM)

Student Name *(as per record)*	VIJAY VIJAYAKUMAR / BHARATH KUMAR	Group No	32	Student No	215030005 / 214383649

	Exceptional	Meets expectations	Issues noted	Improvement needed	Unacceptable
Exec Summary: Q1					
Data Prep: Q2 & Q3					
Text Analytics: Q4					
Predict Models: Q5 & Q6					
Model Compar: Q7					
Brief Comments					Total

Workshop M3: Executive summary and recommendations (with cross-refs)

NHTSA (US National Highway Traffic Safety Administration) is an agency of the Executive branch of the government of USA as part of the Department of Transportation. They are mainly responsible for reducing deaths, injuries and economic losses resulting from motor vehicle crashes. The NHTSA requires an Early Warning System for potential safety issues associated with automotive vehicles due to manufacturing problems. A predictive model is to be developed which is capable of predicting the likelihood of a vehicle crash, based on publicly available vehicle safety complaints. Based on the results from this analytical model, the NHTSA will initiate a total recall of all the affected vehicles where the probability of crashes is high.

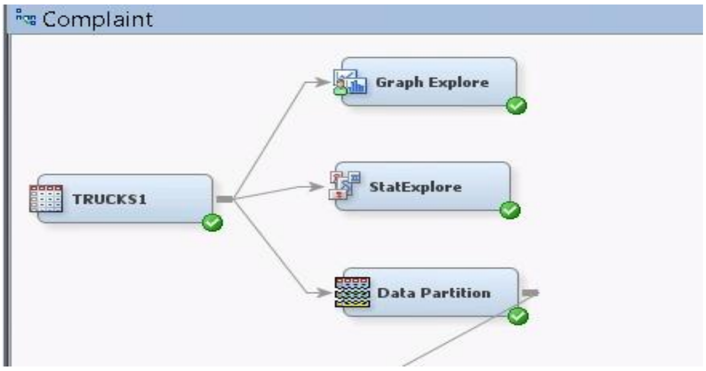
The planned predictive model aims to solve the client's capability to reduce the number of motor vehicular crashes due to various reasons. The proposed predictive model helps the client predict the likelihood of a vehicle crash based on publicly available vehicle safety complaints. The model will help the client make decisions to either recall all of the vehicles prone to crashes due to manufacturing problems or propose economic solutions to combat the problem. This model will serve as an Early Warning System and help to save the lives of many by predicting and preventing issues that may be responsible for a motor vehicle crash posing a threat to human lives and economic losses.

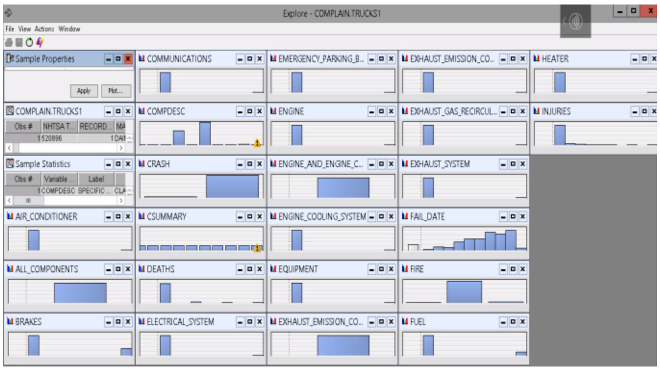
It is proposed to develop a predictive model that can identify factors causing motor vehicle crashes in the past and help prevent them from happening in the future to the lowest error margin. By studying patterns of motor vehicle crashes from past instances from publicly recorded complaint data, we can create a predictive model that can be used to analyze the causes of past motor vehicle crashes. And then the results can be used to forecast and prevent similar causes of these crashes in the future.

It is recommended to take up the predictive model with a Text Rule Builder node. This has been found to be the most effective predictive model with the least margin of error. Text Rule Builder consistently has the lowest ASE value and the best results in the ROC chart as seen from the cross-validation results. And this would provide the most impact to reduce or prevent similar future motor vehicle crashes by serving as an Early Warning System for the NHTSA.

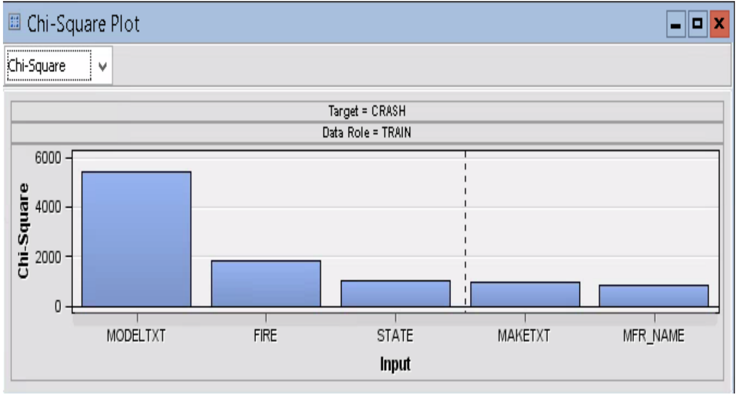
Data Preparation and Exploration in EM

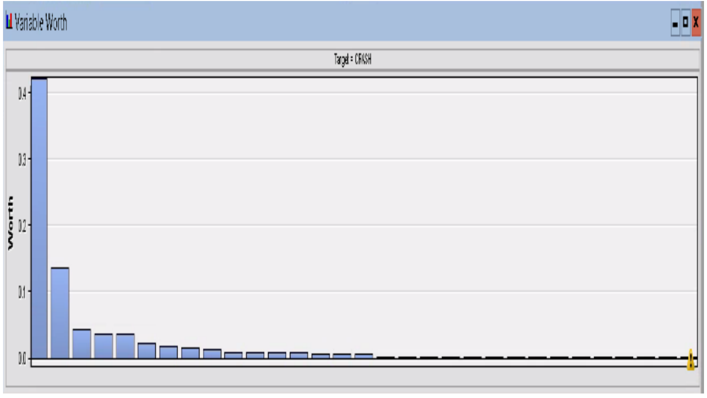
Data Exploration: The NHTSA data consist of 56601 observation or complaints. In this observations, the variable **CRASH** is set as binary target variable which has approximately 30 percent of crash problems.





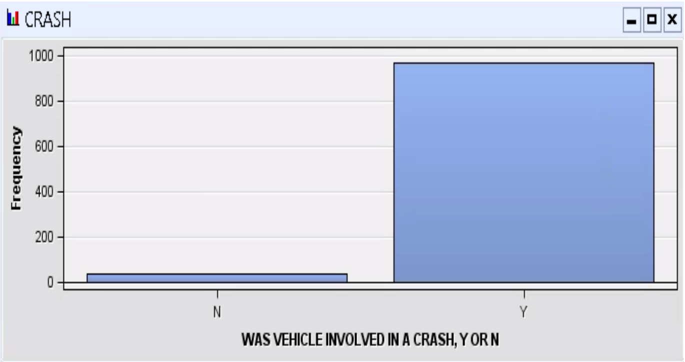
Stat Explore: The below **Chi-Square plot** explains the relationship between the input variables and the target variable (CRASH). Here the strongest correlation variable is Model Txt which holds the Chi square value of **5407.2729**.

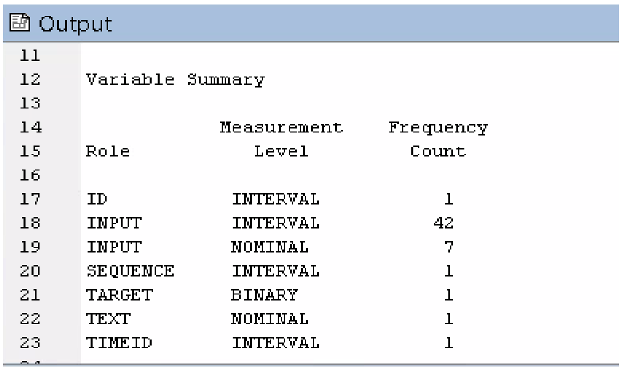




The **Variable worth plot** measures the independent values respective to the calculated worth. In this plot we can see that ‘crash flag’ holds the highest worth.

