Data cleaning in Excel file

Each attribute in each dataset have been thoroughly studied and determined if it is needed or not in related with the objective of this project. Attributes which are used for administrative purpose like vehicle plate, driver license, Municipality code, Area code, Rout No and other irrelevant variables are removed from the dataset that reduced the attributes by significant number. Then, formatting for some attributes have been changed, for example, birth year of the driver "03-10_1970 "was changed into 1970 and name of some attributes are changed as well to understand easily, for example, RD_DIV to ROAD TYPE, CV BODY to VEHICLE BODY. More than 90% of the attributes data entries are coded (01, 02, ….88, 99) and understanding each code helps for the data cleaning purpose, for example, 88 = "OTHER" and 99 = "UNKNOWN". Before the data set was imported to SAS EM, all UNKNOWNS were changed to NULL to make it easy for later cleaning in SAS EM. Some codes are written in decimal, for instance, in vehicle information Body Type code 27.88 = "Snowmobile" changed into 27. Categorical Target Variable (Report Type) was changed into numeric factor by assigning "1" for occurrence of fatal/injury accident and "0" for car damage in planning of using logistic regression as a binary outcome (Appendix 1)

In most of reported accidents there were more than one person involved in the car accident, for example, children or passengers who got killed or injured and registered on personal information dataset with the same Report No. Here, the goal of the project is to identify the cause of the car crash due to the condition of the driver, and all passengers are excluded from the analysis. This was

identified by “Person Type” attribute (Driver, Occupant, Pedestrian). In this project exclusion of Fatal/injury accident on Occupant and Pedestrian can be considered as a limitation.

Creating of a Variable

Data set combination Techniques and Preparing SAS TABELE

The three data sets were imported from Excel file to MS-SQL Database and combined by using INNER JOINT SCRIPT with a common variable of Report No (Appendix 4). After merging the datasets, 26,660 rows and 32 variables were exported to excel file and then imported again into SAS EM library to create SAS Table. The property of SAS Table was set up by assigning the Role of each attribute (Target, Input, ID) and choosing the correct level (Binary, Nominal, Interval). For example, some categorical attributes like County Codes, Accident Time were labeled as interval Variable and edited to Nominal. Out of 32 variables, four of them are interval variables (Speed limit, Reference _Distance, Longitude, and Latitude), Report Type (Target variable) is a Binary variable and all the rest are Nominal (Appendix 5). Advanced Option was used to set up missing percentage Threshold (Default was used where variables with more than 50% of missing value were rejected) and 100 class levels count Threshold was used to avoid rejection of some variables with many classification, for example, car make, county Code, Vehicle body made, accident time, Date of Birth). Out of the total Reported accident 0.27% of them are fatal/injury with Prior probability of 0.2644 for fatal/injury and

0.7336 for car damage. In order to increase the chance of catching rare events adjusted equal prior probability (0.5) was applied and a decision weight of 0.99 for fatal/injury and 0.01 of car damage was used in a sense that fatal/injury event is 99 times more important than car damage. Finally, 25% of the observation are considered to create the Data Source (SAS Table).

Initial Exploration was used to visualize the input data helps to observe possible pattern and check missing value before building a model. The result of the exploration showed the percentage of missing value, Min, Max, Mean, SD, Skewness, and Kurtosis values. The Min and Max values helps to see if there is any outliers, for example speed limit with a min value of 0 and max value of 75 and Reference distance with min and max value of 0 and 75 are found in acceptable range Four variables had a missing value of more than 50% and rejected from being used in the analysis. These variables are Drug Test, Movement code, occ-seat with a missing value of 99%, 96% and 96 %, respectively. The Target variable, Accident Time, Airbag Deployment, Longitude, Latitude, Speed limit have 0 missing value and the rest have missing value with the range of 1.89% -12.62%.

The Stat Explore Node also showed A chi-square plot orders the top 20 variables by their chi-square statistics and a variable worth plot orders the input variables by their importance in predicting target variable (Appendix 6). The stat Explore also provided the summary statistics for class and interval variables. Among the four interval variables, Distance has the highest SD which is more than 100 (146) with high variability (Appendix 7). Variable Clustering Node was run to see the correlation between interval variables and found out that there is high correlation between them, and all are kept in the analysis (Appendix 8).

Replacing and Imputation

Before replacing unknown value and outliers by missing value, data partition (Training dataset of 60%, Validation data set of 30% and Test dataset of 10%) was done. Interval Variables don't have any missing value. With Replacement Node, all class variable and interval variables with unknown values are replaced with missing value (Appendix 8). Then Impute Node has been run to replace all missed class value with the most occurrence event (Mode). For example, Airbag has 400 missing value and imputed with a value of code value of 1 because 63.69% of Airbag deployment was responded on the questioner as code 1.The second class variable with the highest missing value is Alcohol Test code and imputed with code 0 as 92.75% of response were code 0. In general, for bimodal variables, the highest mode was taken as an imputed value for missing value. (Appendix 8).

Creating Additional Variables/ Data Transformation

Rate of Accident

A variable name "Rate of Accident" has been created from Population size in each state in Maryland and total number of accident in each state. Rate of Accident data was calculated by dividing the number of fatal/injury accident by the total number of population in each state. The same thing was done for car damage in each state and divided by the total number of accident to get the rate. This was done by assuming in all state starting age of driving, socioeconomic status and demographic behaviors of the people are the same. However, taking assumption and creating this variable is much more important than counting the number of accident in each state and deliver this statistics to safety planner, insurance company and health organizations. For example, in Baltimore the population is much greater than any other county of the state and getting more car crash than any small populated county. This variable especially very helpful for visualization.

Variable Transformation

Variable transformation for skewed distribution is needed to fulfill the assumption of Normality to build Logistic Regression model. After the missing values are imputed, variable Transformation was done by selecting Transformation Node from SAS EM. The distribution of all variables are shown on (Appendix 5) and the type of transformation is selected according to the distribution of the variable. Latitude, Longitude, Speed limit, Date of Birth have close normal distribution. Some of the variables have pick at the left

or right and flat in the middle which means most of the values are concentrated at the end. For example, person condition has the highest pick at the right (code 1), movement of the person, Airbag deployment, Alcohol test are positively skewed. All the rest of the variables have also negative or positive skewness which violates Normality and needs transformation. In order to approach Normality the log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

The best thing about SAS EM, whenever transformation or imputation are conducted, it keeps the original variable and create a new feature. However, the status of the original variables are changed into rejection. After transformation, the new feature will have less standard deviation or variability and pretty much enough to be representative.

Data Analysis

Tableau is going to be used for visualization which is one of the most important part of data analysis to present the data in an understandable and visually appealing format. In this project, various types of visualization will be provided for end users. For example, fatal /injury data by county, Airbag deployment, age, name of road, in addition, the most interesting part of tableau is providing geospatial information from longitude and latitude variables. The below screenshot showed the exact place where accident was occurred, and the red dots indicates the place where fatal/injury accident occurred. In the map, on Stone Rd in 2016 there were more fatal/ injury than other places. This helps to Maryland safety Highway office to plan and implement on the right

time to reduce accident. More than two variables can be included to get more detailed information about the accident like the time, the driver condition, the type of road and so on. In general, Tableau does a lot of visualization and presenting reality with a simple bar graph, pie chart, line graph, maps, so any stakeholders, business organizations, individual can grasp the information very easily.

SAS EM is the right tool for statistics analysis and in this project, it will be used for model development. SAS EM is the best as it does all the data preparation, cleaning, transformation, model development, evaluation and comparison.

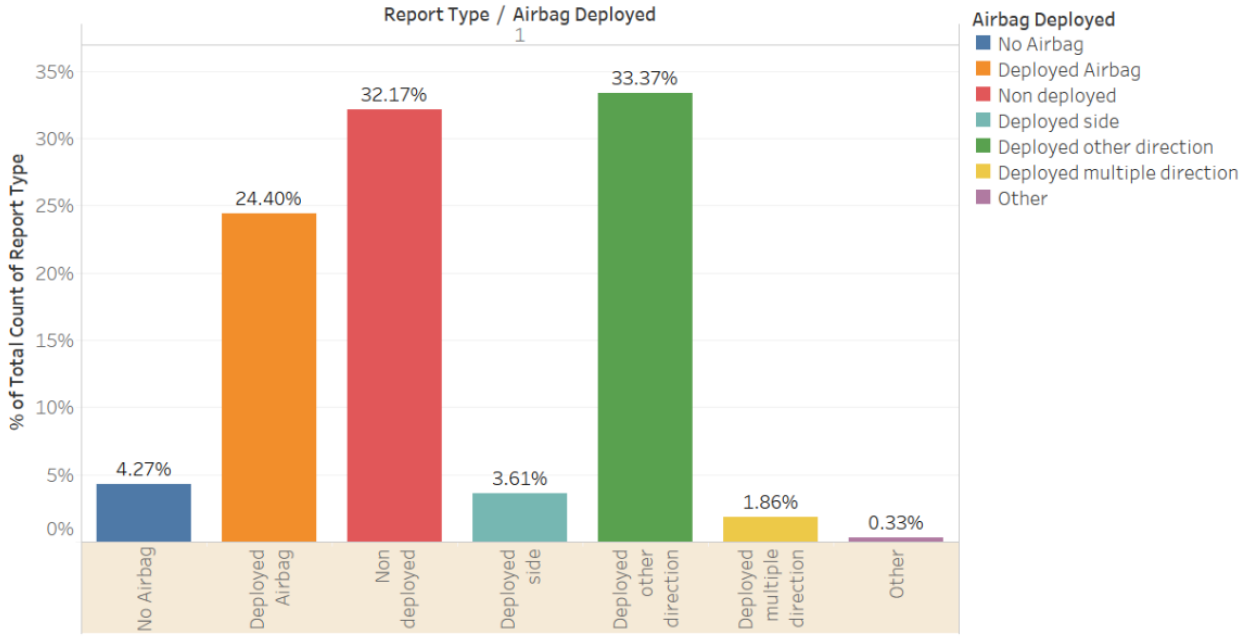
Data Visualization

Visualize large amounts of complex data is easier than presenting spreadsheets or reports as it is easy way to convey concepts in a universal manner. In this project various visualization are prepared in related with fatal/injury car crash in Maryland to answer basic questions like in which county is the highest death due to Car crash? What age groups are mainly involved in Car accident? What is the impact of Air bag performance at the time of accident on the rate of fatal/injury accident and other similar questions?

Photos - Fatal_injury Car Crash Report by Airbag Deployment.png

See all photos | Add to a creation | Edit & Create | Share

For Destroyed Car Damage the Number of Fatal/injury Car Crash by Airbag deployment



% of Total Count of Report Type for each Airbag Deployed broken down by Report Type. Color shows details about Airbag Deployed. The data is filtered on Damage Code, which keeps Destroyed. The view is filtered on Report Type and Airbag Deployed. The Report Type filter keeps 1. The Airbag Deployed filter excludes Null. Percents are based on each row of the table.

The purpose of the air bag is to provide a cushion between the occupants and the vehicle's interior. For air bags to be effective they must be fully inflated in a short amount of time, before the occupants contact them (Jesse, 2018). The above Visualization 1 showed the Percentage of fatal-injury car crash by Air bag deployment performance at the time when the car is damaged. One of the car accident tragedy happens when the car is destroyed. The question is what was the performance of the Air bag at the time of this tragedy? Was the Air bag deployed on the right time and on the right direction? or was the Air bag failed? Or was the Air bag deployed in different direction? Here, it can be possible to see the effect of Air bag deployment by selecting only one type of Car damage which is "Destroy".

The row and the column of the bar graph shows the percentage of fatal/injury and the performance of Air bag deployment. As it can be seen from the graph, out of the total fatal/injury Car crash in knowing that the Car was "destroyed", 33.37% of the fatal/injury was contributed by poor performance of Air bag with a code of "Deployed in other direction". The second highest fatal/injury (32.1%) was contributed again by another type of poor performance of Air bag with a code of "non deployed at all". In contrast, the lowest fatal/injury car crash was happened with Air bag performance of code "Deployed multiple direction "which is only 1.86% of fatal/injury happened out of the total death of car destroy. The second lowest death was showed when the Air bag was deployed in one direction with a percentage of 3.61%.

However, one of the columns of the Air bag deployment (Air bag deployed) showed unexpected result of fatal/injury car crash with a contribution of 24.4% which was considered the highest as far as the air bag was deployed. The reason might be different in such

a way that the performance is not only measured because It was deployed. However, other questions like does the Air bag deployed on time? Or does the Airbag deploy with a perfect inflation? Moreover, other reasons also affect the performance.

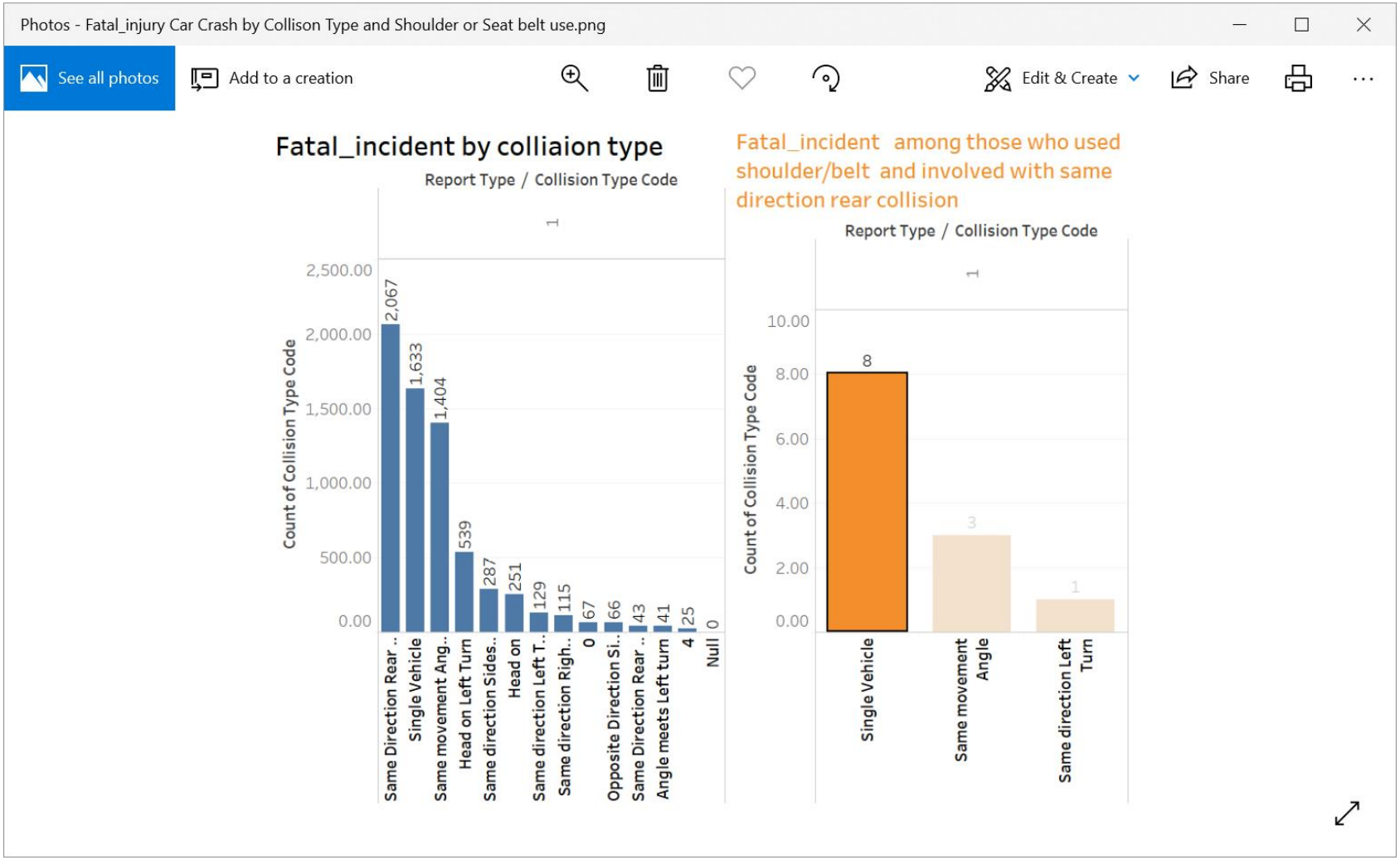
In general, the insights gained from the visualization is that the performance of Air bag deployment plays a significant role in saving the life of people even at the time of big accident tragedy. Good Performance of Air bag is a combination effect of various factors such as deployment time, the size of the inflation, and the direction. Form this result Car Manufactures can get good insight to assess or evaluate the performance of their car by combining the result and their long-time experience in the domain. Car insurance company also get some knowledge about the type of car their customers are driving in calculating the risk of selling their insurance with respect to the type of car with the history of Airbag deployment performance. Least but not last, each and everyone who wants to buy a car search up the performance history of the car.

Earlier, in this project, the performance of Air bag deployment was proposed to be one of the important factors to determine the fatal/injury Car Crash. The result from the visualization also supported the scope of this project which confirms to move forward and develop models by taking Air bag performance as one of the most effective variables to train the model. However, the gap of information of the

One of the lessons that was gained from this result is that detailed information is the base for the quality of analysis. For example, the result from the column of Airbag deployment with a code of “Airbag deployed” resulted a high fatal/injury percentage. This

implies that the question must go further in order to address the problem of the data quality by going deeper and asking questions like the time the Air bag deployed during the accident or the size of the inflation to protect the occupant.

Data Visualization 2: Fatal/injury Car Crash by Collision Type and using Safety equipment



Using safety equipment like Lap or shoulder belt are the most effective way to prevent fatal/injury Car crash. The above visualization displayed two bar graphs. The first one with blue color showed the number of the fatal/injury Car crash with the type of collision. The result revealed that the highest fatal/injury Car Crash was happened when the type of Collison was “same direction with rear end” with 2067 death. The second highest death was happened with a collision of a car with some other object like building, pole or something external other than another car with 1633. The third and fourth highest number of deaths appeared with a collision type of same movement angle and head on left turn with 1405 and 519, respectively. In contrary, the lowest death was observed with same direction sideswipe and angle meet left turn with a total fatal of 43 and 41, respectively.

In order to see the effect of wearing lap/shoulder belt on fatal/ injury Car Crash, the second bar graph was done by taking the number of fatal /injury car crash and experience of wearing safety equipment and filtered by “lap/shoulder” and applied to all type of collusion. Surprising, there was no one who died due to “same direction with rear end Collison” accident because of the safety equipment. The column was reduced into three with the lowest death in all. Over all only 13 people died even though they wear lap/shoulder belt. The total number of deaths by “single car” crash was 1633, among these deaths only 8 people with belt were died. This result proved that even though the accident is serious if a person wear lap/shoulder belt, the chance of being killed by accident is very rare. “Five children were killed, and two adults were injured in a Maryland car crash, officials said the children were not wearing seat belt, but the driver and the passenger were wearing seat belt and were saved” CNN, 2019).

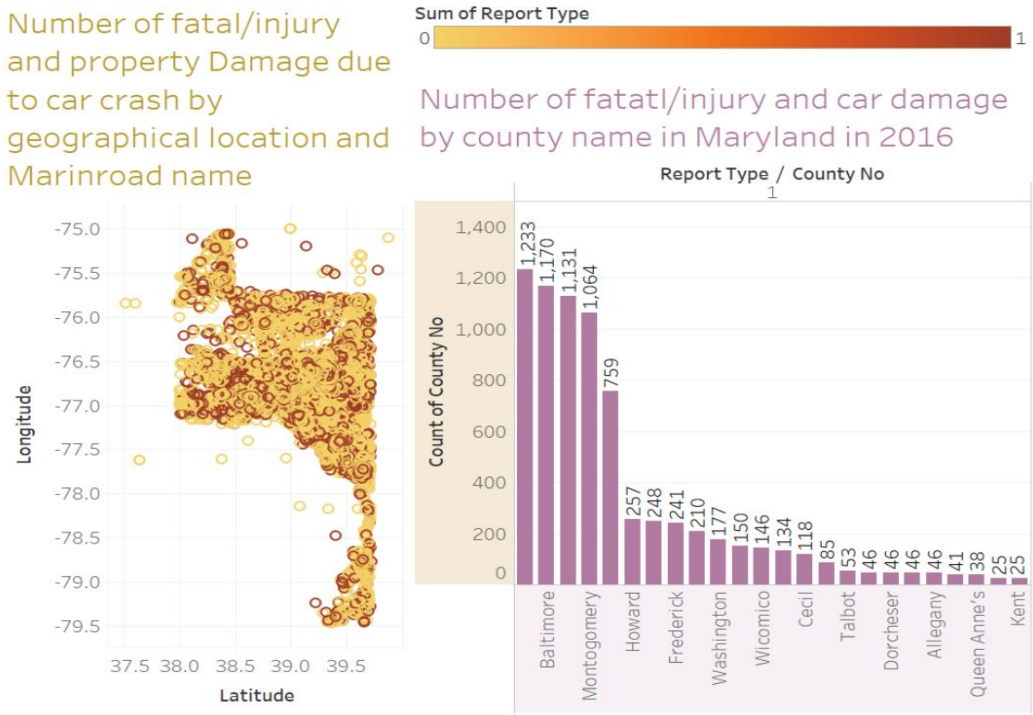
The insight gained from this result is seat belt prevent fatal car crash in a higher chance. The result of the visualization helps to add the attribute “safety equipment” as one of important variable to train the various model (logistics, SVM) to increase the accuracy of predicting fatal/ injury car crash in Maryland.

