University of Maryland University College

DATA 640 – Predictive Modeling

SUMMER 2017

Assignment 2-Logistic Regression

Testing Eight Logistic Models in SAS Enterprise Miner to Identify the Most Important Variables Initiating a Non-Fatal Injury

John Parsons

jparsons20@student.umuc.edu

Professor Knode

<u>Introduction</u>

Laura King (2016) from the Huffington Post discusses the top 15 causes of car accidents. The number one cause is distracted drivers (which was not a variable in this study) and can be witnessed anytime on the interstate today with the popularity of the smart phones. The second cause would be reckless driving and then the article talks about speeding, road conditions, weather, construction zones, age of driver and animal interactions. The US Department of Transportation (NHTSA, 2016) recorded a slight increase (3.8%) in non-fatal crashes from 6,034,000 million in 2014 up to 6,264,000 million in 2015.

The 2005 Accidents Database came from the United States Department of Transportation and was chosen to see which factors are the most important in contributing to non-fatal injuries to motorists. This 2005 database contains a total of 24 attributes and 42,183 data points. The dependent or response variable for this study is **MAX_SEV_I** (type of injury recorded from the accident). The independent variables contain a total of 23 attributes such as; **PED_ACC_R (are pedestrians or cyclists involved), WKDY_I_R** (did the incident occur during the weekday or weekend), **ALCOHOL_I** (was alcohol involved), **SPD_LIM** (how fast the car was traveling), **WEATHER_R** (weather conditions) **and WRK_ZONE** (was this a work zone) to name a few**.** The list of variables can be found in **Figure 1A** (A refers to Appendix) and the screenshot of this database can be found in **Figure 1** below.

All attributes in this study were numerical and the majority these were recorded as a binary response variable (0 = No and 1= Yes). The independent variables had a total of 15 binary levels, 3 Ordinal and 5 Nominal levels (**Figure 1A**). The dependent variable was a trinary variable and the third level was fatality. The total number of fatalities that occurred in this study was 466 out of 42,184 data points (1.1% of the entire dataset). This subset was removed from this study which produced 41,717 data points and will be discussed in the next section.

The goal of this study is to select from a total of eight Logistic Regression models the model that correctly identifies the most accidents that were recorded as a non-fatal injury. The model will determine which factors need to be addressed to increase passenger safety for the US Department of Transportation. The criteria used to select the best model(s) are the True Positive values and Accuracy Rates from the Confusion Matrix, the Validation Misclassification Index from the Fit Statistics, ROC Index and the number of parameters used for the model.

Figure 1: Screenshot of the database in SAS Enterprise Miner.

HOUR_I_R	ALCHL_I	ALIGN_I	STRATUM_R	WRK_ZONE	WKDY_I_R	INT_HWY	LGTCON_I_R	MANCOL_I_R	PED_ACC_R	RELJCT_I_R	REL_RWY_R	PROFIL_I_R	SPD_LIM	SUR_COND	TRAF_CON_R	TRAF_WAY	VEH_INVL	WEATHER_R	INJURY_CRASH	NO_INJ_I	PRPTYDMG_CRASH	FATALITIES	MAX_SEV_IR
1	2	1	0	0	1	1	3	2	0	1	1	1	70	4	0	3	2	2	0	0	1	0	0
1	2	1	0	0	1	0	3	2	0	1	1	1	35	4	1	2	2	2	0	0	1	0	0
1	2	1	1	0	0	0	3	2	0	1	1	1	35	4	1	2	2	1	0	0	1	0	0
1	1	1	0	0	1	0	3	2	0	0	1	1	25	4	0	2	3	1	0	0	1	0	0
1	2	1	0	0	1	1	3	0	0	0	0	1	70	4	0	2	1	2	0	0	1	0	0
1	2	1	1	0	1	0	3	0	0	1	0	1	30	4	0	1	1	2	0	0	1	0	0
0	2	1	0	0	0	0	3	0	0	1	0	1	25	4	0	1	1	2	0	0	1	0	0
1	2	1	0	0	1	0	3	0	0	0	0	1	55	4	0	1	1	2	0	0	1	0	0
1	2	1	1	0	1	0	3	2	0	0	1	1	40	4	2	1	2	1	0	0	1	0	0
1	2	2	0	0	1	0	3	0	0	0	0	1	25	4	0	1	1	1	0	0	1	0	0
1	2	1	1	0	1	0	3	0	0	0	0	1	20	4	0	1	1	2	0	0	1	0	0
1	2	1	1	0	0	0	3	0	0	0	0	1	50	4	0	1	1	2	0	0	1	0	0