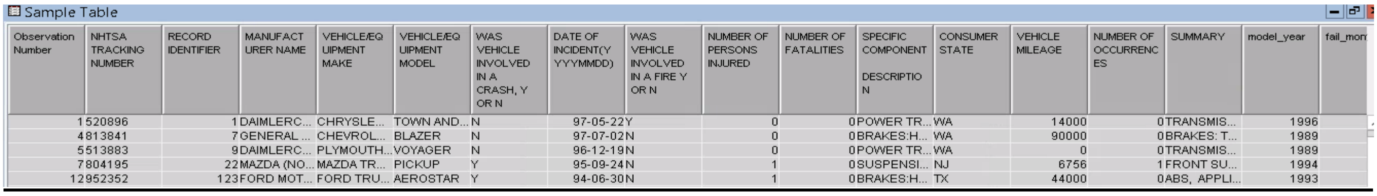
Graph Explore: Taking the Crash plot, we can understand that ‘vehicle involved in crash’ has frequency of 963.





```
Variable Summary

                  Measurement     Frequency
Role                 Level          Count

ID                 INTERVAL           1
INPUT              INTERVAL          42
INPUT              NOMINAL            7
SEQUENCE           INTERVAL           1
TARGET             BINARY             1
TEXT               NOMINAL            1
TIMEID             INTERVAL           1
```

The output table shows the Roles, measurement level and the frequency count of the data set. Here we have made Crash as target binary variable, Sequence variable changed from sequence type to ID type. **ODINO, COMPDESC, CRASH_FLG, INJURIES** variables are identified as potential anomalous or inconsistent data characteristics and thereby were eliminated.


Sample Table

Observation Number	NHTSA TRACKING NUMBER	RECORD IDENTIFIER	MANUFACTURER NAME	VEHICLE/EQUIPMENT MAKE	VEHICLE/EQUIPMENT MODEL	WAS VEHICLE INVOLVED IN A CRASH, Y OR N	DATE OF INCIDENT(YYYMMDD)	WAS VEHICLE INVOLVED IN A FIRE Y OR N	NUMBER OF PERSONS INJURED	NUMBER OF FATALITIES	SPECIFIC COMPONENT DESCRIPTION	CONSUMER STATE	VEHICLE MILEAGE	NUMBER OF OCCURRENCES	SUMMARY	model_year	fail_mon
1	520896	1	DAIMLERC...	CHRYSLE...	TOWN AND...	N	97-05-22	Y	0	0	POWER TR...	WA	14000	0	TRANSMIS...	1996	
4	813841	7	GENERAL ...	CHEVROL...	BLAZER	N	97-07-02	N	0	0	BRAKES:H...	WA	90000	0	BRAKES: T...	1989	
5	513883	9	DAIMLERC...	PLYMOUTH...	VOYAGER	N	96-12-19	N	0	0	POWER TR...	WA	0	0	TRANSMIS...	1989	
7	804195	22	MAZDA (NO...	MAZDA TR...	PICKUP	Y	95-09-24	N	1	0	SUSPENSI...	NJ	6756	1	FRONT SU...	1994	
12	952352	123	FORD MOT...	FORD TRU...	AEROSTAR	Y	94-06-30	N	1	0	BRAKES:H...	TX	44000	0	ABS, APPLI...	1993	

Figure 2.1: Sample data table

Text Analytics in EM (Page 1)

Text Analytics is the process of obtaining high quality of information from the text. This process is also known as Data or Text mining. It involves process of structuring the input by parsing, filtering, clustering and text topic. Text mining is the process of determining anomalies, correlation and patterns using the large data sets for predicting the output. Here NHTSA's sample data is taken (with 56,601 observations) with the target variable as CRASH.

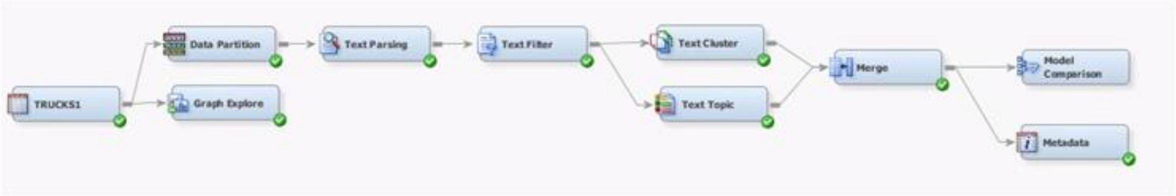


Figure 3.1: Model shows the process of text mining.

Text Parsing: It creates the term or document matrix and also this node enables the data set to parse in order to measure the given information. Text parsing node takes the control over exactly to the text or terms which need to be included in the analysis.

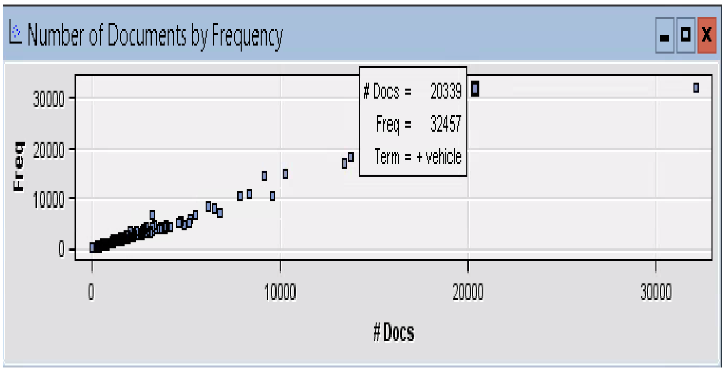


Figure 3.2: Numbers of documents by Frequency

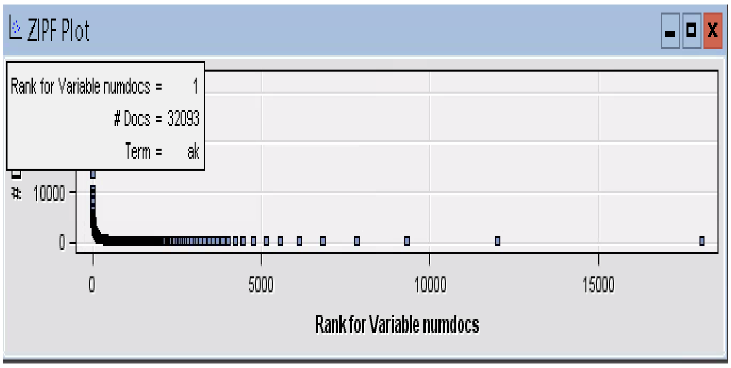
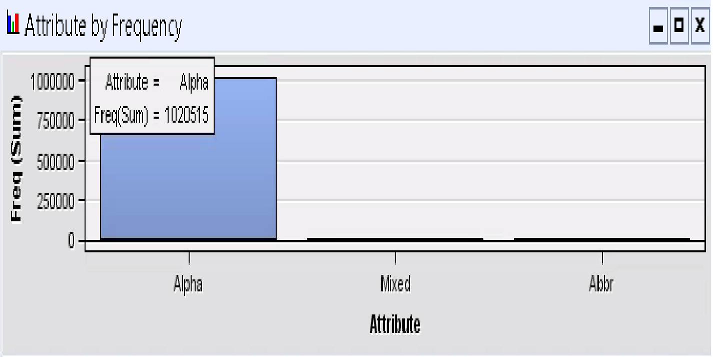
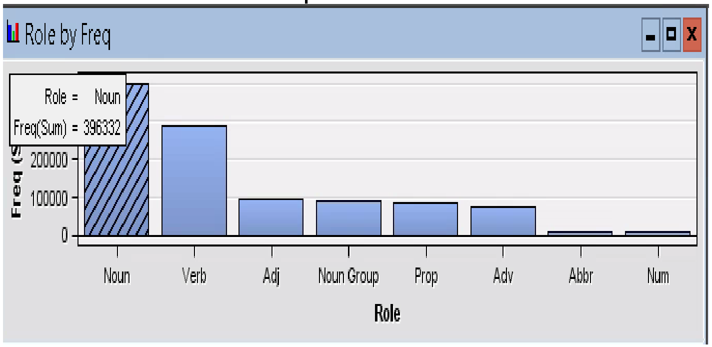


Figure 3.3: ZIPF Plot

This Number of Documents by frequency plot shows the position of the terms like Vehicle, be (from the data set).So, here the term +Vehicle has frequency of 32457.