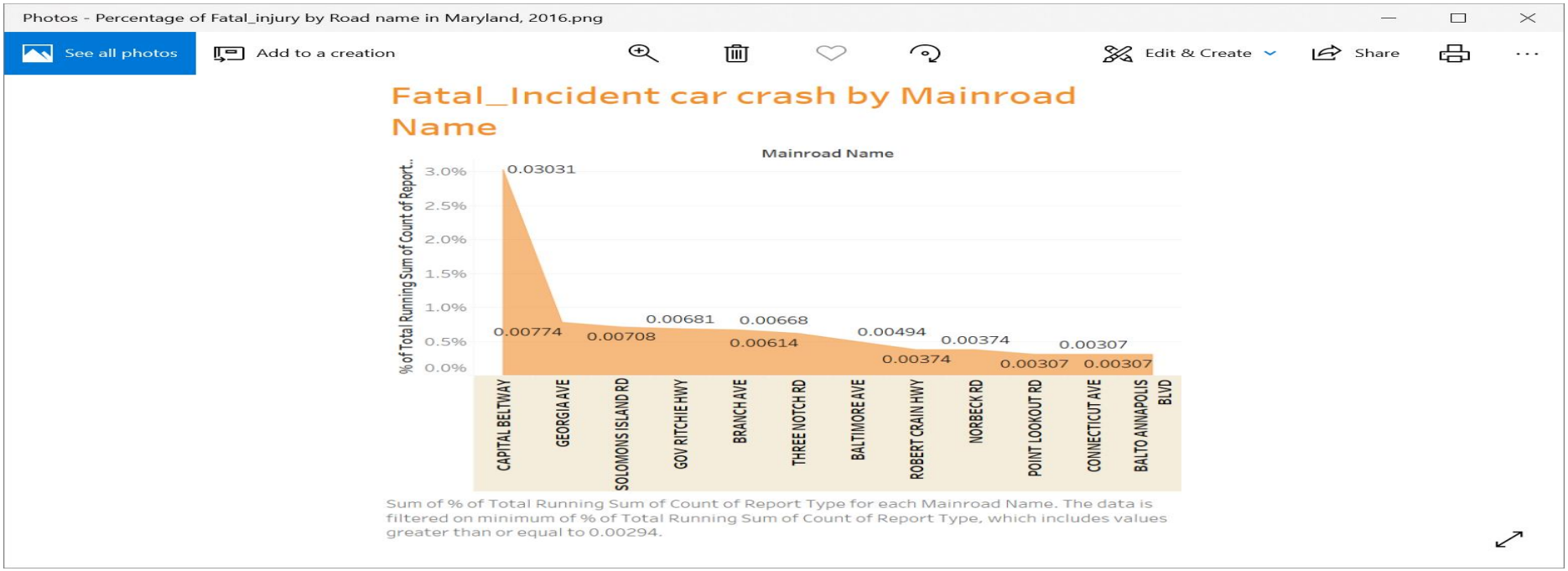
From the point of business objectives, SHSP get knowledge from the result to plan and implement to approach their goal in reduction of death due to car crash. This information also helps Stakeholders like MVA, and other partners can depend on the result to act on the rule of seat belt. For example, car and life insurance can get insight how wearing seat belt can save life of their customer and indirectly reduce the expense of their business. In this aspect, insurance company donate every year to SHSP to the implementation of the rule such as dispatching more officers on the road for the enforcement of seat belt, “Click it or Ticket”.

Data Visualization 3: Fatal/injury Car Crash by County and by Road Main

Photos - Numbr of fatal_injury by County Name and Road Name.png

See all photos | Add to a creation | Edit & Create | Share





The above dashboard displayed two visualization. The Left side shows geographical location of Maryland Roads with fatal/injury accident. On the map, each red dot indicates fatal accident on that specific road which is more red color on one spot means more accident was occurred on that road. In order to see clearly the name of the road with high frequency of accident, another visualization was added below the dashboard which shows the name of Maryland Roads with highest death. Capital Belt Road contributed 3% of

fatal/injury Car Crash in Maryland in the year of 2016. The next highest death or injury happened on the road of Georgia Ave and Solomon Island Rd, with 0.77% and 0.7% respectively. Many fatal/injury car crashes were also observed on the road of Branch Ave, Three North, Baltimore Ave, Robert Ave.

On the Dash board, the second visualization on the right side revealed the distribution of fatal/injury accident by county. Baltimore was the highest with a death of 4087 followed by Baltimore City with 1233 fatal/injury car crash. The next most county with highest death/injury were Prince George's, Montgomery, Anne Arundel, Howard and Harford with death/injury of 1131, 1964, 759, 257, and 248, respectively.

The insight gained from the above result with respect to the project scope, road name can be one of the important variables in predicting the fatal/injury Car crash. Earlier, this attribute was not proposed as a significant variable and now it will be considered in the variable choice in building various models.

From the business objective point of view, the name of the county and road name provided the place where most fatal/ injury accidents occurred. Organization like SHSP who is responsible for planning and implanting car crash in Maryland can use this insight as a starting point to make further research why more accident happened on these places and finding out the reason helps to take the right decision and action.

Proposed Visualization 1: Driver condition by Report type of fatal/injury

The first proposed visualization is displaying driver condition by fatal/injury accident. Driver condition states the physical and mental condition before the accident. The condition code includes “Apparently Normal”, “Had been drinking”, “using drugs”, “Physical Defects”, “other Handicaps”, “Apparently Asleep”, “Emotional depressed or angry”, “influenced by medication or drug”. Alcohol or drug impaired driver is more exposed for fatal/injury accident because the mind of abnormal person can’t process information on time. For example, drunk person is not able to keep enough distance from the vehicle in front of him or press the break on time to avoid any Collison which can be prevented. The same thing with apparently Asleep driver or fatigued person can easily engaged into accident because of failure to control himself.

The number of fatal/injury accident helps SHSP to make the necessary enforcement and rule to save the life of the driver and others. In addition, if this visualization is sorted by age and sex, it will provide good estimate to focus the rule and regulation on these types of people. Car insurance and life insurance can make risk analysis of taking these group of people as a customer or set the sell of their insurance.

Proposed Visualization 2: Road Type by Fatal/injury

The type of the road in Maryland contribute for fatal/injury accident. Looking into the distribution of Road type by fatal/injury gives a lot of insights. The type of roads in the data description are “Two-way Trafficway”, one-way Trafficway”, “Two-way Divided “, “Two way not divided” and many. The display helps the responsible organizations to reconstruct the road to reduce accident.

Predictive Model on Fatal/injury car crash in Maryland, 2016

Intended predictive model is the one which classify the target variable with a higher accuracy rate of sensitivity and overall accuracy. Classification is one of the major problems that we solve while working on standard business problems across industries (Lalit, 2015). In this project, 12 models will be developed including Ensemble model and compared based on ROC value of each models. Before developing the model, the imbalanced Target variable was handled.

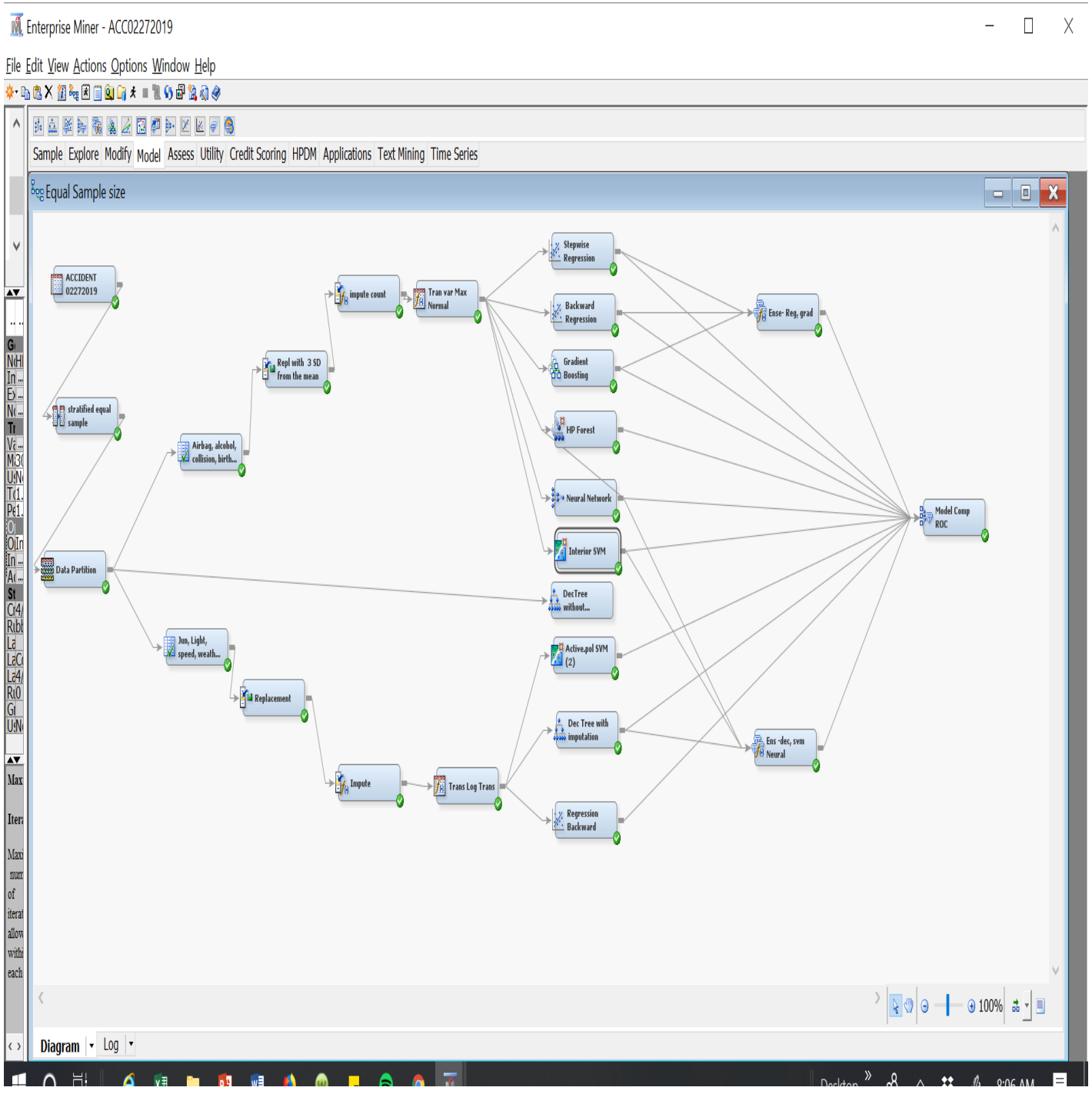
Handling Imbalanced Target variable

The prior probability of the target variable is 0.26 (Fatal/injury accident) and 0.74 (Car damage) which is not deadly imbalanced. However, in order to accomplish the business objective which is reducing death due to car accident towards zero, increasing the proportion of the rare event(fatal/injury) would be the best way by using equal sample size in both groups. This was done by using Sample Node in SAS EM and selecting stratified sampling method with equal sample size from both group of target variable, 5241 for each group.

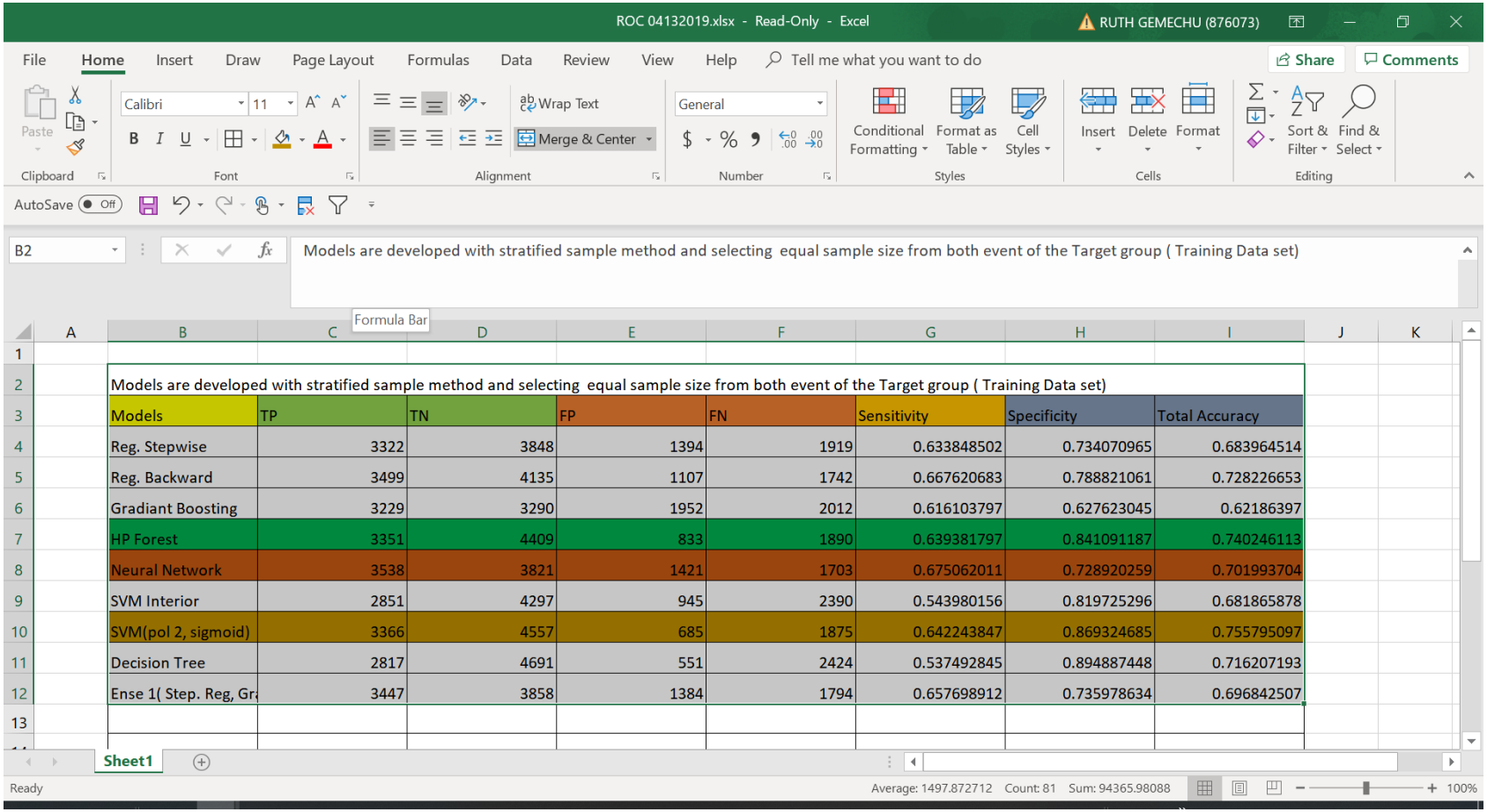
The dataset allocation was done by partitioning the data into training, validation and test dataset with a proportion of 70%, 30% and 10% of the whole sampled dataset. Training set is used for learning to fit the parameters of the classifier and Validation set use to tune the parameters of a classifier and set to find the “optimal” number of hidden which uses to validate the model. Finally, the model was tested with a test dataset.

Altogether, 13 models have been built and compared by ROC value. The top three models are HP Forest, SVM (Active set with pol 2) and Neural Network. The below screenshot showed each step which was used to build the model. In addition, the tables showed the calculated total accuracy rate, sensitivity, specificity and ROC value for Training, Validation and Test dataset. Detailed models properties will be explaining for the three champion models

Models Development

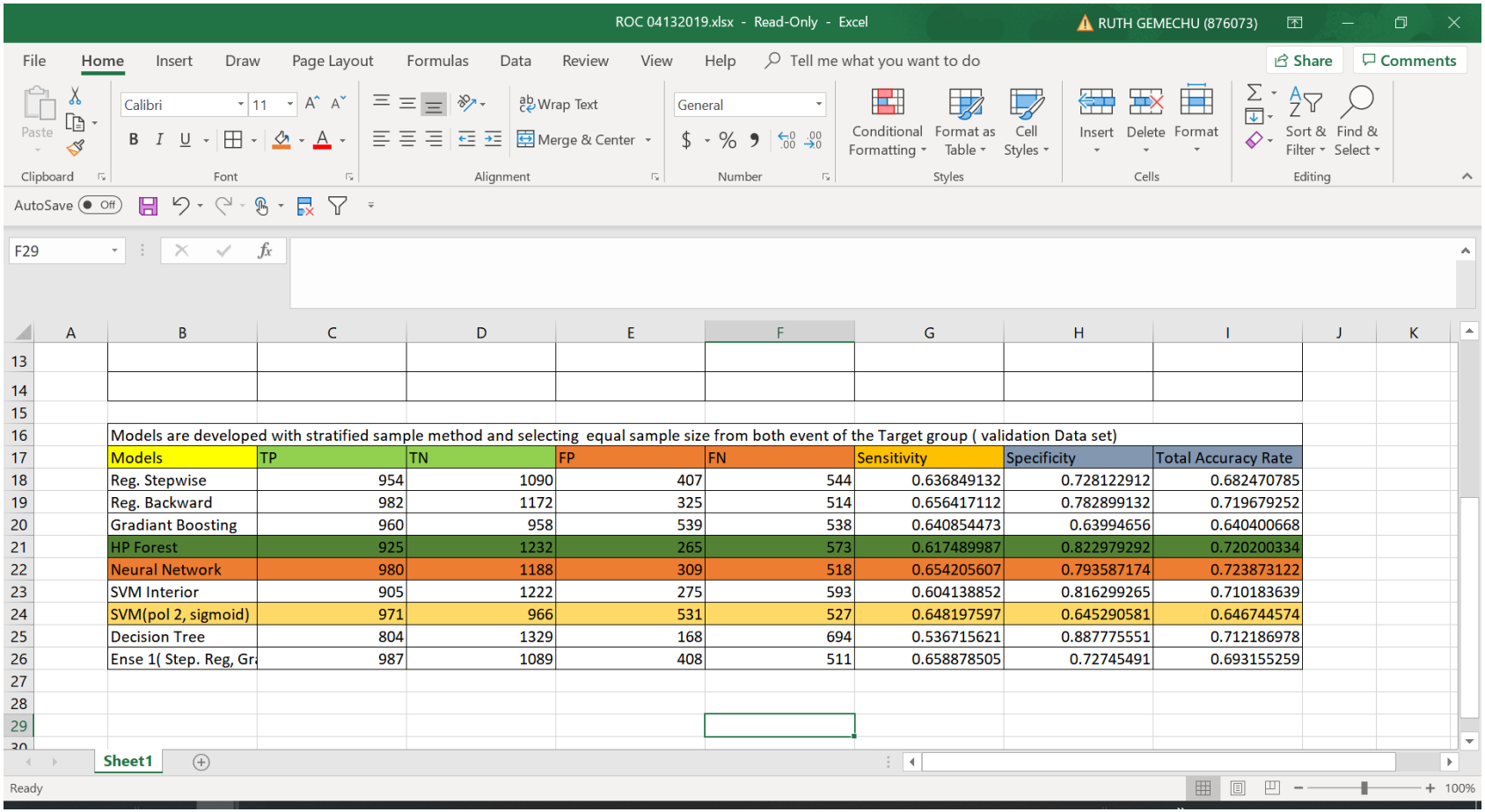

Enterprise Miner - ACC02272019
File Edit View Actions Options Window Help
Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series
Equal Sample size
ACCIDENT 02272019
stratified equal sample
Data Partition
Airbag, alcohol, collision, birth...
Repl with 3 SD from the mean
impute count
Tran var Max Normal
Jun, Light, speed, weath...
Replacement
Impute
Trans Log Trans
Stepwise Regression
Backward Regression
Gradient Boosting
HP Forest
Neural Network
Interior SVM
DecTree without...
Active,pol SVM (2)
Dec Tree with imputation
Regression Backward
Ense- Reg, grad
Ens -dec, svm Neural
Model Comp ROC
100%
Diagram
Log