Data Cleaning and Preparation

The original Accidents database contained 42,183 data points with a total of 24 attributes. The Target Variable in this study was trinomial (with the addition of fatalities) and the number of fatalities was 1.1 percent of the total data points. The fatality data points were removed for two reasons. First, this study is only concerned with non-fatal accidents and which variables are the best predictor in contributing to these accidents. Second, the Target variable was highly recommended to be binary to create several Logistic Models using SAS Enterprise Miner. The best model will then be used to test the original database with a trinomial variable and explained briefly in the Appendix (Figure 10A and 11A).

The **STATEXPLORE** Node was used to look at the vital statistics and generate histograms for the variables (**Figure 2A and 3A**, Appendix). Regressions and Neural Networks need to have imputed values for missing data points and imputation was not needed for this dataset because there are no missing values. The histogram and statistics from

STATEXPLORE shows how the attributes were spread out in this study. The Max Normal would be selected for the Transformation Node, but there were no interval attributes to be smoothed over due to skewness of the data. The five nominal variables in this study were "exploded" into Dummy Variables as recommended by Abbot (2014) and Dr. Knode. The two Nominal Variables in this study (**NO_INJ_I** was removed from this study), **SPD_LIM** and **VEH_INVL** were binned into levels using the Replacement Node in SAS. The **SPD_LMT** was binned into a total of seven bins from a total of 15 categories. The speed limit was divided into units of 10 (not 5) miles per hour. The second replacement node was conducted on **VEH_INVL** and contained a total of 10 levels. If a car had one, two or three cars that were involved in an accident, it was placed in a separate category. All car accidents that involved four or more cars were binned into the fourth category or four.

A total of four attributes were removed from this dataset in the input node and these are: **INJURY_CRASH, NO_INJ_I, PRPTYDMG_CRASH** and **FATALITIES**. The first three variables were highly correlated (Homoscedasticity) to the Target Variable and the **FATALITIES** attribute was removed because this initial report is only concerned with which variables can be used to determine accidents with non-lethal injury. The Target Variable in this study is already measuring if a motorist has an injury or not and this study is not concerned with property damage as the result of the crash so this variable was removed.

<u>Predictive Models Developed</u>

A total of eight models were developed for this study and the model criteria can be seen in **Figure 2** below. The eight models are **A_Gradient Boosting (Figure 8A), B_Default No Transformation, C_Backward, D_Forward, E_Stepwise, F_Cutoff with Stepwise,**

G_Variable Selection with Stepwise and H_Polynomial Logistic Regression. The standard three Regression Models (Backward, Forward and Stepwise) all had the similar values for the Fit Statistics during the initial run, so the stop and stay significance values were changed for the Backward and Stepwise Regression to try and develop more Accurate and True Positive results (Injury Accidents). The Polynomial Regression model was initiated to find all interactions and then only the significant ones as shown in Figure 2 were added to another **H_Polynomial** Regression Model.

Figure 2: Selection Criteria used to build all eight models.

Linear Regression	Model Modifications
A_Gradient Boost Regression	A Logistic Regression model node was added to the Gradient Boosting Node. The data was not transformed and the default settings (none) was used for Model Selection.
B_Default No Transformation Reg.	Logistic Regression with no data transformations and used the default (none) selection criteria.
C_Backward Regression	Logistic used for Backward Regression and Validation Misclassification was used for Selection Criteria. The Use Selection Default was set to NO and the Entry and Stay Significance was set to 0.25 and maximum number of steps was increased to 100.
D_Forward Regression	Logistic used for Forward Regression and Selection Criteria set to Validation Misclassification.
E_Stepwise Regression	Logistic used for Stepwise Regression and Selection Criteria set to Validation Misclassification. The Use Selection Default was set to NO and the Entry and Stay Significance was set to 0.0001 and maximum number of steps was increased to 100.
F_Cutoff	The Cutoff Node was attached to a Stepwise Regression node with Validation Misclassification for Selection Criteria. The Cutoff Method was changed to User Input and the value was set to 0.48.
G_Var_Sel. Stepwise Reg.	The variable selection node was added to the Logistic Stepwise Model (Selection Criteria using Validation Misclassification). The Chi Square (Logistic) was added to the Target Model for training the data.
H_Polyomial Regression	Logistic setting was used and the "User Terms" under Equation was changed to Yes. Default settings used for selection model. The Term Editor was opened and the following interactions were added based on the lowest p values from another Polynomial Regression that tested all interactions: RELJCT_I_R*REL_RWY_R (p value <.0001) ALCHL_I*REL_RWY_R (p value = 0.0016) ALIGN_I*STRATUM_R (p value = 0.0002) HOUR_I_R*WRK_ZONE (p value = 0.0035) STRATUM_R*WKDY_I_R (p value = 0.0280) REP_VEH_INVL*STRATUM_R (p value <.0001)