In this ZIPF plot, select the term and notice the point corresponding to it. Here it also shows the rank of the selected term from the parsed documents. The term 'ak' positioned as 1 for variable numdocs.



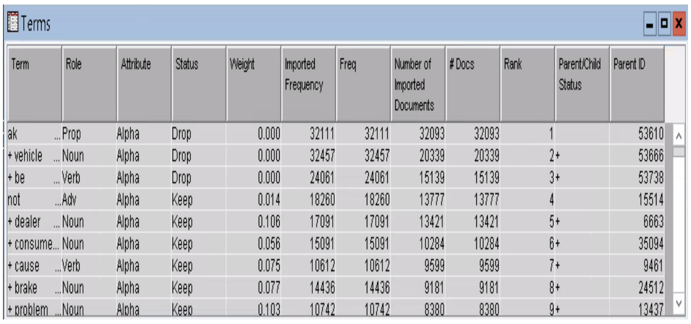


The attribute by frequency chart shows that **Alpha** has the highest number of frequency of 1020515 among attributes in the document collection. The Role by frequency chart shows that **Noun** has the highest frequency of 396332 among roles in the document collection or data set.

After the plots of Attribute and Role of term, we notice that the term 'Vehicle' is not kept in the text parsing analysis. This is shown by the value of **N** in the keep column. We came to know that, not all terms are kept in the analysis when we run the node with default settings. This node enables you to modify the output set of parsed terms by dropping terms which are in certain parts of speech, and other entity or attributes.

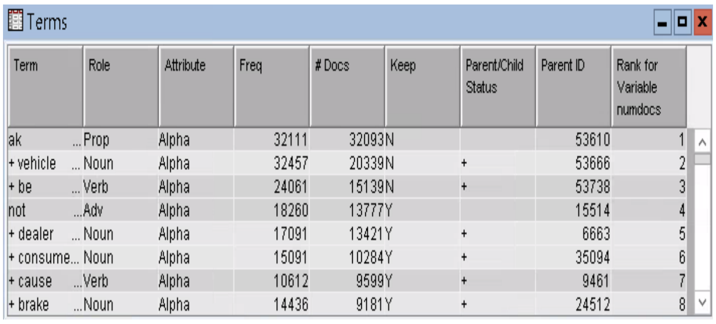
Text Analytics in EM (Page 2)

Text Filter Node: This node is used to decrease the total number of terms (which are parsed). So that we can eliminate the additional information from the docs, this helps to retrieve only the most valuable terms.



Terms

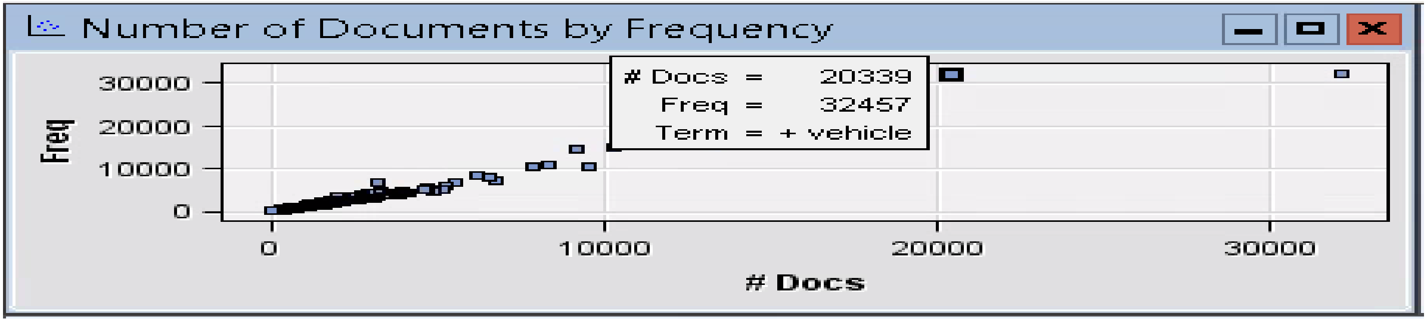
Term	Role	Attribute	Status	Weight	Imported Frequency	Freq	Number of Imported Documents	# Docs	Rank	Parent/Child Status	Parent ID
ak ...	Prop	Alpha	Drop	0.000	32111	32111	32093	32093	1		53610
+ vehicle ...	Noun	Alpha	Drop	0.000	32457	32457	20339	20339	2	+	53666
+ be ...	Verb	Alpha	Drop	0.000	24061	24061	15139	15139	3	+	53738
not ...	Adv	Alpha	Keep	0.014	18260	18260	13777	13777	4		15514
+ dealer ...	Noun	Alpha	Keep	0.106	17091	17091	13421	13421	5	+	6663
+ consume...	Noun	Alpha	Keep	0.056	15091	15091	10284	10284	6	+	35094
+ cause ...	Verb	Alpha	Keep	0.075	10612	10612	9599	9599	7	+	9461
+ brake ...	Noun	Alpha	Keep	0.077	14436	14436	9181	9181	8	+	24512
+ problem ...	Noun	Alpha	Keep	0.103	10742	10742	8380	8380	9	+	13437



Terms

Term	Role	Attribute	Freq	# Docs	Keep	Parent/Child Status	Parent ID	Rank for Variable numdocs
ak ...	Prop	Alpha	32111	32093	N		53610	1
+ vehicle ...	Noun	Alpha	32457	20339	N	+	53666	2
+ be ...	Verb	Alpha	24061	15139	N	+	53738	3
not ...	Adv	Alpha	18260	13777	Y		15514	4
+ dealer ...	Noun	Alpha	17091	13421	Y	+	6663	5
+ consume...	Noun	Alpha	15091	10284	Y	+	35094	6
+ cause ...	Verb	Alpha	10612	9599	Y	+	9461	7
+ brake ...	Noun	Alpha	14436	9181	Y	+	24512	8

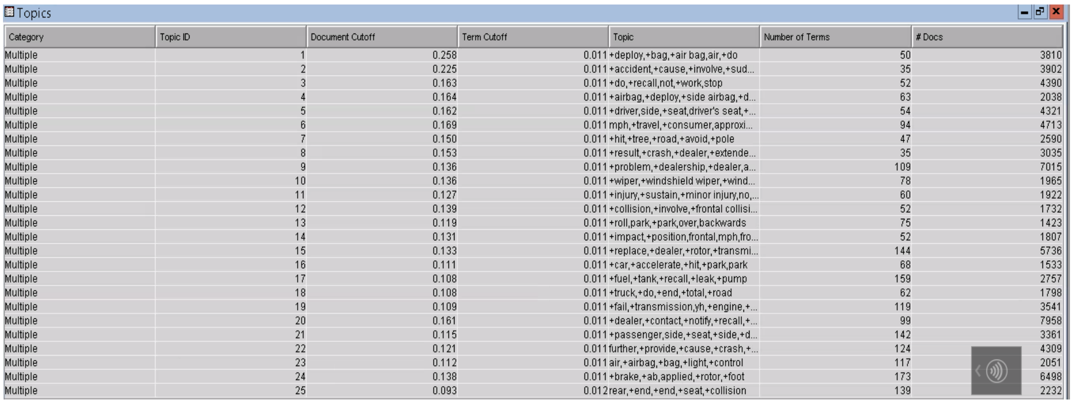
Here based on the assumption, we can also treat two similar terms as one. Like Fail and Failure can be added as one synonym. So here after processing with text filter node, again the frequency of term in the docs will be changed. Here the term 'Vehicle' frequency is not changed as we didn't add any synonyms. Here we can also resize and include only the terms under or equal to particular documents.