Calculated value of Sensitivity, Specificity and overall Accuracy Rate for Training Dataset



Models are developed with stratified sample method and selecting equal sample size from both event of the Target group (Training Data set)							
Models	TP	TN	FP	FN	Sensitivity	Specificity	Total Accuracy
Reg. Stepwise	3322	3848	1394	1919	0.633848502	0.734070965	0.683964514
Reg. Backward	3499	4135	1107	1742	0.667620683	0.788821061	0.728226653
Gradiant Boosting	3229	3290	1952	2012	0.616103797	0.627623045	0.62186397
HP Forest	3351	4409	833	1890	0.639381797	0.841091187	0.740246113
Neural Network	3538	3821	1421	1703	0.675062011	0.728920259	0.701993704
SVM Interior	2851	4297	945	2390	0.543980156	0.819725296	0.681865878
SVM(pol 2, sigmoid)	3366	4557	685	1875	0.642243847	0.869324685	0.755795097
Decision Tree	2817	4691	551	2424	0.537492845	0.894887448	0.716207193
Ense 1(Step. Reg, Gr	3447	3858	1384	1794	0.657698912	0.735978634	0.696842507

Calculated value of Sensitivity, Specificity and overall Accuracy Rate for Validation Dataset



	A	B	C	D	E	F	G	H	I	J	K
13											
14											
15											
16		Models are developed with stratified sample method and selecting equal sample size from both event of the Target group (validation Data set)									
17		Models	TP	TN	FP	FN	Sensitivity	Specificity	Total Accuracy Rate		
18		Reg. Stepwise	954	1090	407	544	0.636849132	0.728122912	0.682470785		
19		Reg. Backward	982	1172	325	514	0.656417112	0.782899132	0.719679252		
20		Gradiant Boosting	960	958	539	538	0.640854473	0.63994656	0.640400668		
21		HP Forest	925	1232	265	573	0.617489987	0.822979292	0.720200334		
22		Neural Network	980	1188	309	518	0.654205607	0.793587174	0.723873122		
23		SVM Interior	905	1222	275	593	0.604138852	0.816299265	0.710183639		
24		SVM(pol 2, sigmoid)	971	966	531	527	0.648197597	0.645290581	0.646744574		
25		Decision Tree	804	1329	168	694	0.536715621	0.887775551	0.712186978		
26		Ense 1(Step. Reg, Gr	987	1089	408	511	0.658878505	0.72745491	0.693155259		
27											
28											
29											

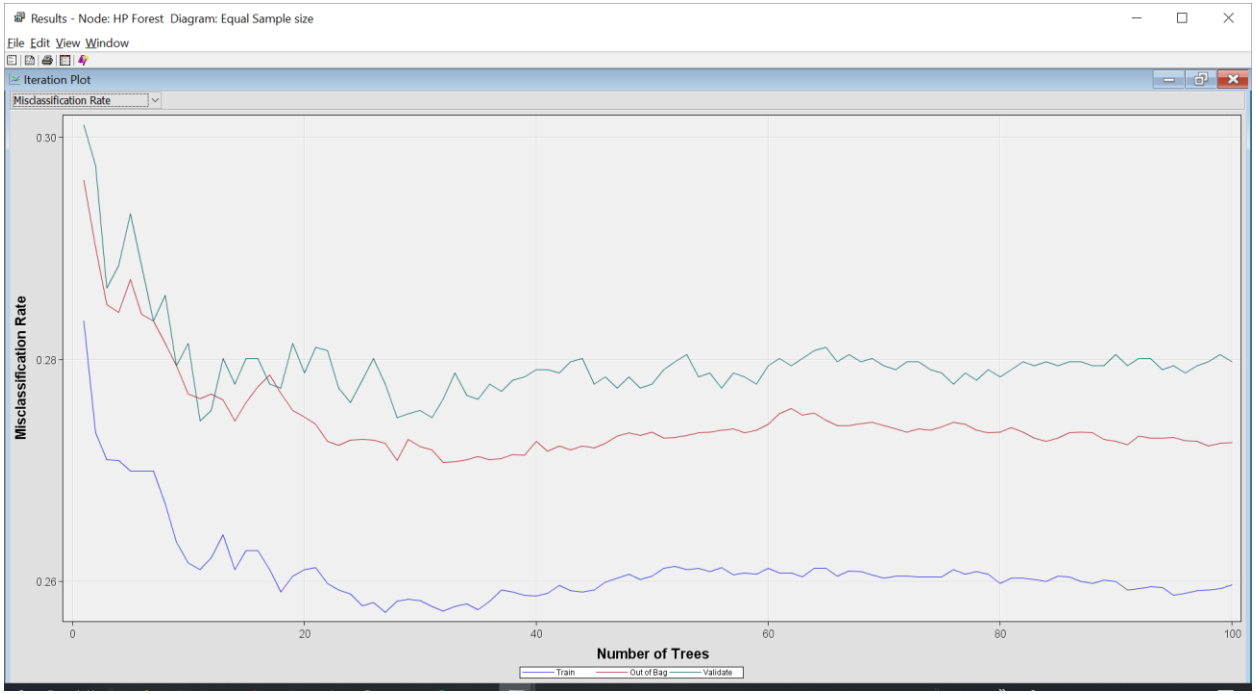
Description of the Three Champion models

First Champion Model: HP Forest

Random forest algorithm is a supervised classification algorithm. The algorithm creates the forest with several trees. The accuracy of the model increases as the number of trees increases. In contrary, average square error decreases as the number of trees increase.

In this project, HP Random was built with 100 trees (Maximum number of trees). The proportion of observations in each tree was 60% of the total observations and the sample was taken with replacement which implies that one observation (one incident report) can have a chance of to be include in more than one tree or in the same tree. The maximum variable used at the start of the tree was 25 and Loss Reduction was used to select important variables with a p- value of 0.05 significant level that is if no association meet this threshold the node will not be split. Based on these properties, the HP Forest model was built. As it can be seen from the below

screenshot, misclass cation rate the datasets (Train, Validation, and Test) decreases as the number of trees increased in reveres the average square error decreased as the trees increases.



Before HP model was developed, all unknowns and outliers (3 SD away from the mean) have been replaced by missing value and then all missing class values were imputed by count. Max Transformation was applied for all skewed variable. The classification table for Training and Validation data have been shown below.

Classification Table for Training data set

Outcome	Predicted fatal/injury	Predicted Car Crash	Total
Actual fatal/injury	**3351**	**1890**	5241
Actual Car Crash	**833**	4409	5242
Total	4992	5491	10483

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 0.74~ 74%

Sensitivity = (TP/TP+FN) = 0.639 ~ 64%

Specificity = (TN/ (TN+FN) = 0.841 ~ 84%

Specificity is predicting fatal/injury accident that showed how much of the events are caught by the model. Out of the total car crash, 3351or 64% of them are predicted as fatal/injury car crash and 1890 or 36 % them are predicted as car damage. Specificity was predicted 84% which is higher than specificity. According to the model the most important variables are Airbag deployment, Alcohol, Type of collision, collision, condition of the driver, Going direction, junction, light code.

Classification Table for Validation data set

Outcome	Predicted fatal/injury	Predicted Car Crash	Total
Actual fatal/injury	**953**	**545**	1498
Actual Car Crash	**414**	1093	1507
Total	1367	1638	3005

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 0.68 ~ 68%

Sensitivity = (TP/TP+FN) = 0.64 ~ 64%

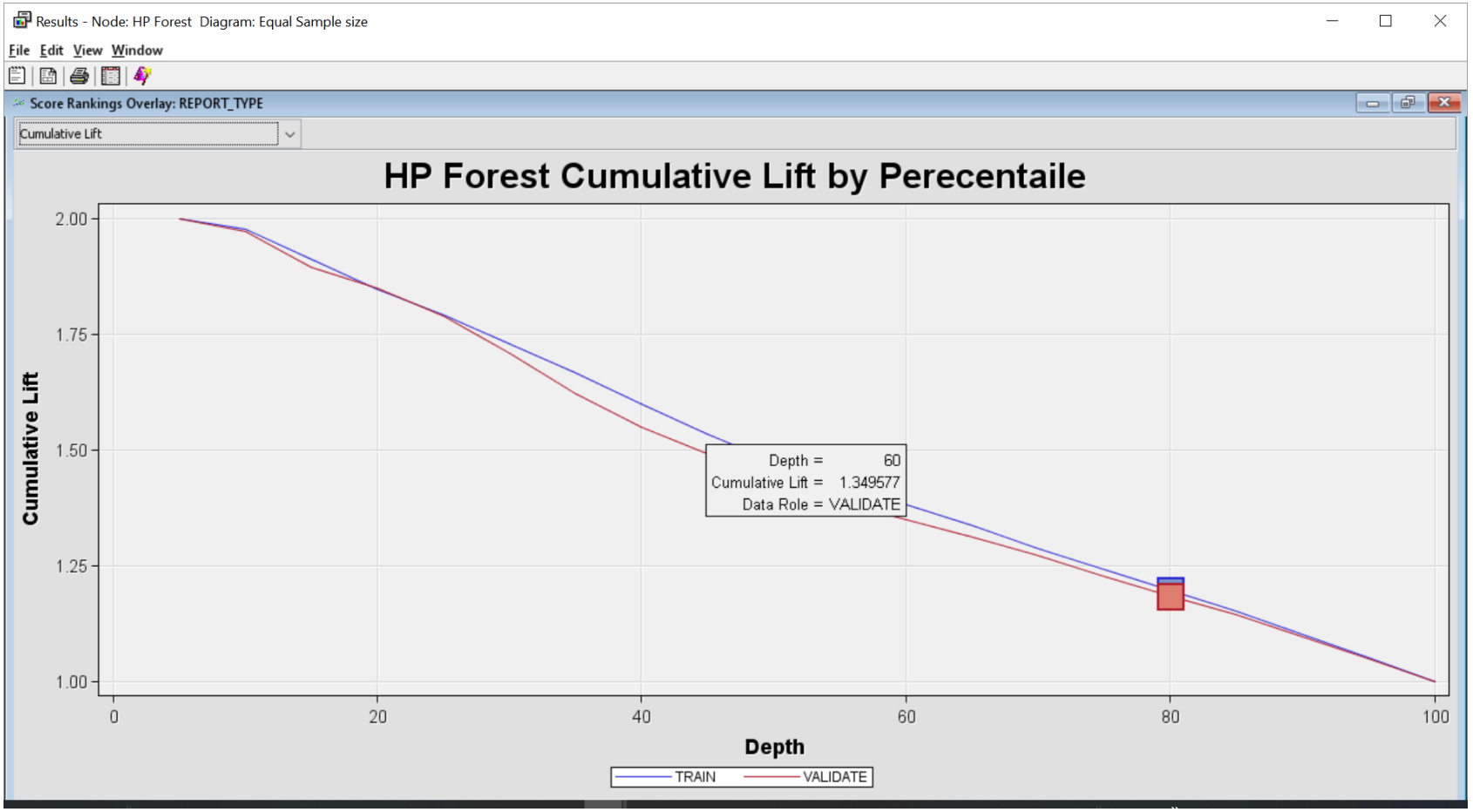
Specificity = (TN/ (TN+FN) = 0.73 ~ 73%

HP Forest model predict 953 fatal/injury car crash correctly with overall accuracy rate of 64% and predict Car Crash correctly with a specificity rate of 0.73 or 73%.

When the accuracy rate of Training dataset was compared with the validation dataset, there is a difference of 0.03 rate

Cumulative Lift for HP Random Forest

Cumulative Lift chart is another evaluating model effectiveness, the x-axis shows the percentile and the y-axis shows lift. The horizontal line intersecting the y-axis at 1 is the chance of identifying the fatal/injury at random is 10% out of the total observation without using model. For example, the Depth 60 and cumulative Lift 1.349, showed that model can capture 60 percentiles (top 60) with a cumulative lift of 1.349 (about 67% accuracy) which is 6 times as many as if no model is used or the lift is above the baseline (1.349. The cumulative lift chart visually shows the advantage of using a predictive model than predicting without the model and do random sample.


Results - Node: HP Forest Diagram: Equal Sample size
File Edit View Window
Score Rankings Overlay: REPORT_TYPE
Cumulative Lift
HP Forest Cumulative Lift by Perecentaile
Cumulative Lift
2.00
1.75
1.50
1.25
1.00
0
20
40
60
80
100
Depth
Depth = 60
Cumulative Lift = 1.349577
Data Role = VALIDATE
TRAIN
VALIDATE

Second model: SVM with Pol 2, sigmoid 2

SVM has a story of success and most popular for nonlinear distribution. One of the reasons for the success is using different optimization method with the option of interior point and active point such as kernel, polynomial degree with is suitable for the distribution.

In this project, SVE with Optimal Method of Active point set of pols 2 with max iteration time of 25 was used. The most important variables with a chi-square value of 3.6 and above were selected to proceed the model to reduce the time of processing and complexity of the model.

SVM Training model classification Table

Outcome	Predicted fatal/injury	Predicted Car Crash	Total

Actual fatal/injury	**3366**	**1875**	5241
Actual Car Crash	**685**	4557	5242
Total	4051	6432	10483

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 0.76 ~ 76%

Sensitivity = (TP/TP+FN) = 0.64 ~ 64%

Specificity = (TN/ (TN+FN) = 0.87 ~87%

In the training dataset, SVM predicted only 64% of fatal/injury accident correctly. However, Car Crash Damage was classified correctly with accuracy rate of 87%. The overall accuracy rate was 76% where 7923 reports were classified correctly out of 10, 483.

SVM Validation Data set Classification Table

Outcome	Predicted fatal/injury	Predicted Car Crash	Total

Actual fatal/injury	**971**	**527**	1498
Actual Car Crash	**531**	966	1497
Total	1502	1493	2995

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 0.65 ~65%

Sensitivity = (TP/TP+FN) = 0.65 ~ 65%

Specificity = (TN/ (TN+FN) = 0.65 ~65 %

In the validation dataset, SVM predicted Sensitivity, Specificity and Total Accuracy rate with a rate of 65%. When the accuracy rate of the Training and Validation dataset was compared, the accuracy rate for sensitivity is 1% greater than the training dataset.

The Third Model: Neural Network

Neural networks are a class of parametric models that can capture a variety of nonlinear relationships between a setoff predictor and a target variable than can a traditional logistic regression model. Building a neural network involves two main stages. First, one must define the network configuration or structure; and then iteratively trains the model based on the given network structure (Kechen et al, na).

In this project, Neural Network was used after all class variable with missing values are replaced by count and transformed with Log Transformation method to reduce the skewness of the distribution. Model selection criteria was used for classification.

Neural Network classification Table for Training dataset

Outcome	Predicted fatal/injury	Predicted Car Crash	Total
Actual fatal/injury	**3538**	**1703**	5241
Actual Car Crash	**1421**	3821	5241
Total	4959	5524	

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 70%

Sensitivity = (TP/TP+FN) = 68%

Specificity = (TN/ (TN+FN) = 73%

Neural Network classification Table for Validation dataset

Outcome	Predicted fatal/injury	Predicted Car Crash	Total
Actual fatal/injury	**980**	**518**	1498
Actual Car Crash	**309**	1188	1497
Total	1289	1706	2995

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 72%

Sensitivity = (TP/TP+FN) = 65 %

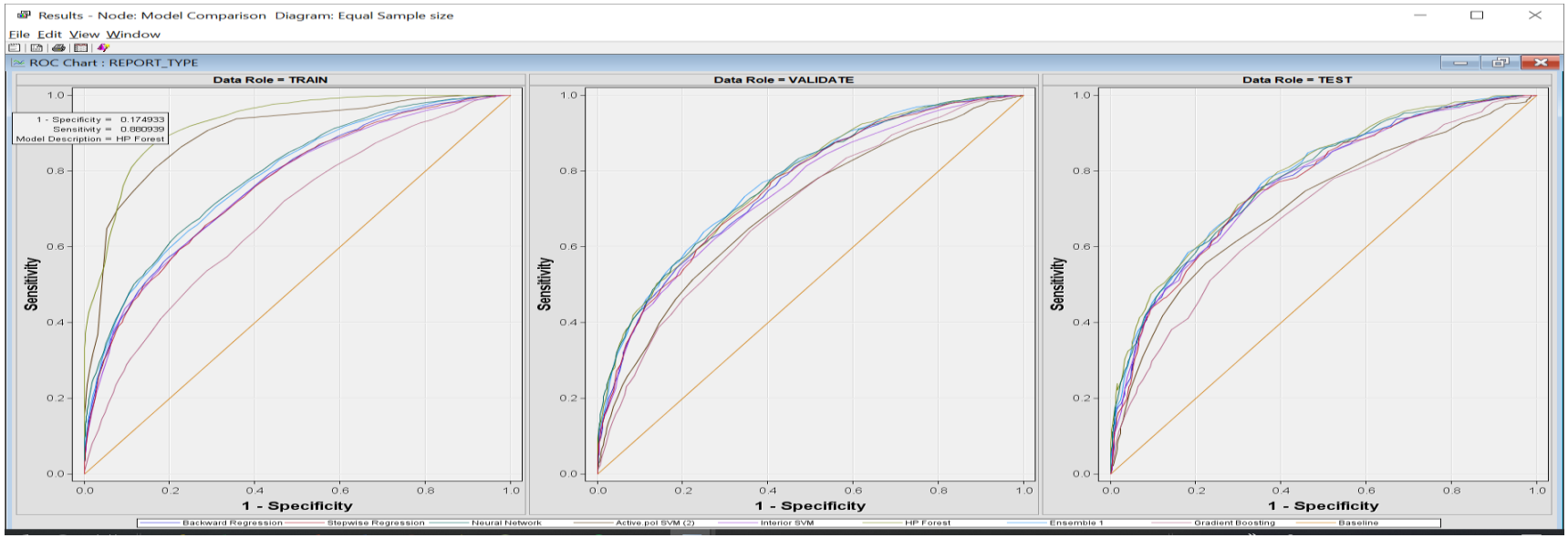
Specificity = (TN/ (TN+FN) = 79 %

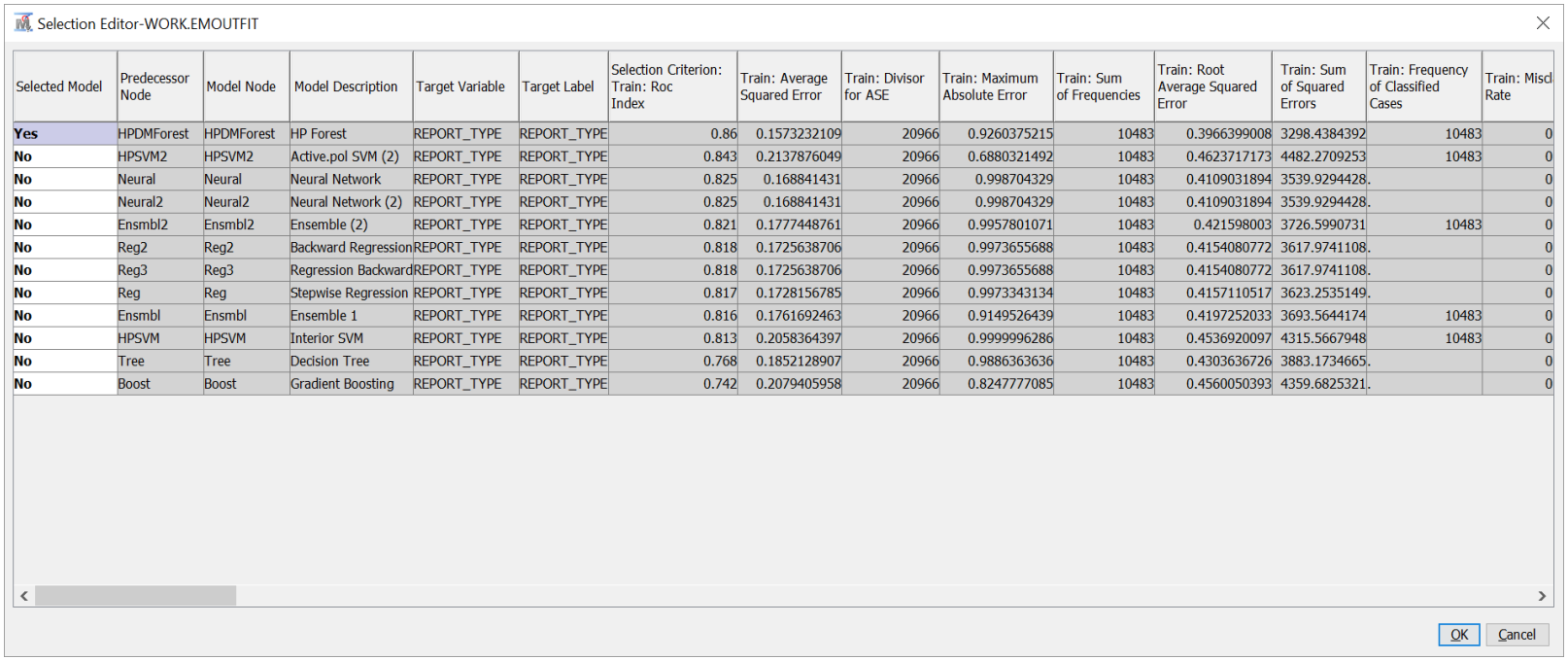
Neural Network model predicated fatal/injury car crash with accuracy rate of 68% on the training dataset and 65% on the validation dataset which shows that a difference of 3% overfit.