The total number of polynomial interactions were not included in this report due to the size of the file. The data was partitioned into the standard 55:25:20 split for the Training, Validation and Testing data and had the random number generator set to 12345 or default setting.

Results

The eight Logistic Regression Models were run in SAS Enterprise Miner 14 and the Fit Statistics generated from the Model Comparison node can be found in Figure 4A (Appendix). The Confusion Matrix results were transferred to a spreadsheet and the output results were used to calculate Accuracy, Specificity and Precision. The number of Iteration Steps and Parameter Estimates used for each model and the Confusion matrix results can be seen in Figure 3.

Figure 3: Confusion Matrix and Parameter Estimates from SAS Enterpriser Miner.

Event Classification Table	False Negative	True Negative	False Postive	True Positive	Accuracy	Specificity	Precision	Iteration Steps	Parameter Estimates
Reg2 B_Default Regression TRAIN MAX_SEV_IR	5226	7537	3859	6321	0.604	0.661	0.621	7	49
Reg2 B_Default Regression VALIDATE MAX_SEV_IR	2367	3409	1771	2882	0.603	0.658	0.619	7	49
Reg A_Regression TRAIN MAX_SEV_IR	5227	7542	3854	6320	0.604	0.662	0.621	6	32
Reg A_Regression VALIDATE MAX_SEV_IR	2367	3411	1769	2882	0.603	0.658	0.620	6	32
Reg4 D_Forward Regression TRAIN MAX_SEV_IR	5087	7389	4007	6460	0.604	0.648	0.617	10	19
Reg4 D_Forward Regression VALIDATE MAX_SEV_IR	2290	3383	1797	2959	0.608	0.653	0.622	10	19
Reg3 C_Backword Regression TRAIN MAX_SEV_IR	5236	7516	3880	6311	0.603	0.660	0.619	6	32
Reg3 C_Backword Regression VALIDATE MAX_SEV_IR	2338	3436	1744	2861	0.607	0.663	0.621	6	32
Reg5 E_Stepwise Regression TRAIN MAX_SEV_IR	5136	7415	3981	6411	0.603	0.651	0.617	9	18
Reg5 E_Stepwise Regression VALIDATE MAX_SEV_IR	2313	3392	1788	2936	0.607	0.655	0.622	9	18
Reg9 Cutoff Stepwise Regession TRAIN MAX_SEV_IR	5087	7389	4007	6460	0.604	0.648	0.617	**	**
Reg9 Cutoff Stepwise Regession VALIDATE MAX_SEV_IR	2290	3383	1797	2959	0.608	0.653	0.622	**	**
Reg8 G_Var. Sel. Stepwise Reg. TRAIN MAX_SEV_IR	5007	7283	4113	6540	0.602	0.639	0.614	7	16
Reg8 G_Var. Sel. Stepwise Reg. VALIDATE MAX_SEV_IR	2255	3282	1898	2994	0.602	0.634	0.612	7	16
Reg6 H_Polynomial Regression TRAIN MAX_SEV_IR	4828	7249	4147	6719	0.609	0.636	0.618	7	51
Reg6 H_Polynomial Regression VALIDATE MAX_SEV_IR	2222	3304	1876	3027	0.607	0.638	0.617	7	51

The goal of this study was to create a Logistic Model that identified the most accidents that resulted in a non-fatal injury by measuring the total number of True Positives, lowest Validation Misclassification Error, best Accuracy and ROC Validation Index. The number of Iteration steps for these models ranged from 6 (**C_Backward**) to 10 steps (**D_Forward**). The parameter estimates ranged from 16 for **G_Variable** to 51 for **H_Polynomial** Regression. The number of True Positives for Validation ranged from 2,861 (**C_Backward**) up to 3,027 for **H_Polynomial**. The Accuracy rates ranged from 0.602 (**G_Variable**) up to 0.608 for two models (**D_Forward and F_Cutoff**). Finally, the ROC Index ranged from 0.638 for **G_Variable** up to 0.650 for the **H_Polynomial**.