Number of Documents by Frequency

# Docs = 20339
Freq = 32457
Term = + vehicle

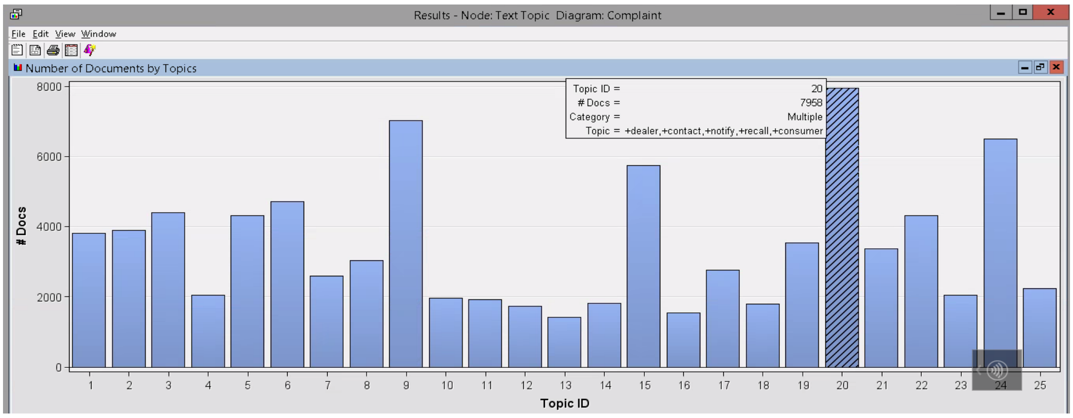
Text topic node : This node helps to explore the docs by automatically taking terms and documents according to both identified and user defined topics (Topics are nothing bit collection of terms that explains about main theme). The process of combining terms to topics can improve the analysis.



Topics

Category	Topic ID	Document Cutoff	Term Cutoff	Topic	Number of Terms	# Docs
Multiple	1	0.258	0.011	+deploy,+bag,+air bag,air,+do	50	3810
Multiple	2	0.225	0.011	+accident,+cause,+involve,+sud...	35	3902
Multiple	3	0.163	0.011	+do,+recall,not,+work,stop	52	4390
Multiple	4	0.164	0.011	+airbag,+deploy,+side airbag,+d...	63	2038
Multiple	5	0.162	0.011	+driver,side,+seat,driver's seat,+...	54	4321
Multiple	6	0.169	0.011	mph,+travel,+consumer,approxi...	94	4713
Multiple	7	0.150	0.011	+hit,+tree,+road,+avoid,+pole	47	2590
Multiple	8	0.153	0.011	+result,+crash,+dealer,+extende...	35	3035
Multiple	9	0.136	0.011	+problem,+dealership,+dealer,a...	109	7015
Multiple	10	0.136	0.011	+wiper,+windshield wiper,+wind...	78	1965
Multiple	11	0.127	0.011	+injury,+sustain,+minor injury,no,...	60	1922
Multiple	12	0.139	0.011	+collision,+involve,+frontal collisi...	52	1732
Multiple	13	0.119	0.011	+roll,park,+park,over,backwards	75	1423
Multiple	14	0.131	0.011	+impact,+position,frontal,mph,fro...	52	1807
Multiple	15	0.133	0.011	+replace,+dealer,+rotor,+transmi...	144	5736
Multiple	16	0.111	0.011	+car,+accelerate,+hit,+park,park	68	1533
Multiple	17	0.108	0.011	+fuel,+tank,+recall,+leak,+pump	159	2757
Multiple	18	0.108	0.011	+truck,+do,+end,+total,+road	62	1798
Multiple	19	0.109	0.011	+fail,+transmission,yh,+engine,+...	119	3541
Multiple	20	0.161	0.011	+dealer,+contact,+notify,+recall,+...	99	7958
Multiple	21	0.115	0.011	+passenger,side,+seat,+side,+d...	142	3361
Multiple	22	0.121	0.011	further,+provide,+cause,+crash,+...	124	4309
Multiple	23	0.112	0.011	air,+airbag,+bag,+light,+control	117	2051
Multiple	24	0.138	0.011	+brake,+ab,applied,+rotor,+foot	173	6498
Multiple	25	0.093	0.012	rear,+end,+end,+seat,+collision	139	2232

This Topics table have been created by using the default run of this text topic node. Here the number of documents by topics plot shows the position or status of the topic by the number of documents that it contains. And also this can be converted into multi-term topics.



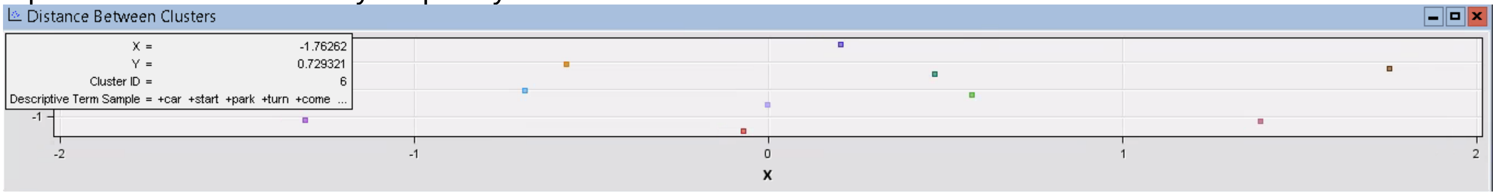
Results - Node: Text Topic Diagram: Complaint

Number of Documents by Topics

Topic ID = 20
# Docs = 7958
Category = Multiple
Topic = +dealer,+contact,+notify,+recall,+consumer

Text Analytics in EM (Page 3)

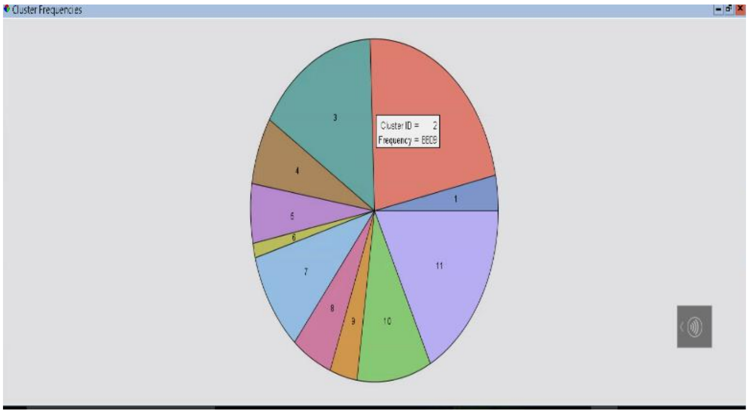
Text cluster node: This node clusters the docs to disjointed sets of documents and reports. The clusters table consist of Cluster ID and Descriptive terms, frequency and range. The Cluster frequency window shows a pie chart of the clusters by frequency.

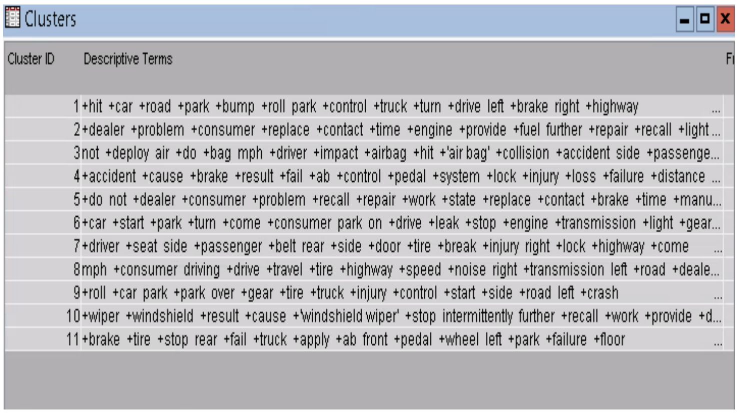


X = -1.76262
Y = 0.729321
Cluster ID = 6
Descriptive Term Sample = +car +start +park +turn +come ...