Comparison of Models

ROC Index and classification chart to compare the champion model

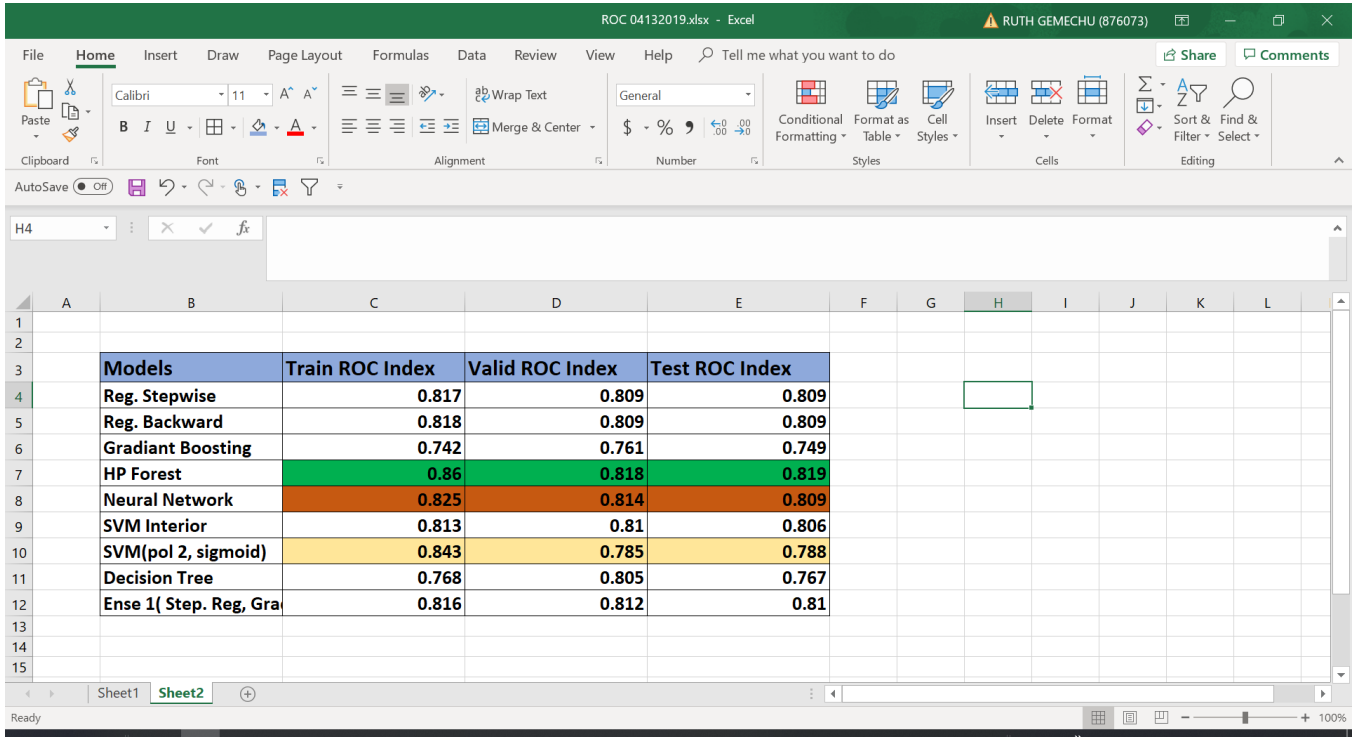
The area under the ROC (Receiver Operating Characteristic) can be used as a criterion to measure the test's discriminative ability. The graph is plotted with true positive rate (sensitivity) as a Y axis and 1-Sensitivity as x-axis which is classifying false negative as true positive. As the Y axis increases, the sensitivity increases, in contrast as the x-axis increase the false positive increase. The goal is to find the optimum point where the model gets the max sensitivity and min false positive. Each point on the ROC curve represents a sensitivity/(1-specificity) pair correspond to a decision threshold. In this project, 0.05 cut off value was used. As it can be seen from the ROC curve and the summary table, the champion model is HP forest with ROC value of 0.86, 0.81, and 0.80 for training, validation, and test dataset. The second champion model is SVM and the ROC values for training, validation and test dataset are 0.84, 0.785 and 0.788 respectively. The third champion is Neural Network, with ROC value of 0.825 for training, 0.814 for validation and 0.80 for test dataset. In general, all three has ROC value above 80 which is catching fatal/injury car crash significantly. Moreover, all three of the models, revealed the most important variables which are Airbag deployment, Alcohol Test, collision, Car Damage code, and safety equipment.


Results - Node: Model Comparison Diagram: Equal Sample size
File Edit View Window
ROC Chart : REPORT_TYPE
Data Role = TRAIN
Data Role = VALIDATE
Data Role = TEST
1 - Specificity = 0.174933
Sensitivity = 0.880939
Model Description = HP Forest
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Backward Regression
Stepwise Regression
Neural Network
Active.pol SVM (2)
Interior SVM
HP Forest
Ensemble 1
Gradient Boosting
Baseline



Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Roc Index	Train: Average Squared Error	Train: Divisor for ASE	Train: Maximum Absolute Error	Train: Sum of Frequencies	Train: Root Average Squared Error	Train: Sum of Squared Errors	Train: Frequency of Classified Cases	Train: Misc Rate
Yes	HPDMForest	HPDMForest	HP Forest	REPORT_TYPE	REPORT_TYPE	0.86	0.1573232109	20966	0.9260375215	10483	0.3966399008	3298.4384392	10483	0
No	HPSVM2	HPSVM2	Active.pol SVM (2)	REPORT_TYPE	REPORT_TYPE	0.843	0.2137876049	20966	0.6880321492	10483	0.4623717173	4482.2709253	10483	0
No	Neural	Neural	Neural Network	REPORT_TYPE	REPORT_TYPE	0.825	0.168841431	20966	0.998704329	10483	0.4109031894	3539.9294428.		0
No	Neural2	Neural2	Neural Network (2)	REPORT_TYPE	REPORT_TYPE	0.825	0.168841431	20966	0.998704329	10483	0.4109031894	3539.9294428.		0
No	Ensmbl2	Ensmbl2	Ensemble (2)	REPORT_TYPE	REPORT_TYPE	0.821	0.1777448761	20966	0.9957801071	10483	0.421598003	3726.5990731	10483	0
No	Reg2	Reg2	Backward Regression	REPORT_TYPE	REPORT_TYPE	0.818	0.1725638706	20966	0.9973655688	10483	0.4154080772	3617.9741108.		0
No	Reg3	Reg3	Regression Backward	REPORT_TYPE	REPORT_TYPE	0.818	0.1725638706	20966	0.9973655688	10483	0.4154080772	3617.9741108.		0
No	Reg	Reg	Stepwise Regression	REPORT_TYPE	REPORT_TYPE	0.817	0.1728156785	20966	0.9973343134	10483	0.4157110517	3623.2535149.		0
No	Ensmbl	Ensmbl	Ensemble 1	REPORT_TYPE	REPORT_TYPE	0.816	0.1761692463	20966	0.9149526439	10483	0.4197252033	3693.5644174	10483	0
No	HPSVM	HPSVM	Interior SVM	REPORT_TYPE	REPORT_TYPE	0.813	0.2058364397	20966	0.9999996286	10483	0.4536920097	4315.5667948	10483	0
No	Tree	Tree	Decision Tree	REPORT_TYPE	REPORT_TYPE	0.768	0.1852128907	20966	0.9886363636	10483	0.4303636726	3883.1734665.		0
No	Boost	Boost	Gradient Boosting	REPORT_TYPE	REPORT_TYPE	0.742	0.2079405958	20966	0.8247777085	10483	0.4560050393	4359.6825321.		0

ROC Table: HP Forest, SVM (pol deg 2), and Neural Network are the three-champion model based on ROC value



Models	Train ROC Index	Valid ROC Index	Test ROC Index
Reg. Stepwise	0.817	0.809	0.809
Reg. Backward	0.818	0.809	0.809
Gradiant Boosting	0.742	0.761	0.749
HP Forest	0.86	0.818	0.819
Neural Network	0.825	0.814	0.809
SVM Interior	0.813	0.81	0.806
SVM(pol 2, sigmoid)	0.843	0.785	0.788
Decision Tree	0.768	0.805	0.767
Ense 1(Step. Reg, Gra	0.816	0.812	0.81

Conclusion

What is the best model? The quality of the model shouldn't be only measured by accuracy rate, but the percentage of overfitting or underfitting, simplicity, easily understandable by decision maker or layman and convenience to implement do matter the most.

According to the result of the ROC chart and cumulative Lift the first champion model is HP forest with a ROC value of 0.86, 0.818, 0.819 for the training, validation and test data respectively. Even though HP Forest is the champion model, most of the models came up with highest ROC value and when these models are compared with HP Forest, the difference is very low. As it can be seen from the above ROC Table, out of 9 models, only Decision Tree (ROC = 0.76) and Gradient Boosting (0.74) have ROC value less than 0.80. These implies that only ROC value can't determine the best model, instead assessing classification matrix and combine the overall result helps to select the working model.

As the classification matrix indicates, Logistic Regression with backward selection method takes the first rank in classifying sensitivity (fatal/injury car crash) with a rate of 0.667, Specificity of 0.79 and overall accuracy rate of 0.72 for training dataset. In contrary, HP forest predicted fatal/injury with a rate of 0.639 (sensitivity), 0.84 (specificity) and 0.74(overall accuracy rate) with a difference of 3%. So, based on the combination effect of ROC value and classification matrix, I personally, suggest that Logistic reg with back ward method is a better model than HP Forest.

When these two models are compared with respect to overfitting, which is when a model learns the detail and noise in the training dataset in such a way that it affects the performance of the model, again Logistic Reg is leading the way with a minimum overfitting of 0.011, (0.667 -0.656). In contrast, HP Forest has overfitting with a value of 0.038 (0.639-0.61). Thus, Logistic Regression is a better model in performing with unknown dataset than HP Forest.

Finally, the two models are compared with respect to complexity, interpretability and ease of implementation. As it is known, one of the strength parts of HP Forest was building the model with 100 trees and a max variable of 25, then dropping the weak variables with loss reduction and posterior probability. Even though, the approach is very systematic in dropping redundant variable or insignificant variables by using loss reduction variable importance with valid Gini or Valid Margin values, it is very complicated interpret the result. Logistic regression with back ward methods used simple likelihood ratio test for Null Hypothesis of Beta value with the input value of effect, Df, chi-square test and probability. In this regard, both models identified the same type of variables, including their codes, to determine the most significant variables. These are Airbag dep, Alcohol test, driver condition, damage code, and safety equipment.

In general, in all aspect of statistical measurement except ROC value, Logistic regression with backward method showed good performance than Hp forest and can be the best models to implement for the reduction of fatal/ injury accident in Maryland. In addition, the sensitivity is below 80% in all models which was the goal of this project to predict as min accuracy rate. However, to compensate the drawback, cost of classification was considered by applying more weight for death or injury event than car damage accident. I believe it is true and acceptable among the decision makers to reduce the risk of death by giving appropriate attention with respect to time, money and knowledge to come to accomplish the vision of ZERO death in Maryland due to car crash.

Computing cost of classification

Why do decision makers use cost classification? Cost classification would be the best solution to bring to management's attention certain costs that are considered more crucial than others. In this project, it is absolutely recommended to impose different weight for fatal/injury accident and car damage because of the consequence of fatal/injury on individual life, family, society and county economy. The below cost matrix was used to calculate the cost of each occurrence based on the Logistic Regression with backward selection method.

Cost classification based on the Logistic Regression backward

Cost Matrix		Predicted class	
Actual class	**C(I/J)**	**+**	**-**
	+	**-1(3499)**	**100(1742)**
	-	**1(1107)**	**0(3821)**

- The cost of predicting fatal/injury car crash incorrectly weight 100 times more than predicting car damage as fatal. Based on this the cost classification will be = -3499 + 1107 +174200 + o= 170, 808. This implies that fatal/injury car crash is very serious which needs to be addressed seriously.

Logistic Regression with backward selection method classification Table for Training dataset

Outcome	Predicted fatal/injury	Predicted Car Crash	Total
Actual fatal/injury	**3499**	**1742**	5241
Actual Car Crash	**1107**	3821	4928
Total	4959	5524	10229

Total Accuracy = (TP +TN) / (TP+TN+FP+FN) = 72%

Sensitivity = (TP/TP+FN) = 667%

Specificity = (TN/ (TN+FN) = 78%

Reference

Brandon. W (2019): Crashes more than twice as common on I-83 as MD Highways.

MHSP (2017): Maryland High way safety office: Retrieved from

http://www.mva.maryland.gov/safety/_docs/FFY2017MarylandHSP.pdf

MDT (2018): Maryland Department of Transportation: Retrieved from http://www.mva.maryland.gov/safety/mhso/index.htm

Six Reasons why data science project fail: Retrieved from https://medium.com/@ODSC/6-reasons-why-data-science-projects-fail-6240bf9326f6

Miller and Zois (2016): Car Accident statistics in Maryland: Retrieved from https://www.millerandzois.com/car-accident-statistics.html

http://www.mva.maryland.gov/safety/mhso/Maryland-Traffic-Safety-Data.htm

http://www.mva.maryland.gov/safety/mhso/index.htm

https://www.cio.com.au/article/401353/how_define_scope_project/

https://rapidbi.com/the-difference-between-goals-objectives/

https://www.pinderplotkin.com/er-vs-urgentcare/

https://www.iii.org/sites/default/files/docs/pdf/auto_rates_wp_092716-62.pdf

http://www.marylandconsumers.org/penn_station/folders/issues/auto/insurance/Auto_Insurance_Brief_with_Infographics_-_for_website.pdf

https://www.hindawi.com/journals/mpe/2013/547904/

United States Census Bureau. B01001 SEX BY AGE, 2017 American Community Survey 5-Year Estimates. U.S. Census Bureau, American Community Survey Office. Web. 6 December 2018. http://www.census.gov/.

http://www.mva.maryland.gov/_resources/docs/MarylandSHSP_2016-2020-Final.pdf

http://jesshampton.com/2013/11/11/model-evaluation-explaining-the-cumulative-lift-chart/

Appendix 1

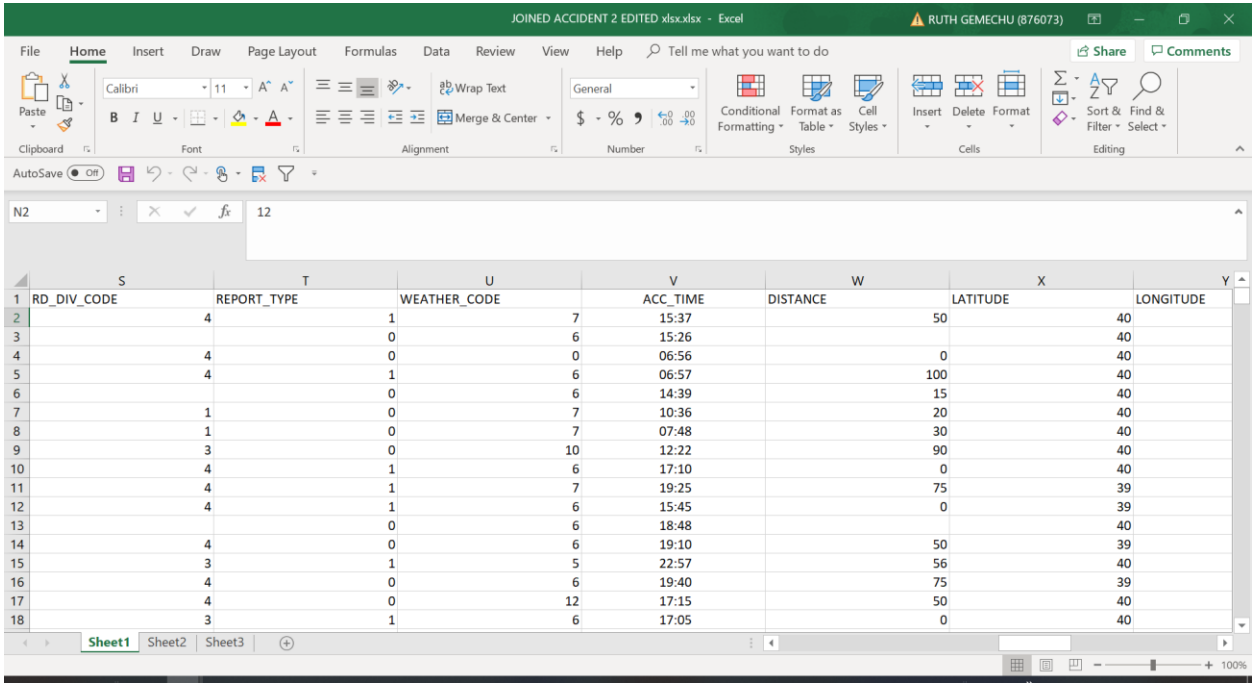
VARIABLE	CATEGORY	DATA TYPE
SEX_CODE	F=FEMALE M=MALE U=UNKOWN	Nominal
CONDITION_CODE	U- UNKOWN 00=NOT APPLICABLE 01=APPARENTLY NORMAL 02=HAD BEEN DRINKING 03=USING DRUGS 04=PHYSICAL DEFECTS 05=OTHER HADICAPS 06=III 07=FATIGUED FAINTED 08=APPARENTLY ASLEEP 09=EMOTIONAL DEPRESSED ANGRY DISTURBED	Nominal

	10=INFLUENCED BY Medications and/or DRUGS AND /OR ALCOHOL 88=OTHER HADICAPS 99UNKNOWN	
OCC-SEAT-POST- CODE	00=NOT APPLICAPLE 01=IN VEHICLE 02=CENTER FRONT SEAT 03=RIGHT FRONT SEAT 04=LEFT REAR/MOTORCYLCE PASSENGER 05=CENTER REAR SEAT 06=RIGHT REAR SEAT 07=OTHER IN VEHICLE 08=CARGO AREA 09=UNENCLOSED CARGO AREA 15=TRAILER UNIT 88= OTHER 99= UNKNOWN	Nominal
ALCHOL _TEST _CODE	00=NOT APPLICABE 01=TEST REFUSED 02=POSITIVE PRELIMINARY TEST 03=EVIDENCE TEST GIVEN 88=OTHE R 99=UNKNOWN	Nominal