The **H_Polynomial** model generated the highest number of True Positives (3,027), had the second highest accuracy (0.607) with two other models, had the second lowest Validation Misclassification Rates (0.3941) and highest ROC Validation Index (0.65). The Polynomial also had the largest number of variables (51) and second lowest Iteration steps. The second-best model looking at True Positives was **G_Variable Selection** at 2,994 but had the highest Validation Misclassification Rates (0.3982). The third best model for the True Positives criteria was the **F_Cutoff** with a total of 2,959 and had the lowest Validation Misclassification Rates (0.392). The **C_Backup** Regression has the lowest True Positives (2861) and this was due to the change made in the stop and stay significance values.

The Receiver Operator Curves (ROC) in Figure 5A shows a visual representation of the Confusion Matrices of the Sensitivity (y-axis True Positive) and Specificity (x-axis False Positive) (Abbott, 2014). The models did not have a lot of separation between these extremes as shown in this graph.

Abbott (2012) and Dr. Knode for Lecture discusses how the best model really depends on the end goal or the Business Decision. The best decision for this study is to use the **G_Variable** Selection Model which had the second highest True Positives and the second least number of Iteration Steps. This model had the least number of parameters that can be easily explained to the public about any changes that need to be made for driver safety. The model variables are: **Intercept, PED_ACC_R, REL_RWY_R, REP_SPD_LIM, REP_VEH_INVL,**

STRATUM_R, TI_MANCOL_I_R2, and TI_SUR_COND1. The Wald Chi squared statistics can be seen in Figure 6A. The top four parameters based on the Wald's Chi square are: **PED_ACC_R** (265.28), **STRATUM** (257.54), **REP_VEH_INV** (245.83) and **REP_SPD_LMT** (69.53). These variables are the top four variables listed in the Chi-Squared Plot for the summary statistics in Figure 7A and are a few variables that need to be addressed to help minimize non-fatal accidents for the DOT.

Conclusions

Abbott (2014), Kattamuri (2013), Robie Video, SAS Books (2016) and Dr. Knode Class notes were used to create the Logistic Models and interpret them based on the criteria such as True Positives, Validation Misclassification and ROC Values. The best model is arbitrary and is needs to be aligned with the business or organization mission statement to be effective for any study. The **G_Variable Model** was chosen as the best model because it had the second best True Positive results from the Confusion Matrix and it had the least number of parameters that could be easily explained to the public when changes are made to lower non-lethal accident rates.

There are two things that I would do differently for the next Logistic Model and this is to find a dataset that has missing values to practice with the imputing node and to add more intervals to the categories to have a nice distribution of inputs variables to develop the models. Finally, the Profit/Loss Selection Criteria is another useful tool that can be used to select models and this is the other area that needs to be explored with the remaining assignments.

References

Abbot, D. (2014). Applied Predictive Analytics. Principals and Techniques for the Professional Data Analyst. John Wiley & Sons, Inc. Indianapolis, Indiana. Chapters 3-4.

Kattamuri, S. (2013). Predictive Modeling with SAS Enterprise Miner. Practical Solutions for Business Applications. SAS Institute, Inc. Cary, NC, USA. Second Edition.

King, L. 2016. Top 15 Causes of Car Accidents and How You Can Prevent Them. Huffington Post. Retrieved From: http://www.huffingtonpost.com/laiza-king-/top-15-causes-of-car-accidents_b_11722196.html

NHTSA. (2016). 2015 Motor Vehicle Crashes: Overview. US Department of Transportation. National Highway Traffic Safety Administration. Retrieved From: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318

Robie, C. (2015, Feb 15). Getting Started with SAS Enterprise Miner: **Building a regression model** and building a neural network [Video file]. Retrieved from https://www.youtube.com/watch?v=TnWRJQb5z4c

SAS books (ND). Data mining using SAS Enterprise Miner: A case study approach. SAS Institute Inc., Cary, N.C. Chapter 1. Retrieved from http://support.sas.com/documentation/cdl/en/emcs/66392/PDF/default/emcs.pdf

SAS books. (2016). Applied Analytics Using SAS Enterprise Miner Course Notes. SAS Institute, Inc. Cary, NC, USA. Chapters 1-6.

US Dept. of Transportation, Bureau of Transportation Statistics, "TranStats,". Retrieved From: www.transtats.bts.gov

Appendix

Figure 1A.