Figure 3.11: Distance between clusters

Here in the above chart, the distance between the clusters X and Y can be identified with Cluster ID and descriptive term sample. We have got 11 clusters and each described in the below table.





Cluster ID	Descriptive Terms	Fr
1	+hit +car +road +park +bump +roll park +control +truck +turn +drive left +brake right +highway ...	
2	+dealer +problem +consumer +replace +contact +time +engine +provide +fuel further +repair +recall +light ...	
3	not +deploy air +do +bag mph +driver +impact +airbag +hit +'air bag' +collision +accident side +passenge...	
4	+accident +cause +brake +result +fail +ab +control +pedal +system +lock +injury +loss +failure +distance ...	
5	+do not +dealer +consumer +problem +recall +repair +work +state +replace +contact +brake +time +manu...	
6	+car +start +park +turn +come +consumer park on +drive +leak +stop +engine +transmission +light +gear...	
7	+driver +seat side +passenger +belt rear +side +door +tire +break +injury right +lock +highway +come ...	
8	mph +consumer driving +drive +travel +tire +highway +speed +noise right +transmission left +road +deale...	
9	+roll +car park +park over +gear +tire +truck +injury +control +start +side +road left +crash ...	
10	+wiper +windshield +result +cause +'windshield wiper' +stop intermittently further +recall +work +provide +d...	
11	+brake +tire +stop rear +fail +truck +apply +ab front +pedal +wheel left +park +failure +floor ...	

Figure 3.12: Pie chart distribution of clusters *Figure 3.13: Cluster ID & Descriptive Terms*

Cluster ID	Description
1	In this cluster the descriptive terms say about a hit or accident that happened on the road/park which is caused by the driver not due to vehicle problem. This has 3 percent from the overall accident cases.
2	In this cluster it describes saying that complaint is raised due to the engine/oil leak/light or other repair issues. This plays a major part in the cluster chart by holding 22 percent.
3	This cluster shows that the injuries or deaths happened due to non-deployment of air bags in the passenger side. This holds 16 percent from the overall cases.
4	This cluster shows that the complaint is raised because of brake failure in the vehicle. This covers 6 percent from the overall cases.
5	This cluster explains that 6 percent of case or complaints because of dealer/ manufacturer problem
6	This cluster holds 1 percent of parking or transmission complaints.
7	10 percent of the complaints or injuries are due to seat belt issues.
8	About 5 percent of the overall complaints are due to wheel or tire problems.
9	About 4 percent of the overall problems are due to hit while parking the vehicle on the road.
10	10 percent of the overall complaints are because of windshield wipers
11	About 17 percent of the accidents are due to brake failures.

Figure 3.14: Description of all clusters

Predictive Models in EM (Page 1)

Screenshot of the overall predictive model

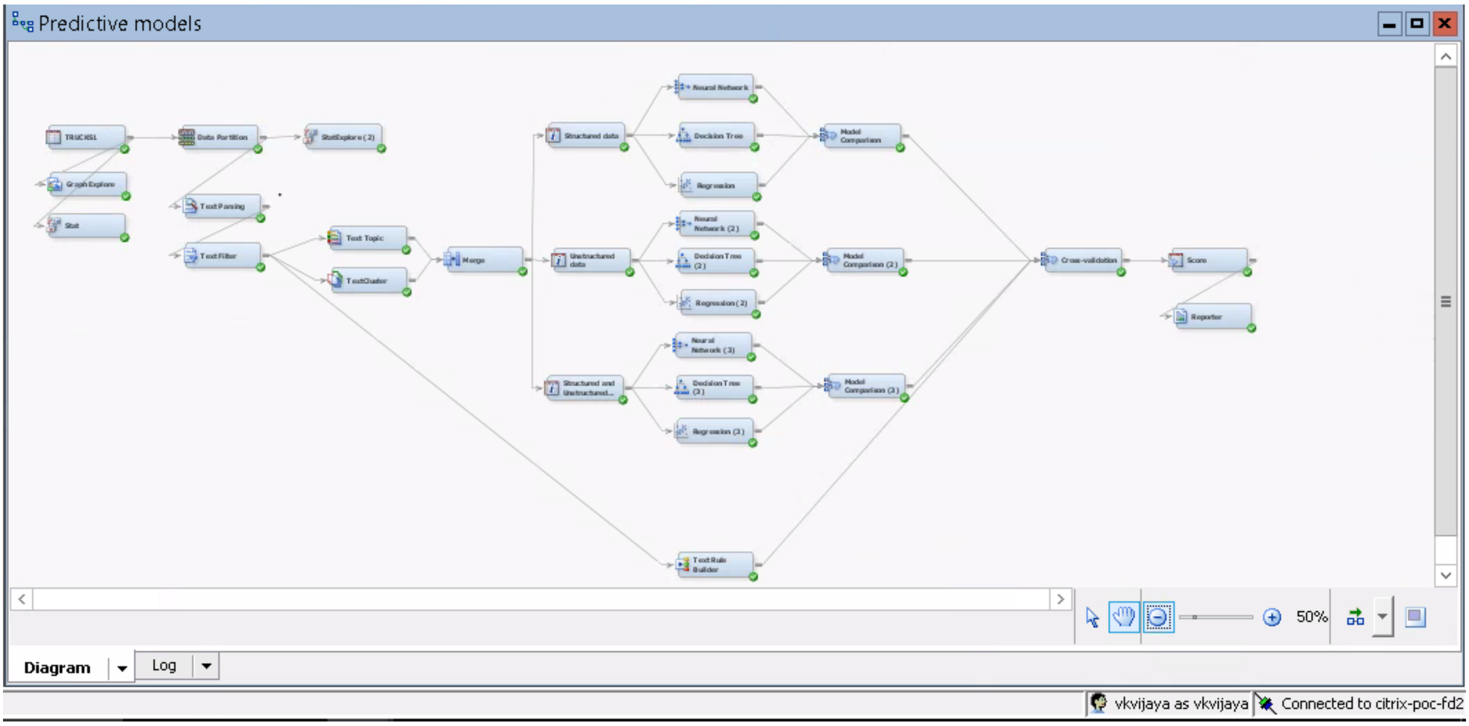


Figure 4.1: Overall predictive model

Structured data predictive model

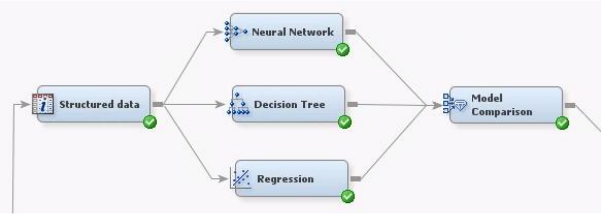
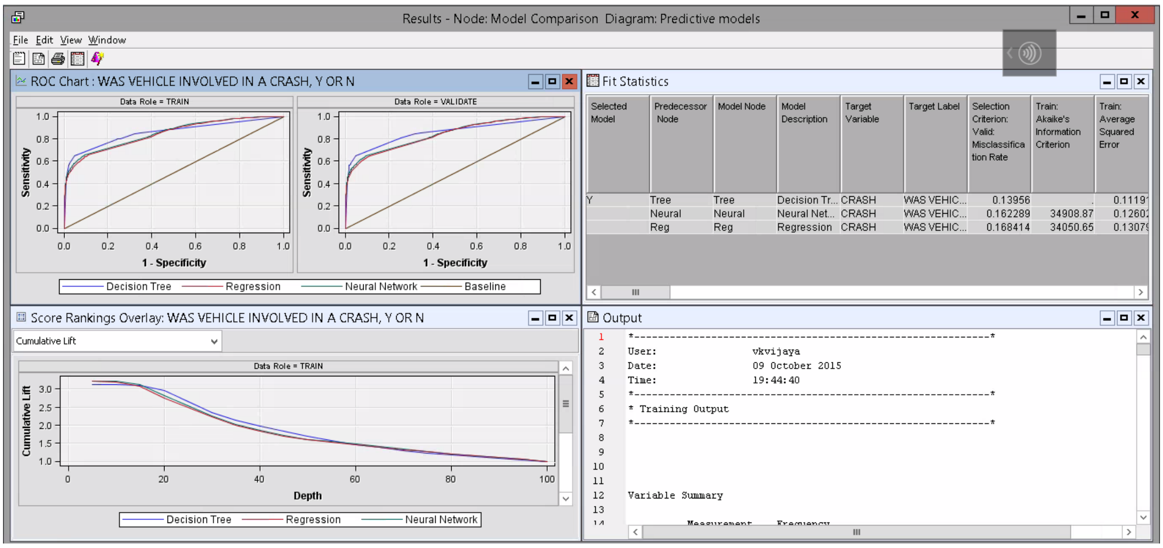