DRUG_TEST RESULT_CODE	00=NOT APPLICABE 01=TEST REFUSED 02=POSITIVE PRELIMINARY TEST 03=EVIDENCE TEST GIVEN 88=OTHE R 99=UNKNOWN	Nominal
EQUIP-PROOB-CODE		Nominal
SAF-EQUIP-CODE		Nominal
AIRBAG_DEPLOYED		Nominal
DATA OF BIRTH		Nominal
EMS-UNIT LABEL		Nominal
LIGHT CODE	00=NOT APPLICABEL 01=DAYLIGHT 03=DARK LIGHT ON 04=DARK NO LIGHTS 05=DAWN 06=DUSK 07=DARK-UNKNOWN 88=OTHER 99=UNKNOWN	Nominal
COUNTY_NO	01=ALLEGANY 02=ANNE ARUNDEL 03=BALTIMORE 04=CALVERT 05=CAROLINE 06=CARROLL 07=CECIL 08=CHARLES 09=DORCHESTER	Nominal

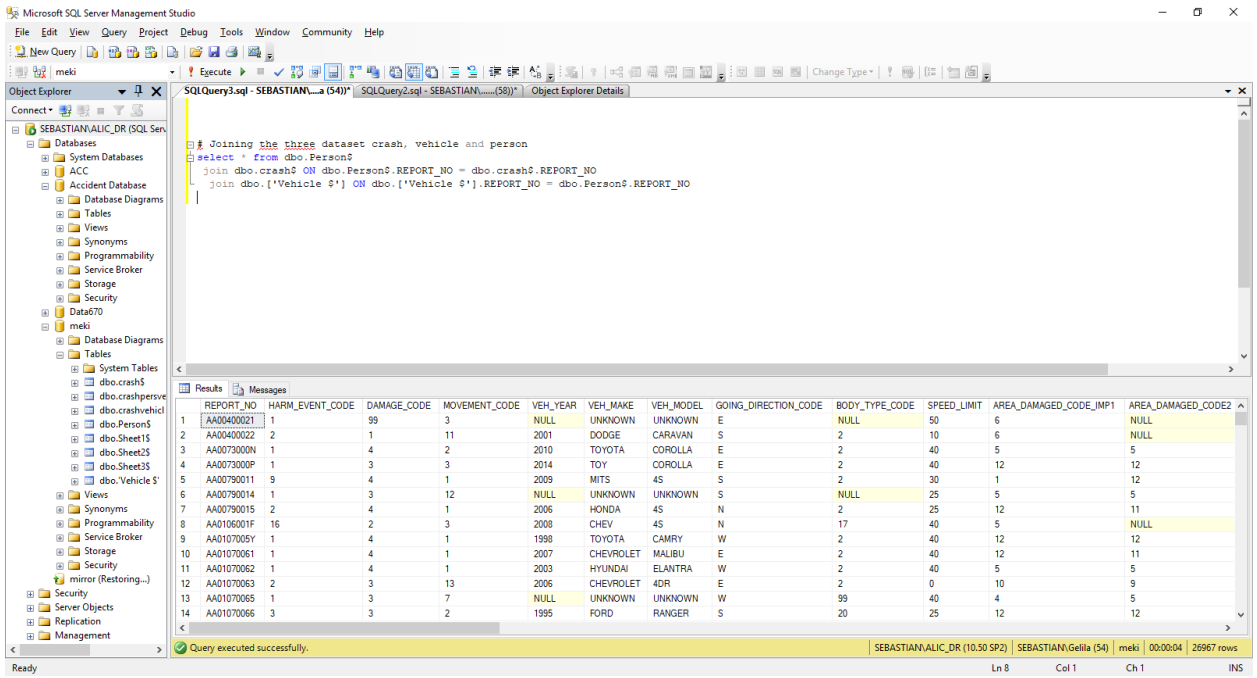
	10=FREDERICK 11=GARRETT 12=HARFORD 13=HOWARD 14=KENT 15=MONTGOMERY 16=PRINCE GEORGE'S 17=QUEEN ANNE'S 18=ST. MARY'S 19=SOMERSET 20=TALBOT 21=WASHINGTON 22=WICOMICO 23=WORCESTER 24=BALTIMORE CITY	
JUNCTION_TYPE_CODE		Nominal
		Nominal
		Nominal
COLLISION_TYPE_CODE		Nominal
SURFACE_CODNITION_CODE		Nominal
RD_COND_CODE		Nominal
RD_DIV_CODE		Nominal
WEATHER_CODE		Nominal
ACC_TIME		Nominal
RTE-NO		Nominal

APPENDIX 2 JOINED DATASET WITH MS-SQL

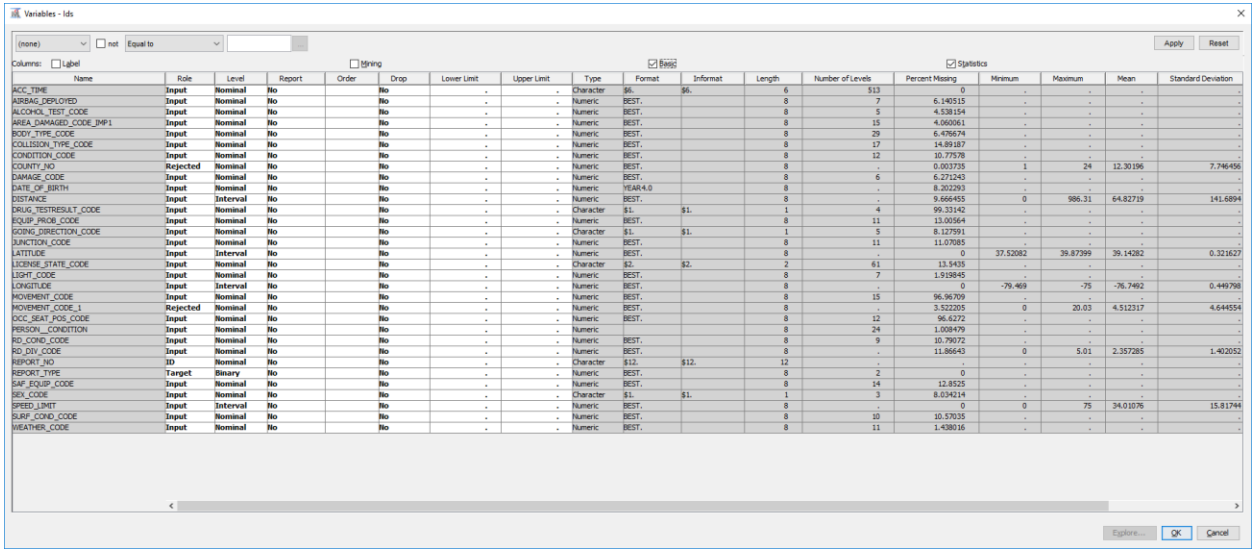


	S	T	U	V	W	X	Y
1	RD_DIV_CODE	REPORT_TYPE	WEATHER_CODE	ACC_TIME	DISTANCE	LATITUDE	LONGITUDE
2	4	1	7	15:37	50	40	
3		0	6	15:26		40	
4	4	0	0	06:56	0	40	
5	4	1	6	06:57	100	40	
6		0	6	14:39	15	40	
7	1	0	7	10:36	20	40	
8	1	0	7	07:48	30	40	
9	3	0	10	12:22	90	40	
10	4	1	6	17:10	0	40	
11	4	1	7	19:25	75	39	
12	4	1	6	15:45	0	39	
13		0	6	18:48		40	
14	4	0	6	19:10	50	39	
15	3	1	5	22:57	56	40	
16	4	0	6	19:40	75	39	
17	4	0	12	17:15	50	40	
18	3	1	6	17:05	0	40	

APPENDIX 3 COMBINING THE THREE DATASETS WITH MS-SQL INNER JOINT WITH COMMON VARIALBE OF REPORT NO

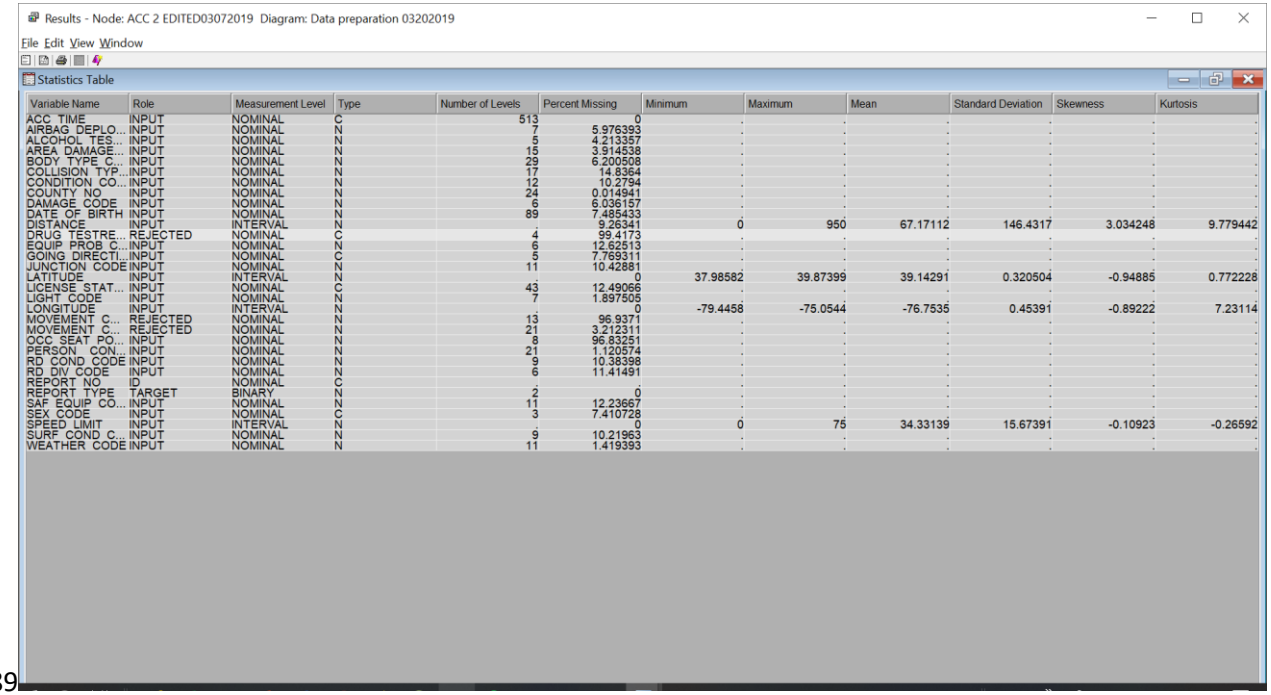


APPENDIX 4: 1 INITIAL EXPLORATION


Variables - Ids
(none)
not
Equal to
Apply
Reset
Columns:
Label
Mining
Basic
Statistics
Name
Role
Level
Report
Order
Drop
Lower Limit
Upper Limit
Type
Format
Informat
Length
Number of Levels
Percent Missing
Minimum
Maximum
Mean
Standard Deviation
ACC_TIME Input Nominal No No Character $6. $6. 6 513 0
AIRBAG_DEPLOYED Input Nominal No No Numeric BEST. 8 7 6.140515
ALCOHOL_TEST_CODE Input Nominal No No Numeric BEST. 8 5 4.538154
AREA_DAMAGED_CODE_IMP1 Input Nominal No No Numeric BEST. 8 15 4.060061
BODY_TYPE_CODE Input Nominal No No Numeric BEST. 8 29 6.476674
COLLISION_TYPE_CODE Input Nominal No No Numeric BEST. 8 17 14.89187
CONDITION_CODE Input Nominal No No Numeric BEST. 8 12 10.77578
COUNTY_NO Rejected Nominal No No Numeric BEST. 8 . 0.003735 1 24 12.30196 7.746456
DAMAGE_CODE Input Nominal No No Numeric BEST. 8 6 6.271243
DATE_OF_BIRTH Input Nominal No No Numeric YEAR4.0 8 . 8.202293
DISTANCE Input Interval No No Numeric BEST. 8 . 9.666455 0 986.31 64.82719 141.6894
DRUG_TESTRESULT_CODE Input Nominal No No Character $1. $1. 1 4 99.33142
EQUIP_PROB_CODE Input Nominal No No Numeric BEST. 8 11 13.00564
GOING_DIRECTION_CODE Input Nominal No No Character $1. $1. 1 5 8.127591
JUNCTION_CODE Input Nominal No No Numeric BEST. 8 11 11.07085
LATITUDE Input Interval No No Numeric BEST. 8 . 0 37.52082 39.87399 39.14282 0.321627
LICENSE_STATE_CODE Input Nominal No No Character $2. $2. 2 61 13.5435
LIGHT_CODE Input Nominal No No Numeric BEST. 8 7 1.919845
LONGITUDE Input Interval No No Numeric BEST. 8 . 0 -79.469 -75 -76.7492 0.449798
MOVEMENT_CODE Input Nominal No No Numeric BEST. 8 15 96.96709
MOVEMENT_CODE_1 Rejected Nominal No No Numeric BEST. 8 . 3.522205 0 20.03 4.512317 4.644554
OCC_SEAT_POS_CODE Input Nominal No No Numeric BEST. 8 12 96.6272
PERSON__CONDITION Input Nominal No No Numeric 8 24 1.008479
RD_COND_CODE Input Nominal No No Numeric BEST. 8 9 10.79072
RD_DIV_CODE Input Nominal No No Numeric BEST. 8 . 11.86643 0 5.01 2.357285 1.402052
REPORT_NO ID Nominal No No Character $12. $12. 12 . .
REPORT_TYPE Target Binary No No Numeric BEST. 8 2 0
SAF_EQUIP_CODE Input Nominal No No Numeric BEST. 8 14 12.8525
SEX_CODE Input Nominal No No Character $1. $1. 1 3 8.034214
SPEED_LIMIT Input Interval No No Numeric BEST. 8 . 0 0 75 34.01076 15.81744
SURF_COND_CODE Input Nominal No No Numeric BEST. 8 10 10.57035
WEATHER_CODE Input Nominal No No Numeric BEST. 8 11 1.438016
Explore...
OK
Cancel

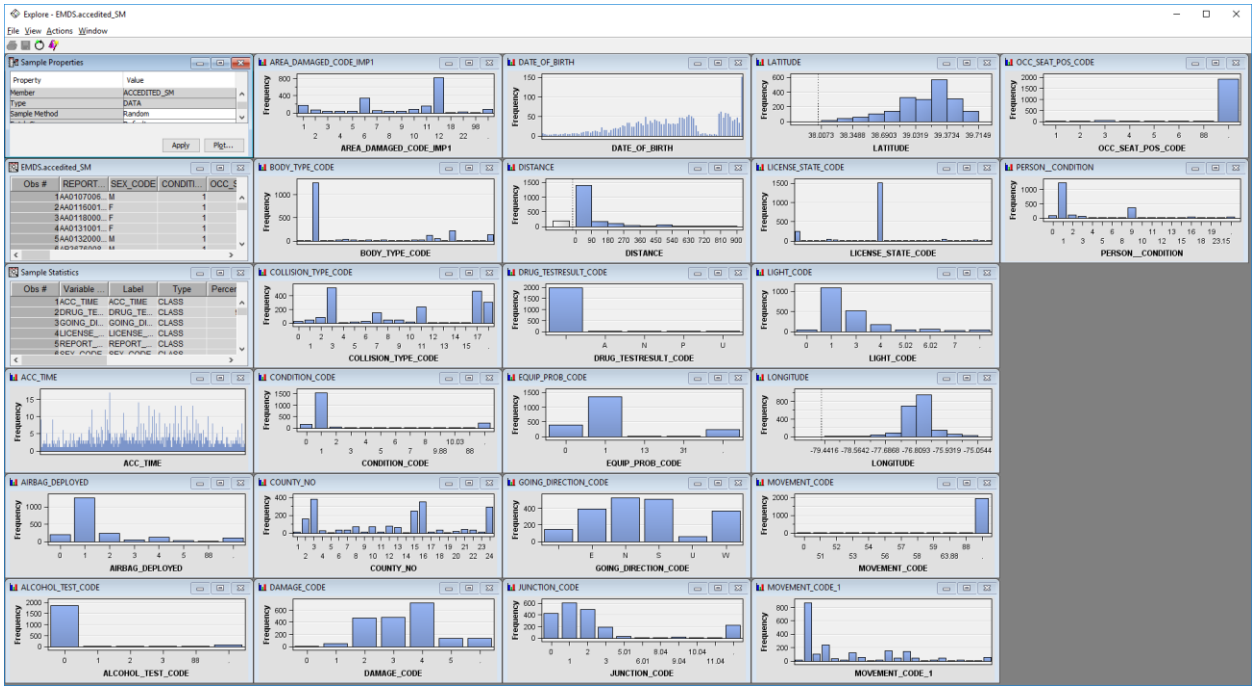
APPENDIX 4:2

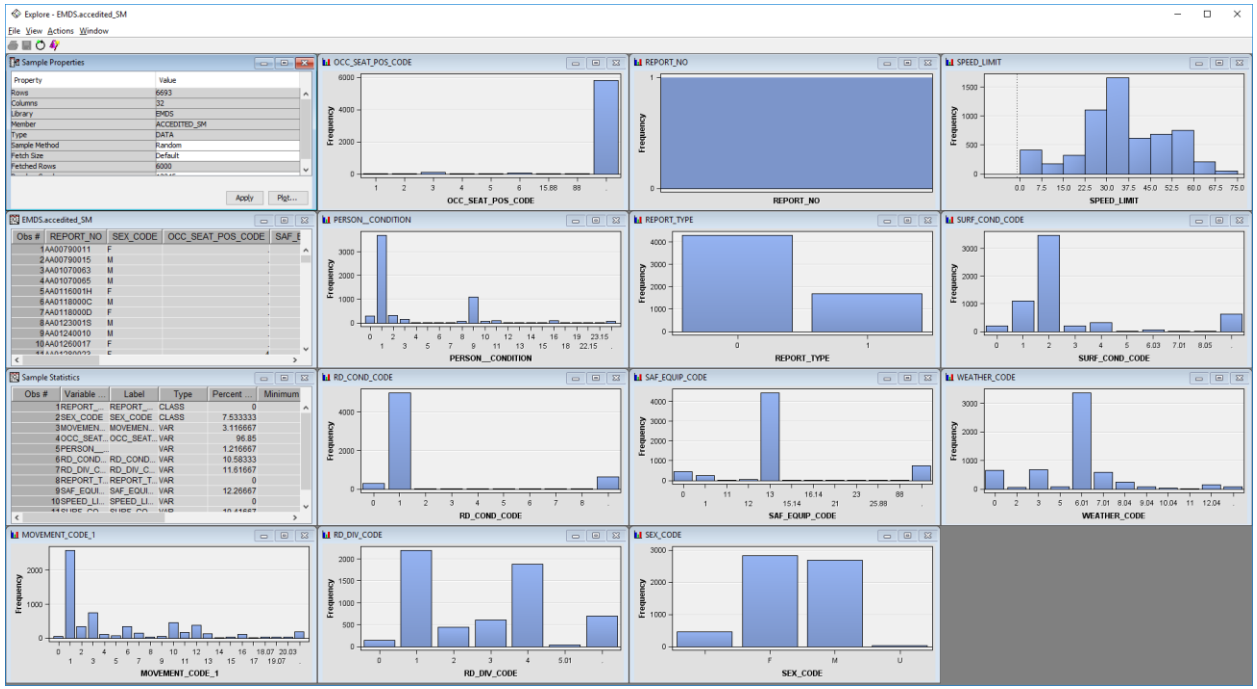
1.89



Variable Name	Role	Measurement Level	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
ACC_TIME	INPUT	NOMINAL	C	513	0						
AIRBAG_DEPLO...	INPUT	NOMINAL	N	7	5.976393						
ALCOHOL_TES...	INPUT	NOMINAL	N	5	4.213357						
AREA_DAMAGE...	INPUT	NOMINAL	N	15	3.914538						
BODY_TYPE_C...	INPUT	NOMINAL	N	29	6.200508						
COLLISION_TYP...	INPUT	NOMINAL	N	17	14.8364						
CONDITION_CO...	INPUT	NOMINAL	N	12	10.2794						
COUNTY_NO	INPUT	NOMINAL	N	24	0.014941						
DAMAGE_CODE	INPUT	NOMINAL	N	6	6.036157						
DATE_OF_BIRTH	INPUT	NOMINAL	N	89	7.485433						
DISTANCE	INPUT	INTERVAL	N		9.26341	0	950	67.17112	146.4317	3.034248	9.779442
DRUG_TESTRE...	REJECTED	NOMINAL	C	4	99.4173						
EQUIP_PROB_C...	INPUT	NOMINAL	N	6	12.62513						
GOING_DIRECTI...	INPUT	NOMINAL	C	5	7.769311						
JUNCTION_CODE	INPUT	NOMINAL	N	11	10.42881						
LATITUDE	INPUT	INTERVAL	N		0	37.98582	39.87399	39.14291	0.320504	-0.94885	0.772228
LICENSE_STAT...	INPUT	NOMINAL	C	43	12.49066						
LIGHT_CODE	INPUT	NOMINAL	N	7	1.897505						
LONGITUDE	INPUT	INTERVAL	N		0	-79.4458	-75.0544	-76.7535	0.45391	-0.89222	7.23114
MOVEMENT_C...	REJECTED	NOMINAL	N	13	96.9371						
MOVEMENT_C...	REJECTED	NOMINAL	N	21	3.212311						
OCC_SEAT_PO...	INPUT	NOMINAL	N	8	96.83251						
PERSON_CON...	INPUT	NOMINAL	N	21	1.120574						
RD_COND_CODE	INPUT	NOMINAL	N	9	10.38398						
RD_DIV_CODE	INPUT	NOMINAL	N	6	11.41491						
REPORT_NO	ID	NOMINAL	C								
REPORT_TYPE	TARGET	BINARY	N	2	0						
SAF_EQUIP_CO...	INPUT	NOMINAL	N	11	12.23667						
SEX_CODE	INPUT	NOMINAL	C	3	7.410728						
SPEED_LIMIT	INPUT	INTERVAL	N		0	0	75	34.33139	15.67391	-0.10923	-0.26592
SURF_COND_C...	INPUT	NOMINAL	N	9	10.21963						
WEATHER_CODE	INPUT	NOMINAL	N	11	1.419393						