Figure 4.2: Structured data predictive model



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Tree	Tree	Decision Tr...	CRASH	WAS VEHIC...	0.13956	.	0.1119
	Neural	Neural	Neural Net...	CRASH	WAS VEHIC...	0.162289	34908.87	0.1260
	Reg	Reg	Regression	CRASH	WAS VEHIC...	0.168414	34050.65	0.1307

Output

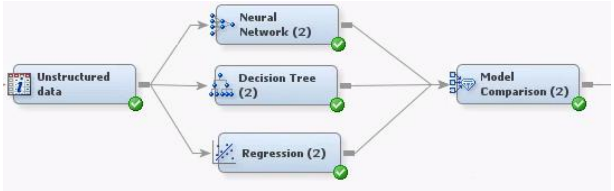
```
*------------------------------------------------------------*
User:                    vkvijaya
Date:                    09 October 2015
Time:                    19:44:40
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

           Measurement     Frequency
```

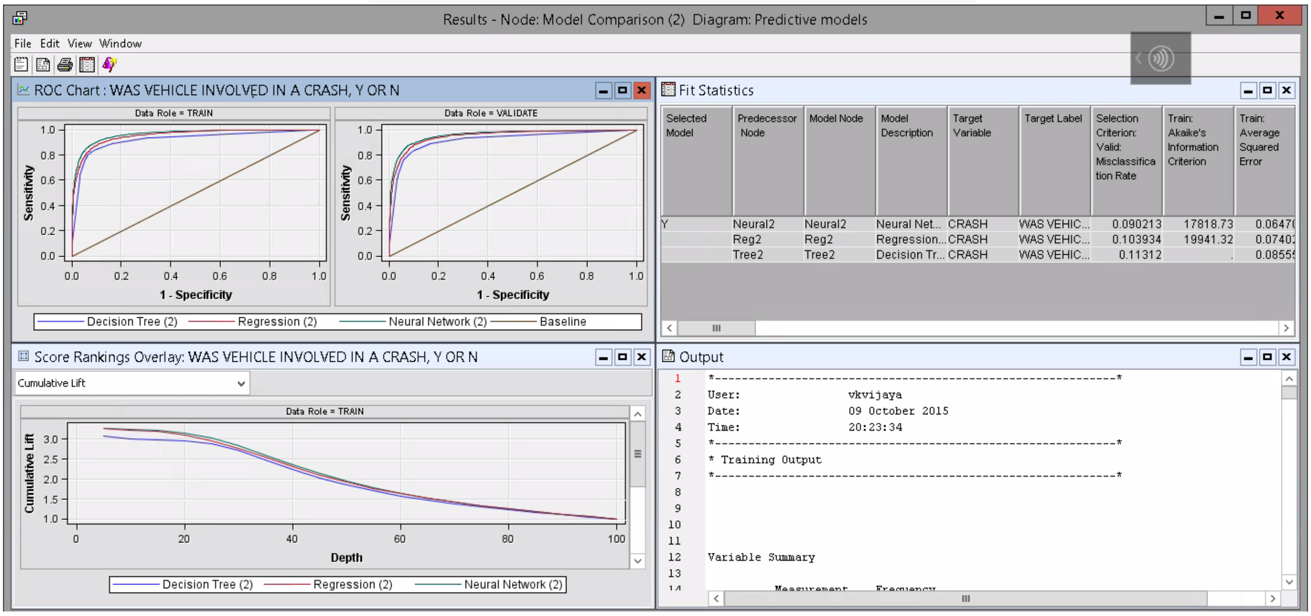
Figure 4.3: Model Comparison node results for the structured data (Decision Tree, Neural network, Regression)

Predictive Models in EM (Page 2)

For structured data, all columns excluding the CSUMMARY and the text topic clusters were excluded.
The ROC comparison chart shows high performance quality of the model and is indicated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified as the area under the ROC curve. The area under the ROC curve, or ROC Index, is summarized in the Output window of the Model Comparison node. From Figure x, we can see that the **Decision Tree** chart line is closest to the upward left corner making it the best predictive model of the three. Furthermore, the average squared error (ASE) of the decision tree model is 0.111912 which is the least of the three models. Both the other two models, Neural Network and Regression, have very similar characteristics and hence are not quite the best representation of structured data.
For the binary target variable, Crash, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order for each model.
The Fit Statistics displays the Decision Tree model as the champion model represented by the Y in the table making it the best predictive model to be represented as the structured data predictive model.

<u>Unstructured data predictive model</u>





Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Neural2	Neural2	Neural Net...	CRASH	WAS VEHIC...	0.090213	17818.73	0.0647
	Reg2	Reg2	Regression...	CRASH	WAS VEHIC...	0.103934	19941.32	0.0740
	Tree2	Tree2	Decision Tr...	CRASH	WAS VEHIC...	0.11312	.	0.0855

```
*------------------------------------------------------------*
User:                    vkvijaya
Date:                    09 October 2015
Time:                    20:23:34
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
```

Figure 4.4: Unstructured data predictive model results (Decision Tree, Neural Network and Regression)

For unstructured data, all generated text topic clusters were considered while excluding the remaining structured variables. The ROC comparison chart shows high performance quality of the model and is indicated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified as the area under the ROC curve. The area under the ROC curve, or ROC Index, is summarized in the Output window of the Model Comparison node. From Figure x, we can see that the ROC curve line for **Neural Network** model is the closest to the upper left hand of the ROC chart making it the best predictive model.
Furthermore, the Fit Statistics table shows Neural Network Node2 as the champion model and this is represented by the Y next to the node name in the table. Alternatively, Decision Tree model was comparatively the least effective predictive model for unstructured data. The average squared error for neural network is 0.064707 which is the least error margin when compared to the other two models.
For the binary target variable, Crash, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order for each model. Additionally, Score Rankings Overlay also shows neural network model as the most effective predictive model for unstructured data.

Predictive Models in EM (Page 3)

Structured and Unstructured data predictive model

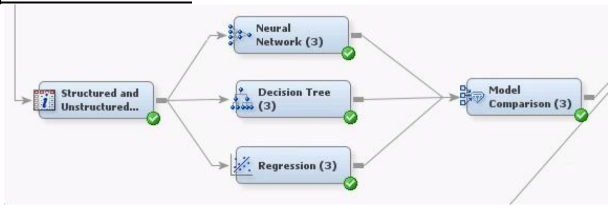
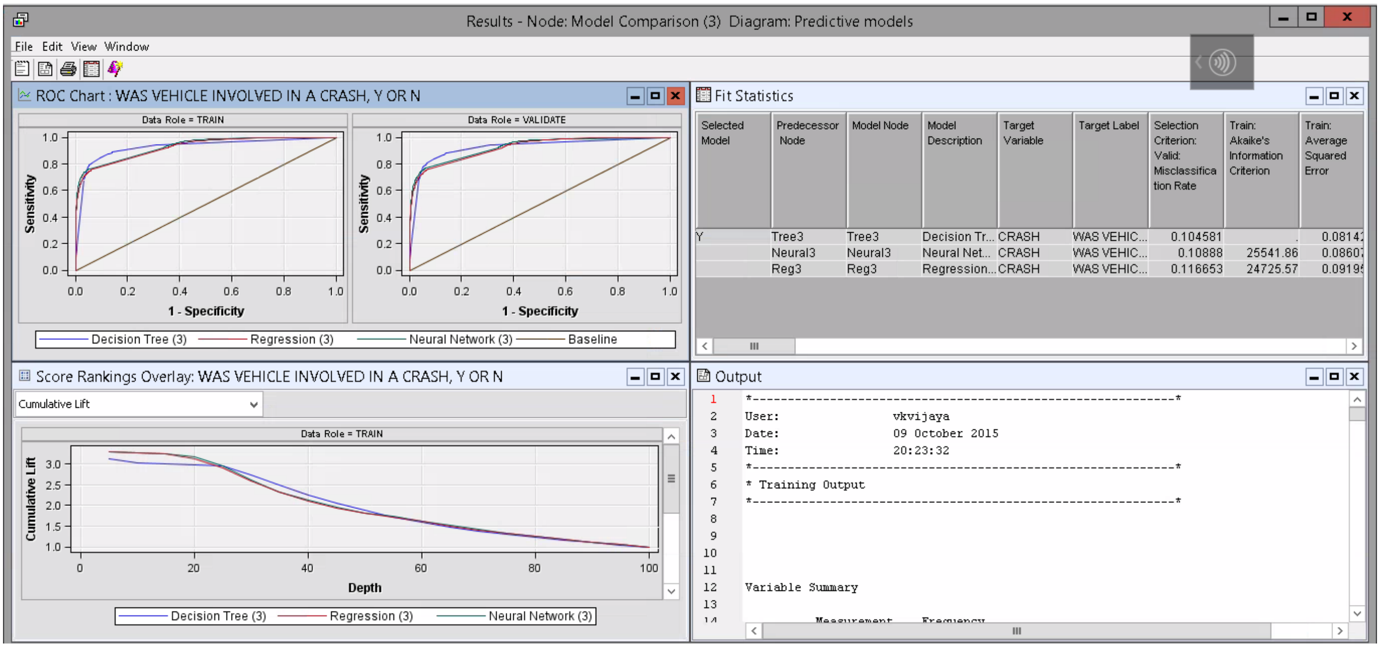


Figure 4.5: Structured and Unstructured data predictive model



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error
Y	Tree3	Tree3	Decision Tr...	CRASH	WAS VEHIC...	0.104581	.	0.08142
	Neural3	Neural3	Neural Net...	CRASH	WAS VEHIC...	0.10888	25541.86	0.08607
	Reg3	Reg3	Regression...	CRASH	WAS VEHIC...	0.116653	24725.57	0.09195

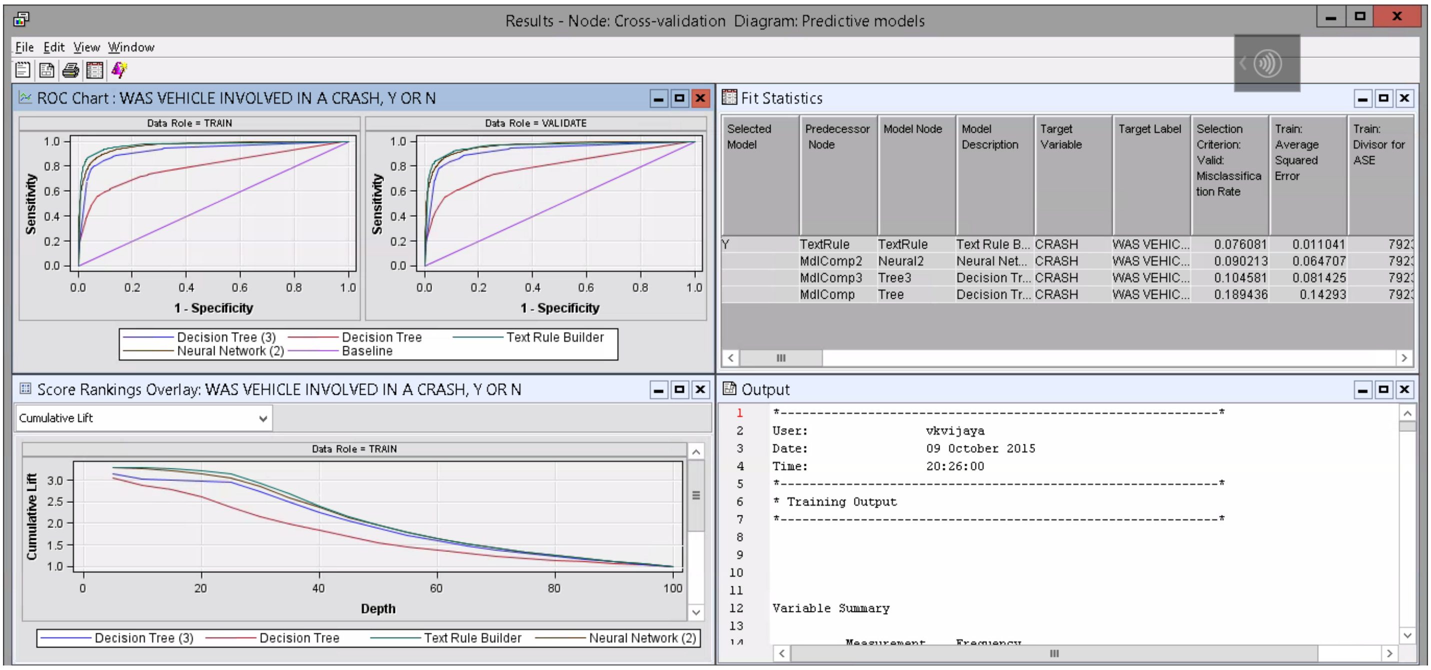
```
*------------------------------------------------------------*
User:                    vkvijaya
Date:                    09 October 2015
Time:                    20:23:32
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
```

Figure 4.6: Predictive model results for both structured and unstructured data (Decision Tree, Neural Network and Regression)

For both structured and unstructured data, all generated text topic clusters and structured variables were considered. The ROC comparison chart shows high performance quality of the model and is indicated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified as the area under the ROC curve. The area under the ROC curve, or ROC Index, is summarized in the Output window of the Model Comparison node. From Figure x, we can see that the ROC curve line for **Decision Tree** model is the closest to the upper left hand of the ROC chart making it the best predictive model.
Furthermore, the Fit Statistics table shows Decision Tree Node3 as the champion model and this is represented by the Y next to the node name in the table. Alternatively, Regression model was comparatively the least effective predictive model when considering both structured and unstructured data. The average squared error for decision tree model is 0.081425 which is the least error margin when compared to the other two models.
For the binary target variable, Crash, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order for each model. Additionally, Score Rankings Overlay also shows Decision Tree model as the most effective predictive model when considering both structured and unstructured data.

Model Comparison and Ensemble Models (Page 1)

Summary of the model assessment statistics over training and validation partitions

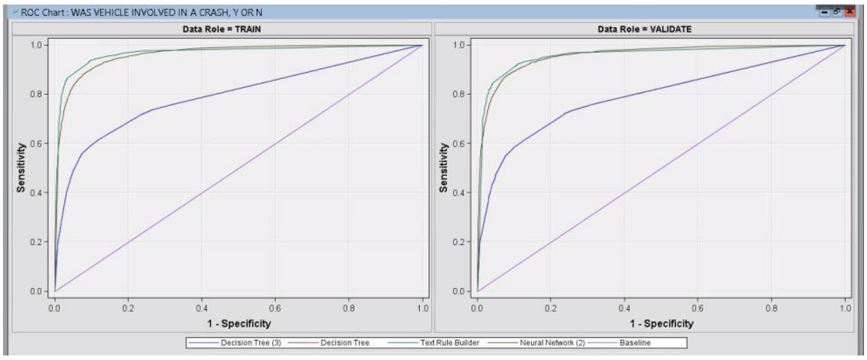


Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Train: Average Squared Error	Train: Divisor for ASE
Y	TextRule	TextRule	Text Rule B...	CRASH	WAS VEHIC...	0.076081	0.011041	792
	MdlComp2	Neural2	Neural Net...	CRASH	WAS VEHIC...	0.090213	0.064707	792
	MdlComp3	Tree3	Decision Tr...	CRASH	WAS VEHIC...	0.104581	0.081425	792
	MdlComp	Tree	Decision Tr...	CRASH	WAS VEHIC...	0.189436	0.14293	792

```
*------------------------------------------------------------*
User:                    vkvijaya
Date:                    09 October 2015
Time:                    20:26:00
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*




Variable Summary

          Measurement    Frequency
```