Appendix 5 Initial Exploration


Explore - EMDS.accedited_SM
File View Actions Window
Sample Properties
Property
Value
Member
ACCEDITED_SM
Type
DATA
Sample Method
Random
Apply
Plot...
EMDS.accedited_SM
Obs #
REPORT...
SEX_CODE
CONDITI...
Sample Statistics
Obs #
Variable
Label
Type
ACC_TIME
DRUG_TE...
GOING_DI...
LICENSE_...
REPORT_...
AREA_DAMAGED_CODE_IMP1
DATE_OF_BIRTH
LATITUDE
OCC_SEAT_POS_CODE
BODY_TYPE_CODE
DISTANCE
LICENSE_STATE_CODE
PERSON__CONDITION
COLLISION_TYPE_CODE
DRUG_TESTRESULT_CODE
LIGHT_CODE
ACC_TIME
CONDITION_CODE
EQUIP_PROB_CODE
LONGITUDE
AIRBAG_DEPLOYED
COUNTY_NO
GOING_DIRECTION_CODE
MOVEMENT_CODE
ALCOHOL_TEST_CODE
DAMAGE_CODE
JUNCTION_CODE
MOVEMENT_CODE_1
Frequency

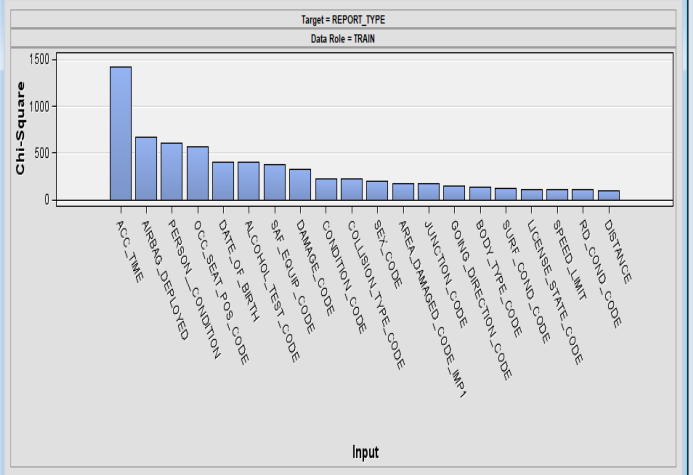


APPENDIX 6 STAT EXPLORE NODE WITH CHI-SQUARE PLOT AND WORTH

Results - Node: StatExplore Diagram: Data preparation 03202019

File Edit View Window

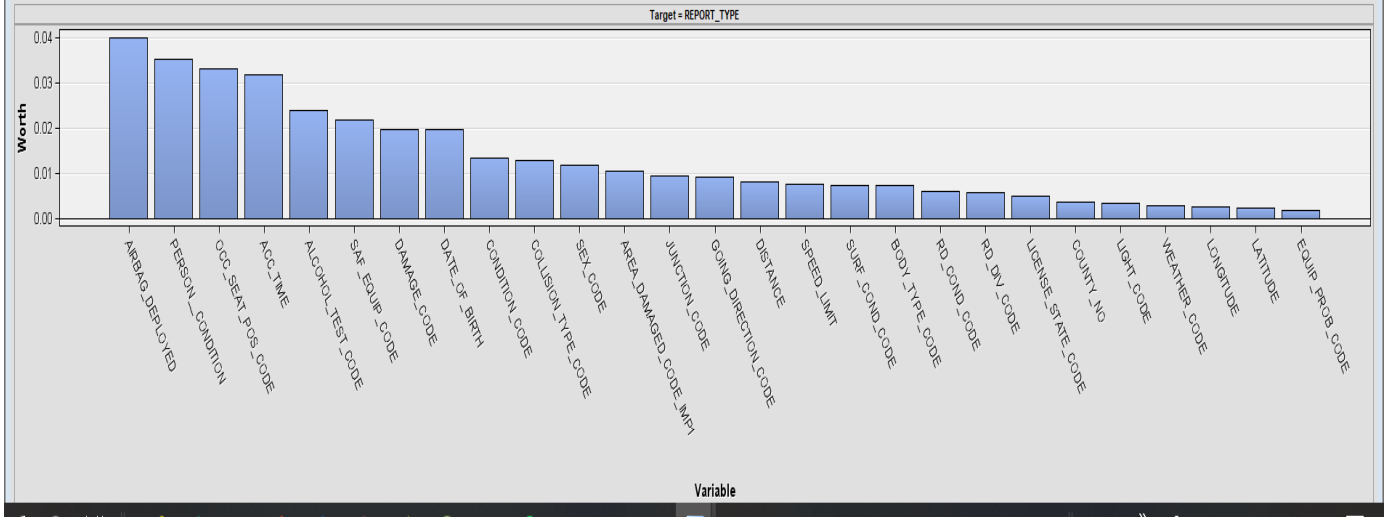
Chi-Square Plot



Output

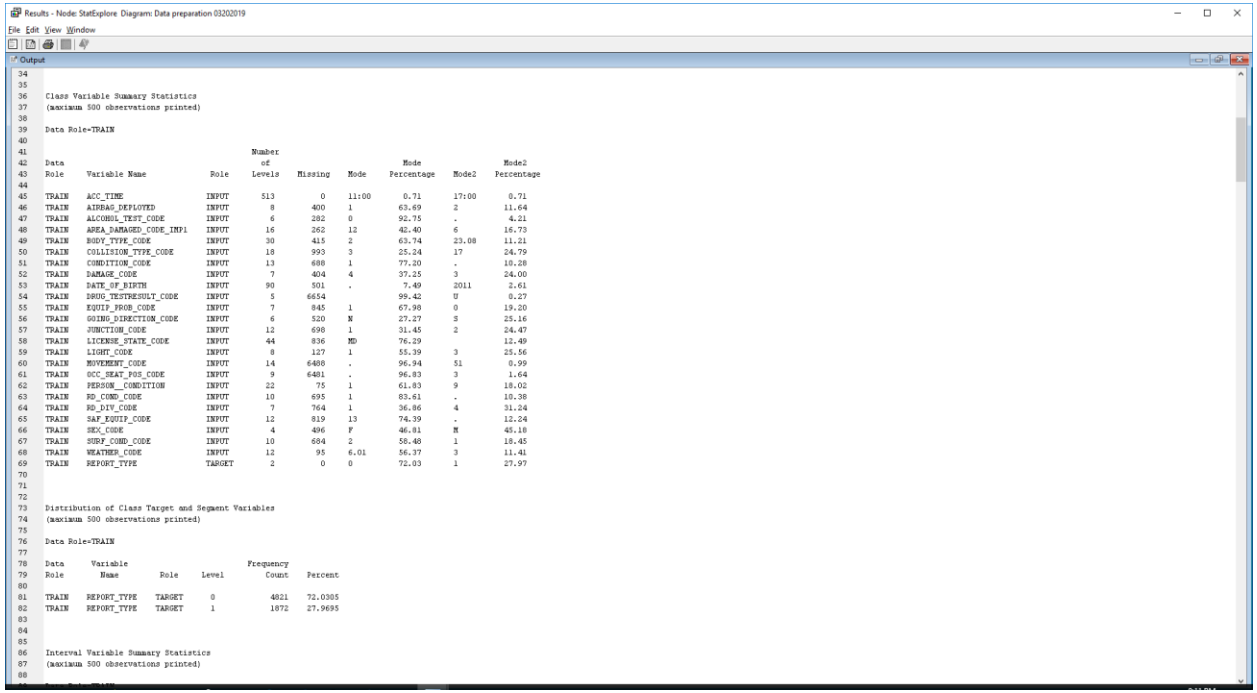
```
*------------------------------------------------------------*



Variable Summary

             Measurement    Frequency
Role           Level         Count

ID           NOMINAL           1
INPUT        INTERVAL          4
INPUT        NOMINAL          23
REJECTED     NOMINAL           3
TARGET       BINARY            1


Variable Levels Summary
(maximum 500 observations printed)

                           Frequency
Variable      Role          Count
```

Variable Worth



Appendix 5

Class Variable Summary Statistics



Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	ACC_TIME	INPUT	513	0	11:00	0.71	17:00	0.71
TRAIN	AIRBAG_DEPLOYED	INPUT	8	400	1	63.69	2	11.64
TRAIN	ALCOHOL_TEST_CODE	INPUT	6	282	0	92.75	.	4.21
TRAIN	AREA_DAMAGED_CODE_IMP1	INPUT	16	262	12	42.40	6	16.73
TRAIN	BODY_TYPE_CODE	INPUT	30	415	2	63.74	23.08	11.21
TRAIN	COLLISION_TYPE_CODE	INPUT	18	993	3	25.24	17	24.79
TRAIN	CONDITION_CODE	INPUT	13	688	1	77.20	.	10.28
TRAIN	DAMAGE_CODE	INPUT	7	404	4	37.25	3	24.00
TRAIN	DATE_OF_BIRTH	INPUT	90	501	.	7.49	2011	2.61
TRAIN	DRUG_TESTRESULT_CODE	INPUT	5	6654		99.42	U	0.27
TRAIN	EQUIP_PROB_CODE	INPUT	7	845	1	67.98	0	19.20
TRAIN	GOING_DIRECTION_CODE	INPUT	6	520	N	27.27	S	25.16
TRAIN	JUNCTION_CODE	INPUT	12	698	1	31.45	2	24.47
TRAIN	LICENSE_STATE_CODE	INPUT	44	836	MD	76.29		12.49
TRAIN	LIGHT_CODE	INPUT	8	127	1	55.39	3	25.56
TRAIN	MOVEMENT_CODE	INPUT	14	6488	.	96.94	51	0.99
TRAIN	OCC_SEAT_POS_CODE	INPUT	9	6481	.	96.83	3	1.64
TRAIN	PERSON__CONDITION	INPUT	22	75	1	61.83	9	18.02
TRAIN	RD_COND_CODE	INPUT	10	695	1	83.61	.	10.38
TRAIN	RD_DIV_CODE	INPUT	7	764	1	36.86	4	31.24
TRAIN	SAF_EQUIP_CODE	INPUT	12	819	13	74.39	.	12.24
TRAIN	SEX_CODE	INPUT	4	496	F	46.81	M	45.18
TRAIN	SURF_COND_CODE	INPUT	10	684	2	58.48	1	18.45
TRAIN	WEATHER_CODE	INPUT	12	95	6.01	56.37	3	11.41
TRAIN	REPORT_TYPE	TARGET	2	0	0	72.03	1	27.97

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

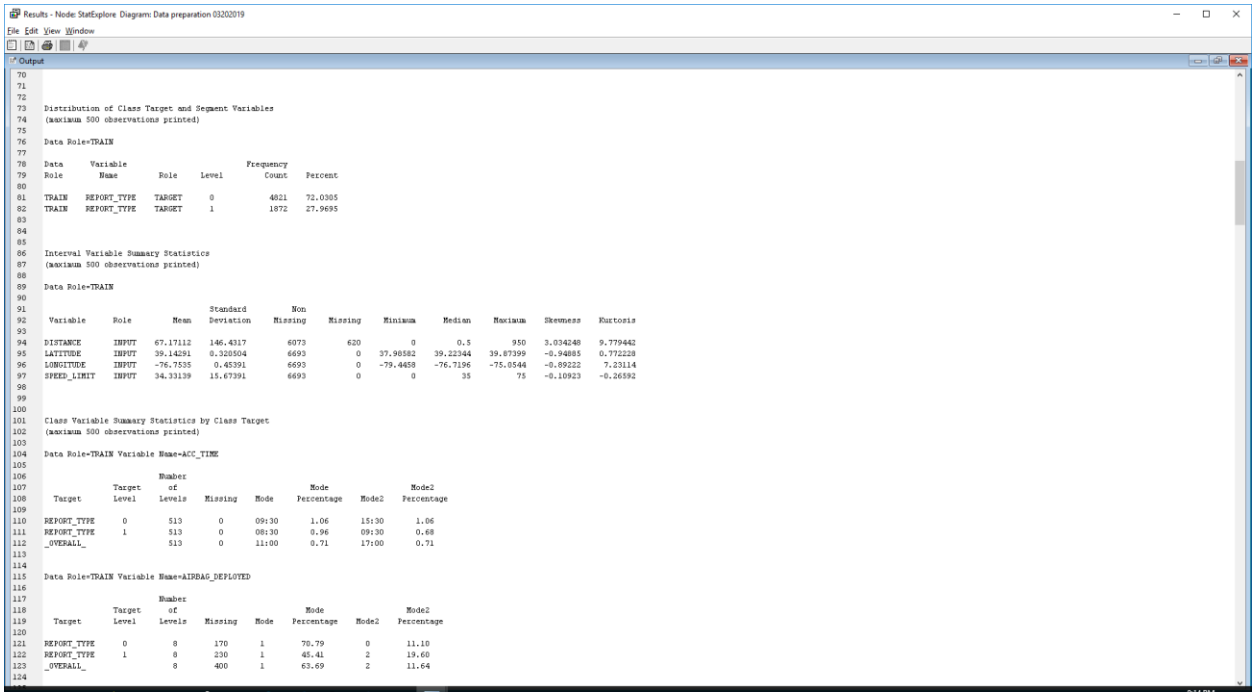
Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	REPORT_TYPE	TARGET	0	4821	72.0305
TRAIN	REPORT_TYPE	TARGET	1	1872	27.9695

Interval Variable Summary Statistics
(maximum 500 observations printed)

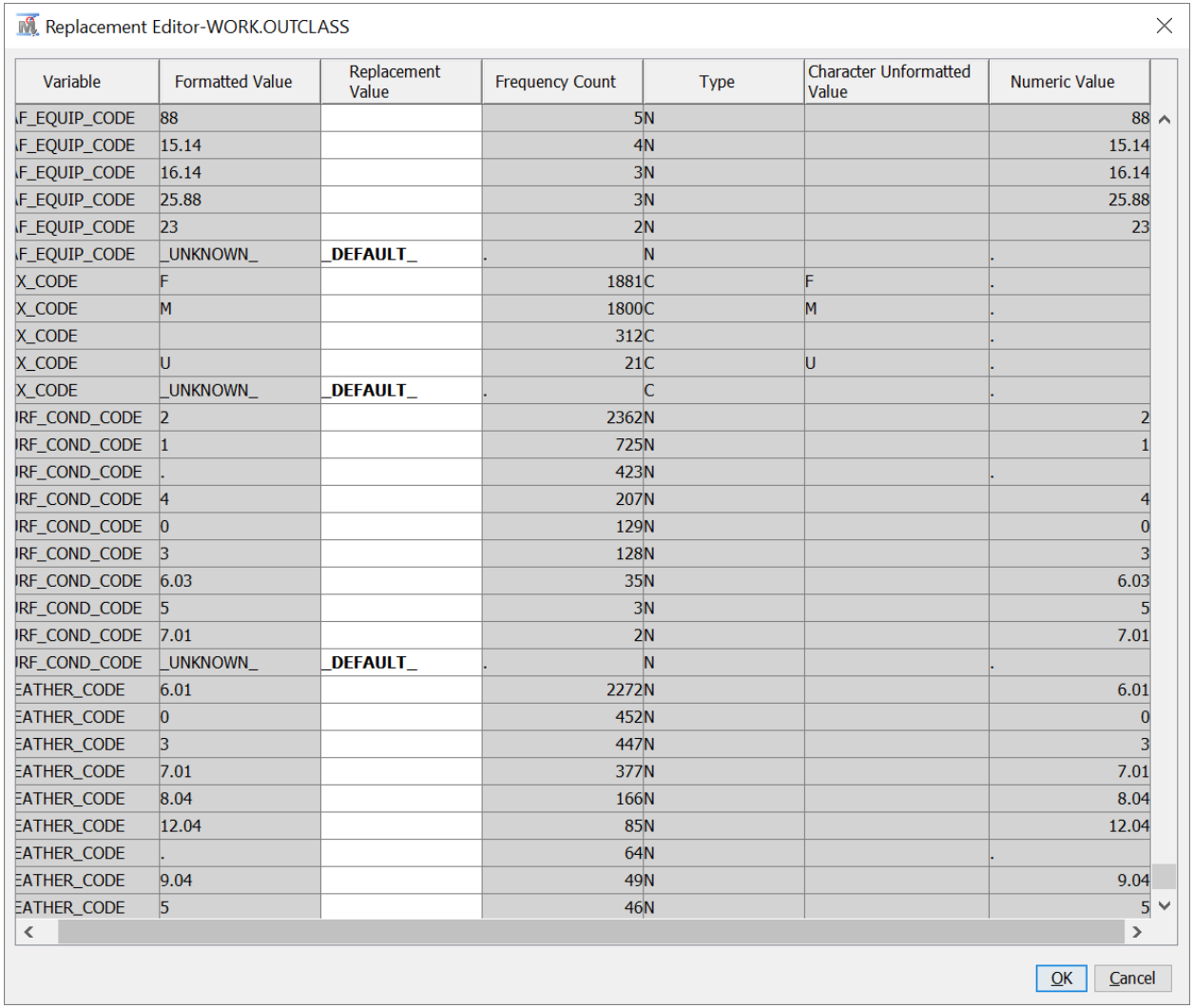
Interval Variable Summary Statistics

Interval Variable Summary Statistics



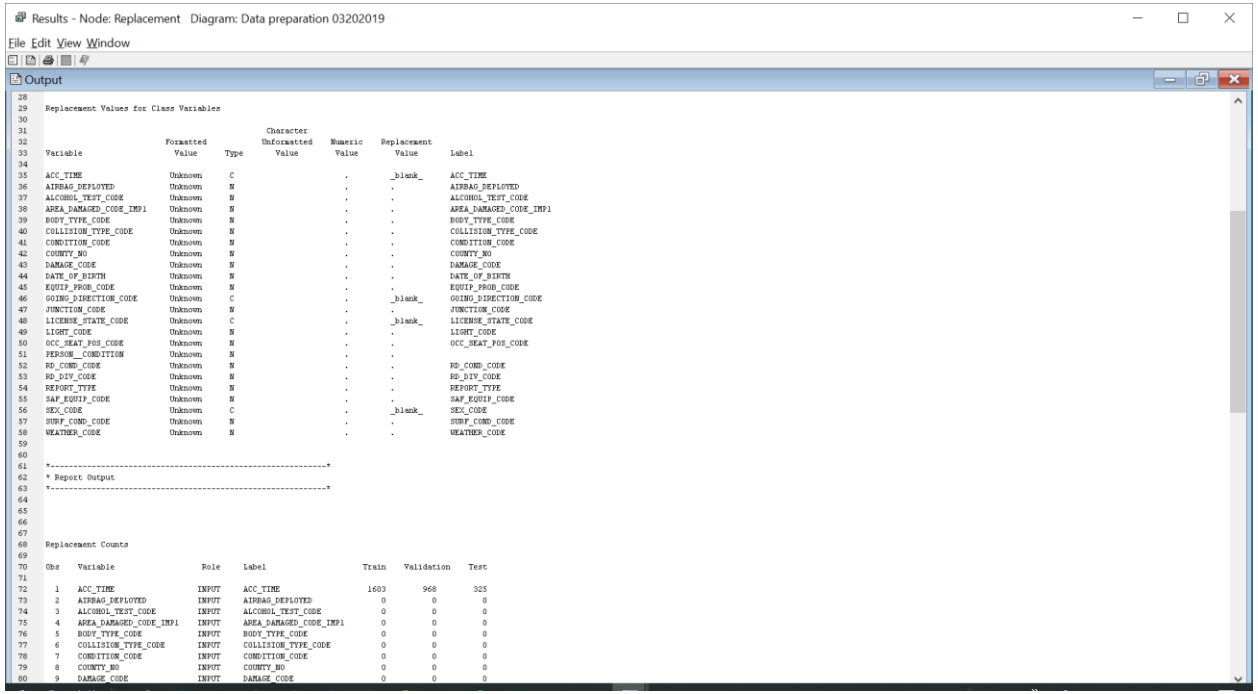
Appendix 6 Variable Clustering to see correlation between interval variable

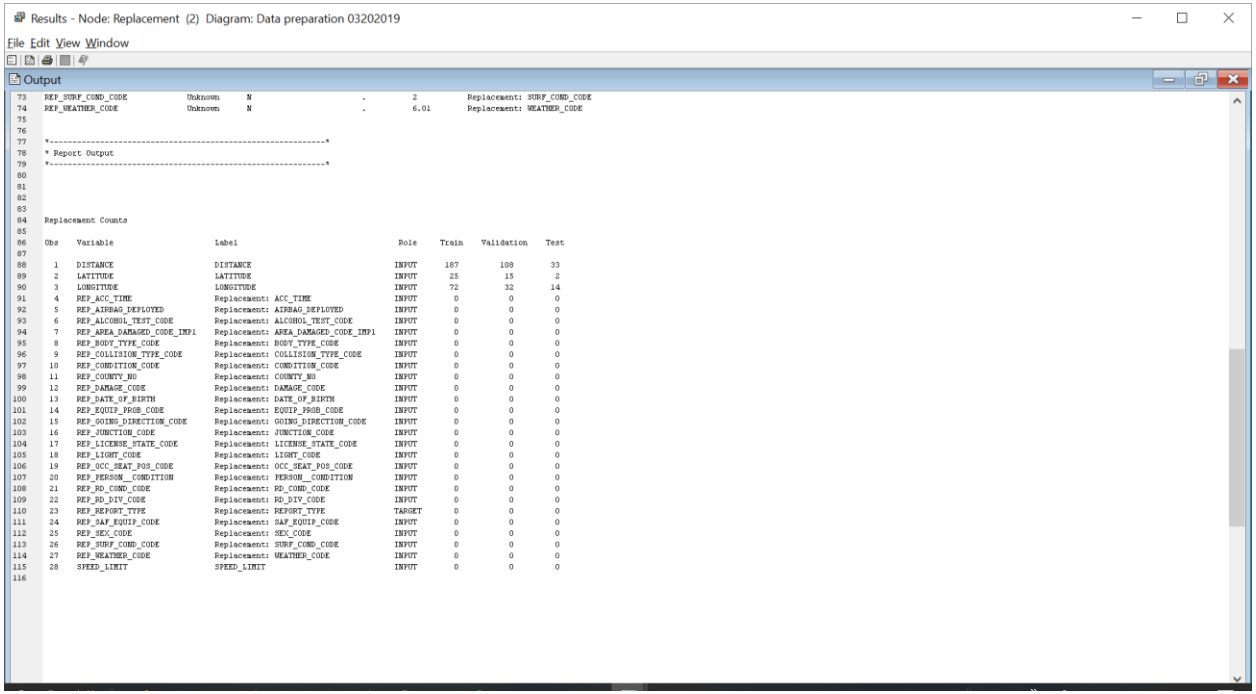
Appendix 7 Replacement Editor

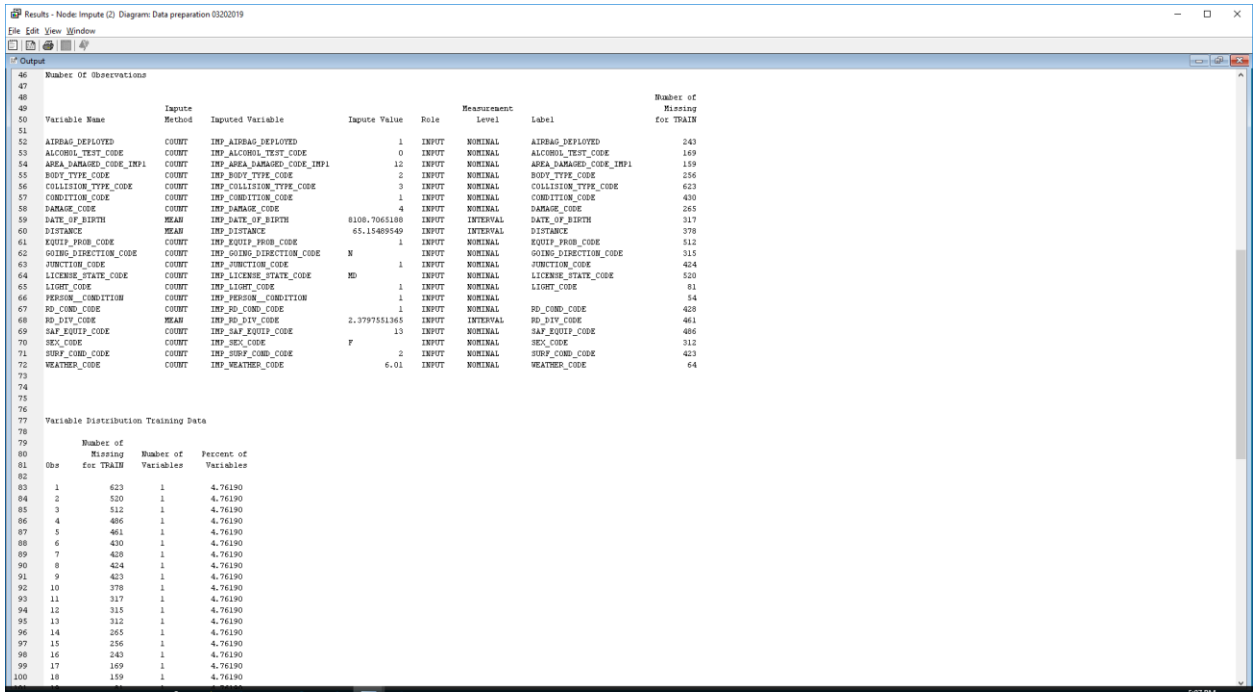

Replacement Editor-WORK.OUTCLASS
Variable
Formatted Value
Replacement Value
Frequency Count
Type
Character Unformatted Value
Numeric Value
AF_EQUIP_CODE 88 5 N 88
AF_EQUIP_CODE 15.14 4 N 15.14
AF_EQUIP_CODE 16.14 3 N 16.14
AF_EQUIP_CODE 25.88 3 N 25.88
AF_EQUIP_CODE 23 2 N 23
AF_EQUIP_CODE _UNKNOWN_ _DEFAULT_ . N .
AX_CODE F 1881 C F .
AX_CODE M 1800 C M .
AX_CODE 312 C .
AX_CODE U 21 C U .
AX_CODE _UNKNOWN_ _DEFAULT_ . C .
AIRF_COND_CODE 2 2362 N 2
AIRF_COND_CODE 1 725 N 1
AIRF_COND_CODE . 423 N .
AIRF_COND_CODE 4 207 N 4
AIRF_COND_CODE 0 129 N 0
AIRF_COND_CODE 3 128 N 3
AIRF_COND_CODE 6.03 35 N 6.03
AIRF_COND_CODE 5 3 N 5
AIRF_COND_CODE 7.01 2 N 7.01
AIRF_COND_CODE _UNKNOWN_ _DEFAULT_ . N .
AEATHER_CODE 6.01 2272 N 6.01
AEATHER_CODE 0 452 N 0
AEATHER_CODE 3 447 N 3
AEATHER_CODE 7.01 377 N 7.01
AEATHER_CODE 8.04 166 N 8.04
AEATHER_CODE 12.04 85 N 12.04
AEATHER_CODE . 64 N .
AEATHER_CODE 9.04 49 N 9.04
AEATHER_CODE 5 46 N 5
OK
Cancel

Appendix 8

Replacing unknown with missing value or mode for class variable




Results - Node: Replacement (2) Diagram: Data preparation 03202019
File Edit View Window
Output
REP_SURF_COND_CODE Unknown N . 2 Replacement: SURF_COND_CODE
REP_WEATHER_CODE Unknown N . 6.01 Replacement: WEATHER_CODE
* Report Output
Replacement Counts
Obs Variable Label Role Train Validation Test
1 DISTANCE DISTANCE INPUT 187 108 33
2 LATITUDE LATITUDE INPUT 25 15 2
3 LONGITUDE LONGITUDE INPUT 72 32 14
4 REP_ACC_TIME Replacement: ACC_TIME INPUT 0 0 0
5 REP_AIRBAG_DEPLOYED Replacement: AIRBAG_DEPLOYED INPUT 0 0 0
6 REP_ALCOHOL_TEST_CODE Replacement: ALCOHOL_TEST_CODE INPUT 0 0 0
7 REP_AREA_DAMAGED_CODE_IMP1 Replacement: AREA_DAMAGED_CODE_IMP1 INPUT 0 0 0
8 REP_BODY_TYPE_CODE Replacement: BODY_TYPE_CODE INPUT 0 0 0
9 REP_COLLISION_TYPE_CODE Replacement: COLLISION_TYPE_CODE INPUT 0 0 0
10 REP_CONDITION_CODE Replacement: CONDITION_CODE INPUT 0 0 0
11 REP_COUNTY_NO Replacement: COUNTY_NO INPUT 0 0 0
12 REP_DAMAGE_CODE Replacement: DAMAGE_CODE INPUT 0 0 0
13 REP_DATE_OF_BIRTH Replacement: DATE_OF_BIRTH INPUT 0 0 0
14 REP_EQUIP_PROB_CODE Replacement: EQUIP_PROB_CODE INPUT 0 0 0
15 REP_GOING_DIRECTION_CODE Replacement: GOING_DIRECTION_CODE INPUT 0 0 0
16 REP_JUNCTION_CODE Replacement: JUNCTION_CODE INPUT 0 0 0
17 REP_LICENSE_STATE_CODE Replacement: LICENSE_STATE_CODE INPUT 0 0 0
18 REP_LIGHT_CODE Replacement: LIGHT_CODE INPUT 0 0 0
19 REP_OCC_SEAT_POS_CODE Replacement: OCC_SEAT_POS_CODE INPUT 0 0 0
20 REP_PERSON__CONDITION Replacement: PERSON__CONDITION INPUT 0 0 0
21 REP_RD_COND_CODE Replacement: RD_COND_CODE INPUT 0 0 0
22 REP_RD_DIV_CODE Replacement: RD_DIV_CODE INPUT 0 0 0
23 REP_REPORT_TYPE Replacement: REPORT_TYPE TARGET 0 0 0
24 REP_SAF_EQUIP_CODE Replacement: SAF_EQUIP_CODE INPUT 0 0 0
25 REP_SEX_CODE Replacement: SEX_CODE INPUT 0 0 0
26 REP_SURF_COND_CODE Replacement: SURF_COND_CODE INPUT 0 0 0
27 REP_WEATHER_CODE Replacement: WEATHER_CODE INPUT 0 0 0
28 SPEED_LIMIT SPEED_LIMIT INPUT 0 0 0

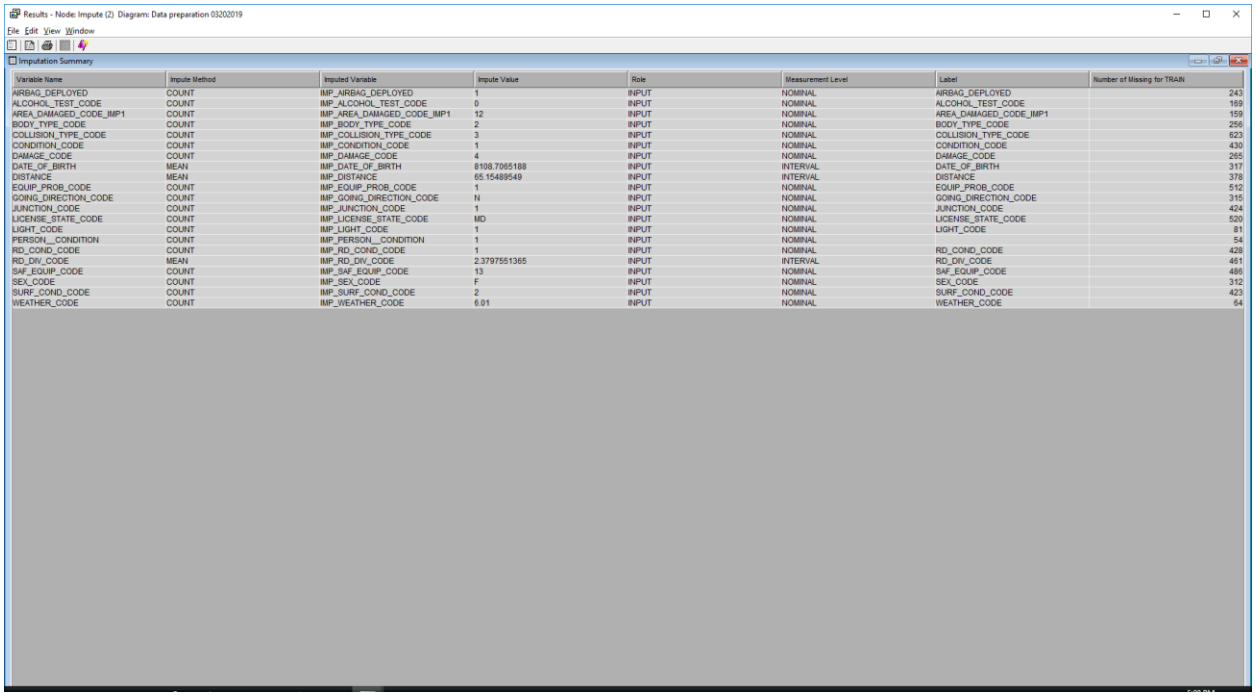


Number Of Observations

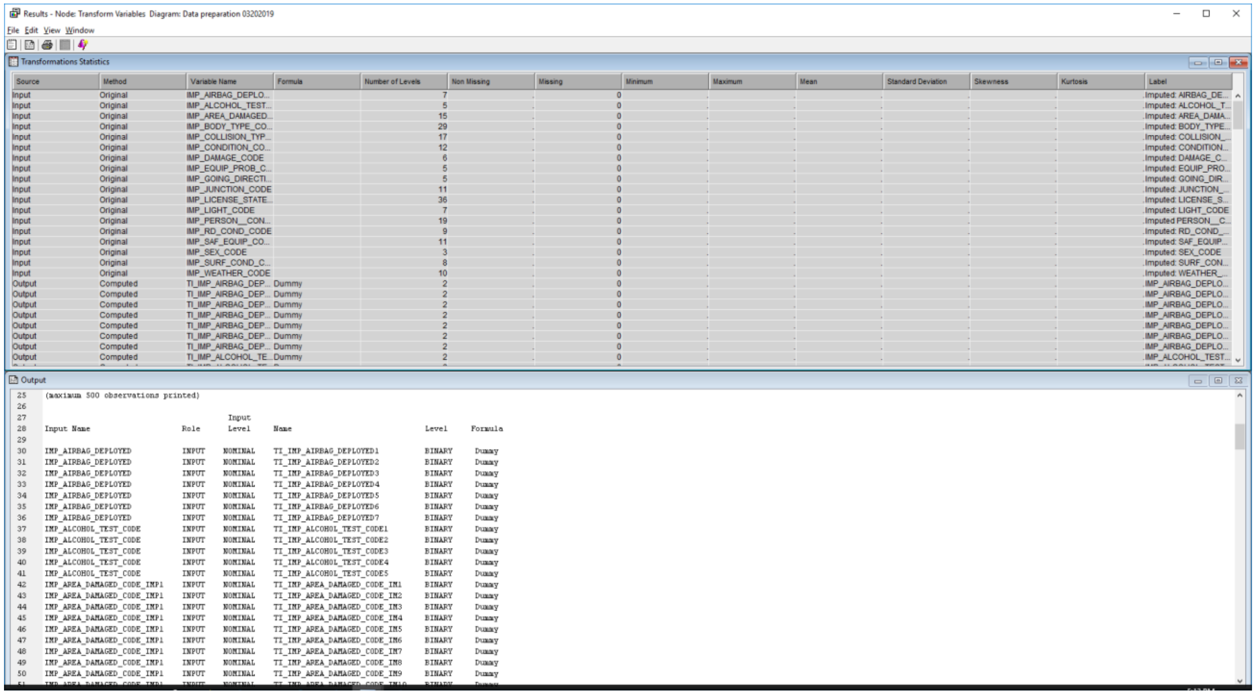
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
AIRBAG_DEPLOYED	COUNT	IMP_AIRBAG_DEPLOYED	1	INPUT	NOMINAL	AIRBAG_DEPLOYED	243
ALCOHOL_TEST_CODE	COUNT	IMP_ALCOHOL_TEST_CODE	0	INPUT	NOMINAL	ALCOHOL_TEST_CODE	169
AREA_DAMAGED_CODE_IMP1	COUNT	IMP_AREA_DAMAGED_CODE_IMP1	12	INPUT	NOMINAL	AREA_DAMAGED_CODE_IMP1	159
BODY_TYPE_CODE	COUNT	IMP_BODY_TYPE_CODE	2	INPUT	NOMINAL	BODY_TYPE_CODE	256
COLLISION_TYPE_CODE	COUNT	IMP_COLLISION_TYPE_CODE	3	INPUT	NOMINAL	COLLISION_TYPE_CODE	623
CONDITION_CODE	COUNT	IMP_CONDITION_CODE	1	INPUT	NOMINAL	CONDITION_CODE	430
DAMAGE_CODE	COUNT	IMP_DAMAGE_CODE	4	INPUT	NOMINAL	DAMAGE_CODE	265
DATE_OF_BIRTH	MEAN	IMP_DATE_OF_BIRTH	8108.7065180	INPUT	INTERVAL	DATE_OF_BIRTH	317
DISTANCE	MEAN	IMP_DISTANCE	65.15489549	INPUT	INTERVAL	DISTANCE	378
EQUIP_PROB_CODE	COUNT	IMP_EQUIP_PROB_CODE	1	INPUT	NOMINAL	EQUIP_PROB_CODE	512
GOING_DIRECTION_CODE	COUNT	IMP_GOING_DIRECTION_CODE	N	INPUT	NOMINAL	GOING_DIRECTION_CODE	315
JUNCTION_CODE	COUNT	IMP_JUNCTION_CODE	1	INPUT	NOMINAL	JUNCTION_CODE	424
LICENSE_STATE_CODE	COUNT	IMP_LICENSE_STATE_CODE	MD	INPUT	NOMINAL	LICENSE_STATE_CODE	520
LIGHT_CODE	COUNT	IMP_LIGHT_CODE	1	INPUT	NOMINAL	LIGHT_CODE	81
PERSON__CONDITION	COUNT	IMP_PERSON__CONDITION	1	INPUT	NOMINAL		54
RD_COND_CODE	COUNT	IMP_RD_COND_CODE	1	INPUT	NOMINAL	RD_COND_CODE	428
RD_DIV_CODE	MEAN	IMP_RD_DIV_CODE	2.3797551365	INPUT	INTERVAL	RD_DIV_CODE	461
SAF_EQUIP_CODE	COUNT	IMP_SAF_EQUIP_CODE	13	INPUT	NOMINAL	SAF_EQUIP_CODE	486
SEX_CODE	COUNT	IMP_SEX_CODE	F	INPUT	NOMINAL	SEX_CODE	312
SURF_COND_CODE	COUNT	IMP_SURF_COND_CODE	2	INPUT	NOMINAL	SURF_COND_CODE	423
WEATHER_CODE	COUNT	IMP_WEATHER_CODE	6.01	INPUT	NOMINAL	WEATHER_CODE	64