Figure 5.1: Cross-validation of the results of previous 3 types of model comparisons

On cross-validation of all the previous three types of model comparisons with text rule builder node results over training and validation partitions. The ROC comparison chart shows high performance quality of the model and is indicated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified as the area under the ROC curve. The area under the ROC curve, or ROC Index, is summarized in the Output window of the Model Comparison node. From Figure x, we can see that the ROC curve line for **Text Rule Builder** model is the closest to the upper left hand of the ROC chart making it the best predictive model which is just behind Neural Network model. Furthermore, the Fit Statistics table shows Text Rule Builder Node as the champion model and this is represented by the Y next to the node name in the table. Surprisingly, Decision Tree model was comparatively the least effective predictive model during cross-validation of the models with text rule builder node. The average squared error (ASE) for text rule builder is 0.011041 which is the least error margin when compared to the other three models.
For the binary target variable, Crash, all observations in the scored data set are sorted by the posterior probabilities of the event level in descending order for each model. Additionally, Score Rankings Overlay also shows Text Rule Builder model as the most effective predictive model from the cross-validation results.

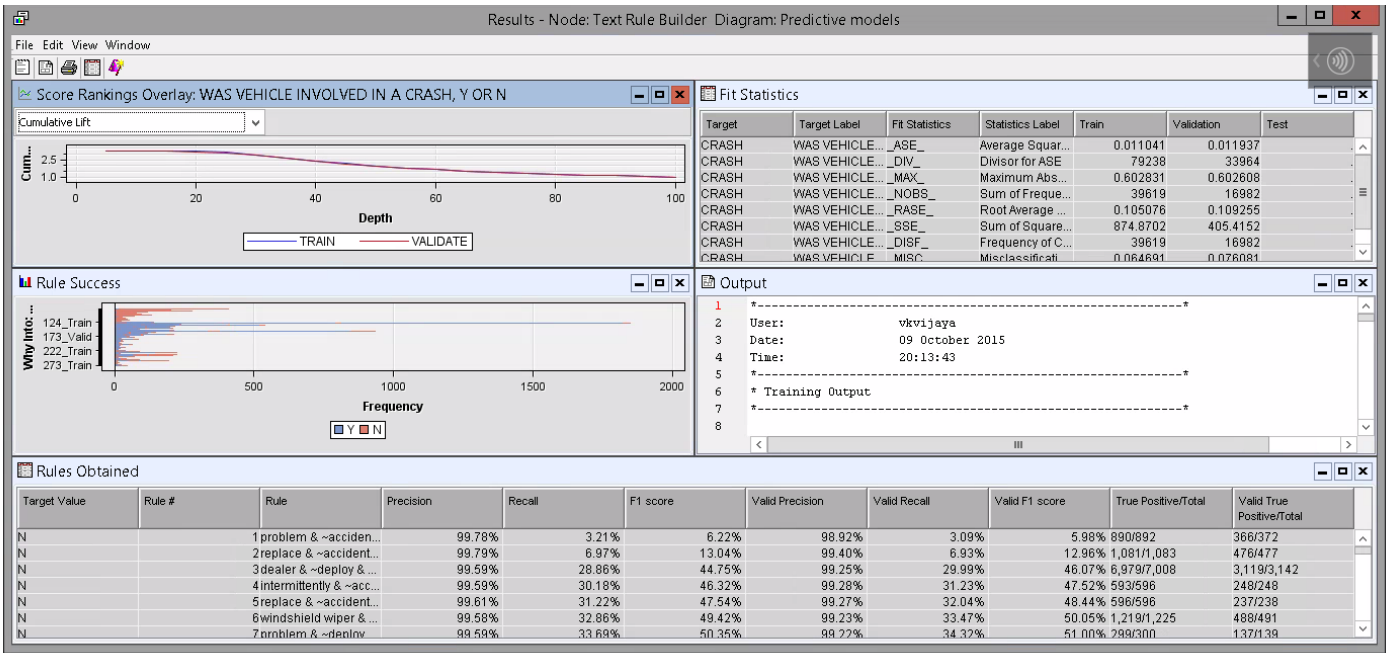


Area under curve (AUC) from ROC chart for the training chart is 0.967264 and 0.963281 from the validation chart. The AUC values are the closest to 1 when compared to the other three model results making it the best predictive model.

Figure 5.2: ROC Chart from text rule builder node result

Model Comparison and Ensemble Models (Page 2)

Best predictive model and summary of the model and its performance



Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
CRASH	WAS VEHICLE...	_ASE_	Average Squar...	0.011041	0.011937	.
CRASH	WAS VEHICLE...	_DIV_	Divisor for ASE	79238	33964	.
CRASH	WAS VEHICLE...	_MAX_	Maximum Abs...	0.602831	0.602608	.
CRASH	WAS VEHICLE...	_NOBS_	Sum of Freque...	39619	16982	.
CRASH	WAS VEHICLE...	_RASE_	Root Average ...	0.105076	0.109255	.
CRASH	WAS VEHICLE...	_SSE_	Sum of Square...	874.8702	405.4152	.
CRASH	WAS VEHICLE...	_DISF_	Frequency of C...	39619	16982	.
CRASH	WAS VEHICLE...	_MISC_	Misclassificati...	0.064691	0.076081	

```
*------------------------------------------------------------*
User:                vkvijaya
Date:                09 October 2015
Time:                20:13:43
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

Target Value	Rule #	Rule	Precision	Recall	F1 score	Valid Precision	Valid Recall	Valid F1 score	True Positive/Total	Valid True Positive/Total
N	1	problem & ~acciden...	99.78%	3.21%	6.22%	98.92%	3.09%	5.98%	890/892	366/372
N	2	replace & ~accident...	99.79%	6.97%	13.04%	99.40%	6.93%	12.96%	1,081/1,083	476/477
N	3	dealer & ~deploy & ...	99.59%	28.86%	44.75%	99.25%	29.99%	46.07%	6,979/7,008	3,119/3,142
N	4	intermittently & ~acc...	99.59%	30.18%	46.32%	99.28%	31.23%	47.52%	593/596	248/248
N	5	replace & ~accident...	99.61%	31.22%	47.54%	99.27%	32.04%	48.44%	596/596	237/238
N	6	windshield wiper & ...	99.58%	32.86%	49.42%	99.23%	33.47%	50.05%	1,219/1,225	488/491
N	7	problem & ~deploy	99.59%	33.69%	50.35%	99.22%	34.32%	51.00%	299/300	137/139

Figure 5.2: Results of the Text Rule Builder node

The Score Rankings Overlay shows very similar training and validation data lines.
From Rules Obtained table, True positive column shows the number of crashes that were assigned to the keywords from the rule column. The Total shows the total number of positives. The Remaining positive shows the number of total remaining entries in the dataset.
From rule number 1, we see 366 entries have the target value of 'N' which implies no crash of the vehicle of the total 372 entries. It has an estimated precision of 99.78% and a valid precision of 98.92% and includes a large number of multiple keywords under this rule listed and certain keywords that do not appear in the entry.
While from rule number 125, 135 out of 139 entries have the target value of 'Y' which implies crash of the vehicle. Rule number 125 shows frontal collision as the solo keyword which has an estimated precision of 99.22% and a valid precision of 97.12%.

Text Rule Builder model was further improved using the settings to Change Target Values property. Initially it was run at low settings for the Generalization Error, Purity of Rules, and Exhaustiveness properties. The results were analyzed and then run at medium settings and then analyzed again.
The results at medium were found to be more optimal than at low settings and hence were then saved.

To summarize, a predictive model using Text Rule Builder node would be the best predictive model for the given situation. Text Rule Builder consistently has the lowest ASE value and the best results in the ROC chart as seen from the cross-validation results.