Variable Distribution Training Data

Obs	Number of Missing for TRAIN	Number of Variables	Percent of Variables
1	623	1	4.76190
2	520	1	4.76190
3	512	1	4.76190
4	486	1	4.76190
5	461	1	4.76190
6	430	1	4.76190
7	428	1	4.76190
8	424	1	4.76190
9	423	1	4.76190
10	378	1	4.76190
11	317	1	4.76190
12	315	1	4.76190
13	312	1	4.76190
14	265	1	4.76190
15	256	1	4.76190
16	243	1	4.76190
17	169	1	4.76190
18	159	1	4.76190



Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
AIRBAG_DEPLOYED	COUNT	IMP_AIRBAG_DEPLOYED	1	INPUT	NOMINAL	AIRBAG_DEPLOYED	243
ALCOHOL_TEST_CODE	COUNT	IMP_ALCOHOL_TEST_CODE	0	INPUT	NOMINAL	ALCOHOL_TEST_CODE	169
AREA_DAMAGED_CODE_IMP1	COUNT	IMP_AREA_DAMAGED_CODE_IMP1	12	INPUT	NOMINAL	AREA_DAMAGED_CODE_IMP1	159
BODY_TYPE_CODE	COUNT	IMP_BODY_TYPE_CODE	2	INPUT	NOMINAL	BODY_TYPE_CODE	256
COLLISION_TYPE_CODE	COUNT	IMP_COLLISION_TYPE_CODE	3	INPUT	NOMINAL	COLLISION_TYPE_CODE	623
CONDITION_CODE	COUNT	IMP_CONDITION_CODE	1	INPUT	NOMINAL	CONDITION_CODE	430
DAMAGE_CODE	COUNT	IMP_DAMAGE_CODE	4	INPUT	NOMINAL	DAMAGE_CODE	265
DATE_OF_BIRTH	MEAN	IMP_DATE_OF_BIRTH	8108.7065188	INPUT	INTERVAL	DATE_OF_BIRTH	317
DISTANCE	MEAN	IMP_DISTANCE	65.15489549	INPUT	INTERVAL	DISTANCE	378
EQUIP_PROB_CODE	COUNT	IMP_EQUIP_PROB_CODE	1	INPUT	NOMINAL	EQUIP_PROB_CODE	512
GOING_DIRECTION_CODE	COUNT	IMP_GOING_DIRECTION_CODE	N	INPUT	NOMINAL	GOING_DIRECTION_CODE	315
JUNCTION_CODE	COUNT	IMP_JUNCTION_CODE	1	INPUT	NOMINAL	JUNCTION_CODE	424
LICENSE_STATE_CODE	COUNT	IMP_LICENSE_STATE_CODE	MD	INPUT	NOMINAL	LICENSE_STATE_CODE	520
LIGHT_CODE	COUNT	IMP_LIGHT_CODE	1	INPUT	NOMINAL	LIGHT_CODE	81
PERSON__CONDITION	COUNT	IMP_PERSON__CONDITION	1	INPUT	NOMINAL		54
RD_COND_CODE	COUNT	IMP_RD_COND_CODE	1	INPUT	NOMINAL	RD_COND_CODE	428
RD_DIV_CODE	MEAN	IMP_RD_DIV_CODE	2.3797551365	INPUT	INTERVAL	RD_DIV_CODE	461
SAF_EQUIP_CODE	COUNT	IMP_SAF_EQUIP_CODE	13	INPUT	NOMINAL	SAF_EQUIP_CODE	486
SEX_CODE	COUNT	IMP_SEX_CODE	F	INPUT	NOMINAL	SEX_CODE	312
SURF_COND_CODE	COUNT	IMP_SURF_COND_CODE	2	INPUT	NOMINAL	SURF_COND_CODE	423
WEATHER_CODE	COUNT	IMP_WEATHER_CODE	6.01	INPUT	NOMINAL	WEATHER_CODE	64



Results - Node: Transform Variables Diagram: Data preparation 03202019

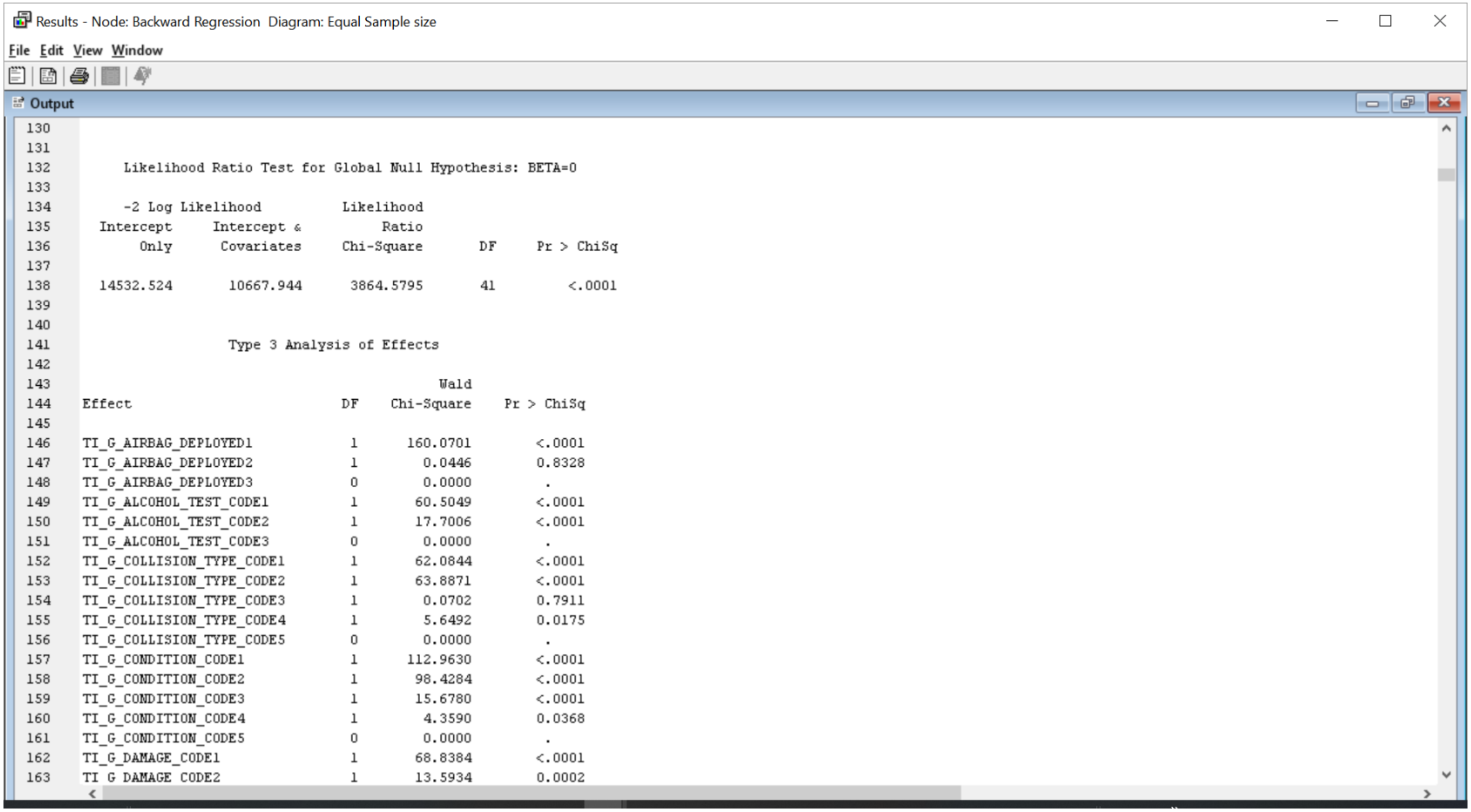
File Edit View Window

Transformations Statistics

Source	Method	Variable Name	Formula	Number of Levels	Non Missing	Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Label
Input	Original	IMP_AIRBAG_DEPLO...		7	.	0	.	.	.	.	.	.	Imputed: AIRBAG_DE...
Input	Original	IMP_ALCOHOL_TEST...		5	.	0	.	.	.	.	.	.	Imputed: ALCOHOL_T...
Input	Original	IMP_AREA_DAMAGED...		15	.	0	.	.	.	.	.	.	Imputed: AREA_DAMA...
Input	Original	IMP_BODY_TYPE_CO...		29	.	0	.	.	.	.	.	.	Imputed: BODY_TYPE...
Input	Original	IMP_COLLISION_TYP...		17	.	0	.	.	.	.	.	.	Imputed: COLLISION_...
Input	Original	IMP_CONDITION_CO...		12	.	0	.	.	.	.	.	.	Imputed: CONDITION...
Input	Original	IMP_DAMAGE_CODE		6	.	0	.	.	.	.	.	.	Imputed: DAMAGE_C...
Input	Original	IMP_EQUIP_PROB_C...		5	.	0	.	.	.	.	.	.	Imputed: EQUIP_PRO...
Input	Original	IMP_GOING_DIRECTI...		5	.	0	.	.	.	.	.	.	Imputed: GOING_DIR...
Input	Original	IMP_JUNCTION_CODE		11	.	0	.	.	.	.	.	.	Imputed: JUNCTION_...
Input	Original	IMP_LICENSE_STATE...		36	.	0	.	.	.	.	.	.	Imputed: LICENSE_S...
Input	Original	IMP_LIGHT_CODE		7	.	0	.	.	.	.	.	.	Imputed: LIGHT_CODE
Input	Original	IMP_PERSON__CON...		19	.	0	.	.	.	.	.	.	Imputed PERSON__C...
Input	Original	IMP_RD_COND_CODE		9	.	0	.	.	.	.	.	.	Imputed: RD_COND_...
Input	Original	IMP_SAF_EQUIP_CO...		11	.	0	.	.	.	.	.	.	Imputed: SAF_EQUIP...
Input	Original	IMP_SEX_CODE		3	.	0	.	.	.	.	.	.	Imputed: SEX_CODE
Input	Original	IMP_SURF_COND_C...		8	.	0	.	.	.	.	.	.	Imputed: SURF_CON...
Input	Original	IMP_WEATHER_CODE		10	.	0	.	.	.	.	.	.	Imputed: WEATHER_...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_AIRBAG_DEP...	Dummy	2	.	0	.	.	.	.	.	.	IMP_AIRBAG_DEPLO...
Output	Computed	TI_IMP_ALCOHOL_TE...	Dummy	2	.	0	.	.	.	.	.	.	IMP_ALCOHOL_TEST...

Output

```
(maximum 500 observations printed)


                                           Input
Input Name                      Role       Level       Name                              Level      Formula

IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED1           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED2           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED3           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED4           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED5           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED6           BINARY     Dummy
IMP_AIRBAG_DEPLOYED             INPUT      NOMINAL     TI_IMP_AIRBAG_DEPLOYED7           BINARY     Dummy
IMP_ALCOHOL_TEST_CODE           INPUT      NOMINAL     TI_IMP_ALCOHOL_TEST_CODE1         BINARY     Dummy
IMP_ALCOHOL_TEST_CODE           INPUT      NOMINAL     TI_IMP_ALCOHOL_TEST_CODE2         BINARY     Dummy
IMP_ALCOHOL_TEST_CODE           INPUT      NOMINAL     TI_IMP_ALCOHOL_TEST_CODE3         BINARY     Dummy
IMP_ALCOHOL_TEST_CODE           INPUT      NOMINAL     TI_IMP_ALCOHOL_TEST_CODE4         BINARY     Dummy
IMP_ALCOHOL_TEST_CODE           INPUT      NOMINAL     TI_IMP_ALCOHOL_TEST_CODE5         BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM1      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM2      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM3      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM4      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM5      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM6      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM7      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM8      BINARY     Dummy
IMP_AREA_DAMAGED_CODE_IMP1      INPUT      NOMINAL     TI_IMP_AREA_DAMAGED_CODE_IM9      BINARY     Dummy
```

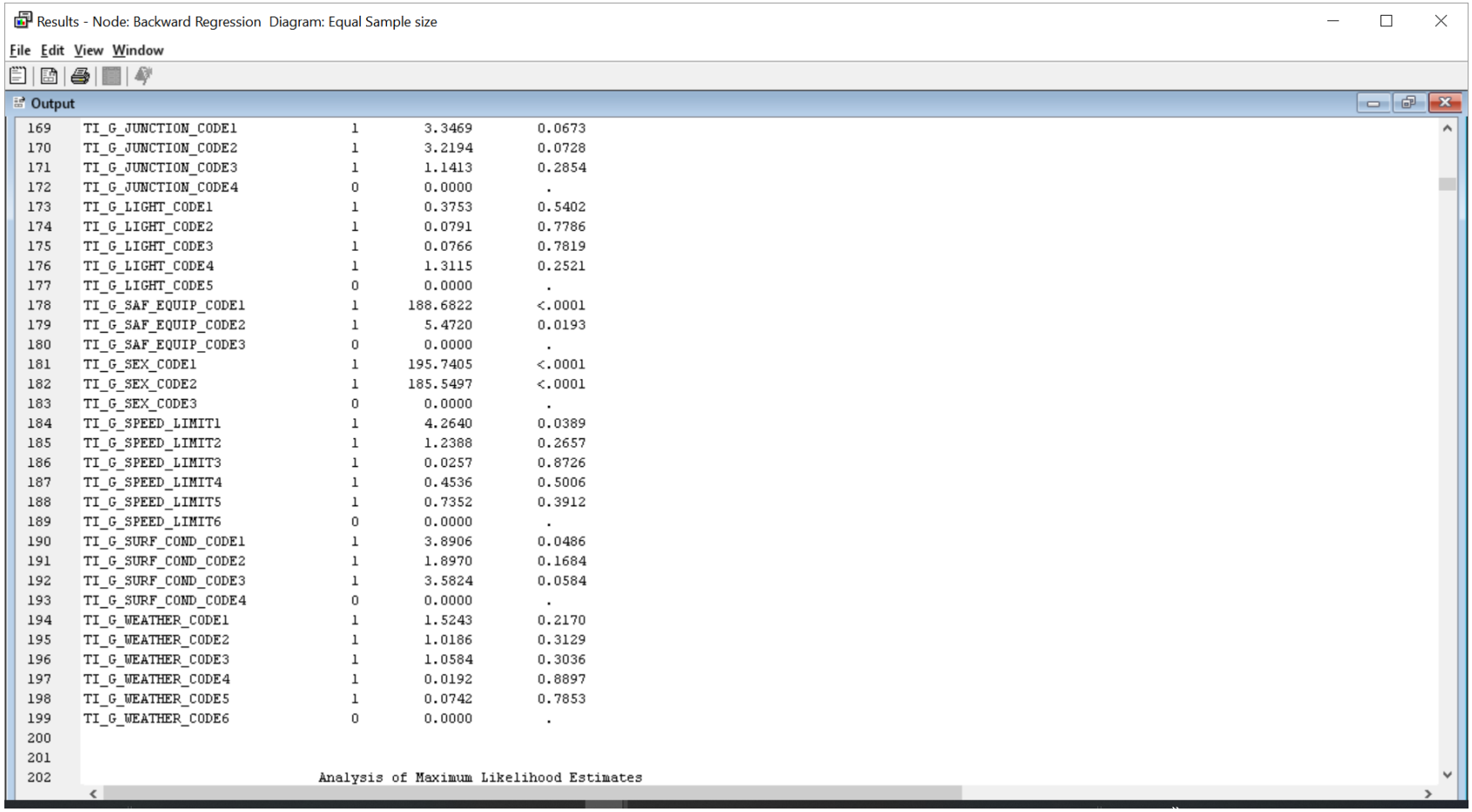


Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
14532.524	10667.944	3864.5795	41	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
TI_G_AIRBAG_DEPLOYED1	1	160.0701	<.0001
TI_G_AIRBAG_DEPLOYED2	1	0.0446	0.8328
TI_G_AIRBAG_DEPLOYED3	0	0.0000	.
TI_G_ALCOHOL_TEST_CODE1	1	60.5049	<.0001
TI_G_ALCOHOL_TEST_CODE2	1	17.7006	<.0001
TI_G_ALCOHOL_TEST_CODE3	0	0.0000	.
TI_G_COLLISION_TYPE_CODE1	1	62.0844	<.0001
TI_G_COLLISION_TYPE_CODE2	1	63.8871	<.0001
TI_G_COLLISION_TYPE_CODE3	1	0.0702	0.7911
TI_G_COLLISION_TYPE_CODE4	1	5.6492	0.0175
TI_G_COLLISION_TYPE_CODE5	0	0.0000	.
TI_G_CONDITION_CODE1	1	112.9630	<.0001
TI_G_CONDITION_CODE2	1	98.4284	<.0001
TI_G_CONDITION_CODE3	1	15.6780	<.0001
TI_G_CONDITION_CODE4	1	4.3590	0.0368
TI_G_CONDITION_CODE5	0	0.0000	.
TI_G_DAMAGE_CODE1	1	68.8384	<.0001
TI_G_DAMAGE_CODE2	1	13.5934	0.0002



169	TI_G_JUNCTION_CODE1	1	3.3469	0.0673
170	TI_G_JUNCTION_CODE2	1	3.2194	0.0728
171	TI_G_JUNCTION_CODE3	1	1.1413	0.2854
172	TI_G_JUNCTION_CODE4	0	0.0000	.
173	TI_G_LIGHT_CODE1	1	0.3753	0.5402
174	TI_G_LIGHT_CODE2	1	0.0791	0.7786
175	TI_G_LIGHT_CODE3	1	0.0766	0.7819
176	TI_G_LIGHT_CODE4	1	1.3115	0.2521
177	TI_G_LIGHT_CODE5	0	0.0000	.
178	TI_G_SAF_EQUIP_CODE1	1	188.6822	<.0001
179	TI_G_SAF_EQUIP_CODE2	1	5.4720	0.0193
180	TI_G_SAF_EQUIP_CODE3	0	0.0000	.
181	TI_G_SEX_CODE1	1	195.7405	<.0001
182	TI_G_SEX_CODE2	1	185.5497	<.0001
183	TI_G_SEX_CODE3	0	0.0000	.
184	TI_G_SPEED_LIMIT1	1	4.2640	0.0389
185	TI_G_SPEED_LIMIT2	1	1.2388	0.2657
186	TI_G_SPEED_LIMIT3	1	0.0257	0.8726
187	TI_G_SPEED_LIMIT4	1	0.4536	0.5006
188	TI_G_SPEED_LIMIT5	1	0.7352	0.3912
189	TI_G_SPEED_LIMIT6	0	0.0000	.
190	TI_G_SURF_COND_CODE1	1	3.8906	0.0486
191	TI_G_SURF_COND_CODE2	1	1.8970	0.1684
192	TI_G_SURF_COND_CODE3	1	3.5824	0.0584
193	TI_G_SURF_COND_CODE4	0	0.0000	.
194	TI_G_WEATHER_CODE1	1	1.5243	0.2170
195	TI_G_WEATHER_CODE2	1	1.0186	0.3129
196	TI_G_WEATHER_CODE3	1	1.0584	0.3036
197	TI_G_WEATHER_CODE4	1	0.0192	0.8897
198	TI_G_WEATHER_CODE5	1	0.0742	0.7853
199	TI_G_WEATHER_CODE6	0	0.0000	.
200				
201				
202	Analysis of Maximum Likelihood Estimates			