Variable Name	Role	Level	Description
ALCHL_P	Input	Binary	Alcohol involved = 1, not involved = 2
ALIGN_I	Input	Binary	1 = straight, 2 = curve
FATALITIES	Rejected	Binary	1= yes, 0= no
HOUR_I_R	Input	Binary	1=rush hour: 0=not (rush = 6-9 am, 4-7 pm
INJURY_CRASH	Rejected	Binary	1=yes, 0= no
INT_HWY	Input	Binary	Interstate Travel: 1=yes, 0= no
LGTCON_I_R	Input	Nominal	Light conditions - 1=day, 2=dark (including dawn/dusk), 3=dark, but lighted,4=dawn or dusk
MANCOL_I_R	Input	Nominal	0=no collision, 1=head-on, 2=other form of collision
MAX_SER_IR	TARGET	Binary	0=no injury, 1=non-fatal inj., 2=fatal inj. ** Please note that Fatality was removed for the study).
NO_INJ_I	Rejected	Ordinal	Number of injuries (1 to 20)
PED_ACC_R	Input	Binary	1=pedestrian/cyclist involved, 0=not
PROFIL_I_R	Input	Binary	1= level, 0=other
PRPTYDMG_CRASH	Rejected	Binary	1=property damage, 2=no property damage
RELJCT_I_R	Input	Binary	1=accident at intersection/interchange, 0=not at intersection
REL_RWY_R	Input	Binary	1=accident on roadway, 0=not on roadway
SPD_LIM	Input	Ordinal	Speed limit, miles per hour
STRATUM_R	Input	Binary	1= NASS Crashes Involving At Least One Passenger Vehicle, i.e., A Passenger Car, Sport Utility Vehicle, Pickup Truck or Van) Towed Due To Damage From The Crash Scene And No Medium Or Heavy Trucks Are Involved.
SUR_COND	Input	Nominal	Surface conditions (1=dry, 2=wet, 3=snow/slush, 4=ice, 5=sand/dirt/oil, 8=other, 9=unknown)
TRAF_CON_R	Input	Nominal	Traffic control device: 0=none, 1=signal, 2=other (sign, officer …)
TRAF_WAY	Input	Nominal	1=two-way traffic, 2=divided hwy, 3=one-way road
VEH_INVL	Input	Ordinal	Number of vehicles involved
WEATHER_R	Input	Binary	1=no adverse conditions, 2=rain, snow or other adverse condition
WKDY_I_R	Input	Binary	1=weekday, 0=weekend
WRK_ZONE	Input	Binary	1= yes, 0= no

Figure 2A: STATEXPLORE Statistics Output.

Name	Use	Report	Role	Level	Type	Format	Informat	Length	Number of Levels	Percent Missing
ALCHL_I	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
ALIGN_I	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
FATALITIES	Default	No	Rejected	Unary	Numeric	BEST12.0	BEST32.0	8	1	0
HOUR_I_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
INJURY_CRASH	Default	No	Rejected	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
INT_HWY	Default	No	Input	Nominal	Numeric	BEST12.0	BEST32.0	8	3	0
LGTCON_I_R	Default	No	Input	Nominal	Numeric	BEST12.0	BEST32.0	8	3	0
MANCOL_I_R	Default	No	Input	Nominal	Numeric	BEST12.0	BEST32.0	8	3	0
MAX_SEV_IR	Default	No	Target	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
NO_INJ_I	Default	No	Rejected	Nominal	Numeric	BEST12.0	BEST32.0	8	13	0
PED_ACC_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
PROFIL_I_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
PRPTYDMG_CRA	Default	No	Rejected	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
RELJCT_I_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
REL_RWY_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
SPD_LIM	Default	No	Input	Ordinal	Numeric	BEST12.0	BEST32.0	8	15	0
STRATUM_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
SUR_COND	Default	No	Input	Ordinal	Numeric	BEST12.0	BEST32.0	8	5	0
TRAF_CON_R	Default	No	Input	Nominal	Numeric	BEST12.0	BEST32.0	8	3	0
TRAF_WAY	Default	No	Input	Nominal	Numeric	BEST12.0	BEST32.0	8	3	0
VEH_INVL	Default	No	Input	Ordinal	Numeric	BEST12.0	BEST32.0	8	10	0
WEATHER_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
WKDY_I_R	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0
WRK_ZONE	Default	No	Input	Binary	Numeric	BEST12.0	BEST32.0	8	2	0

Figure 3A: Histogram of the Input Variables.

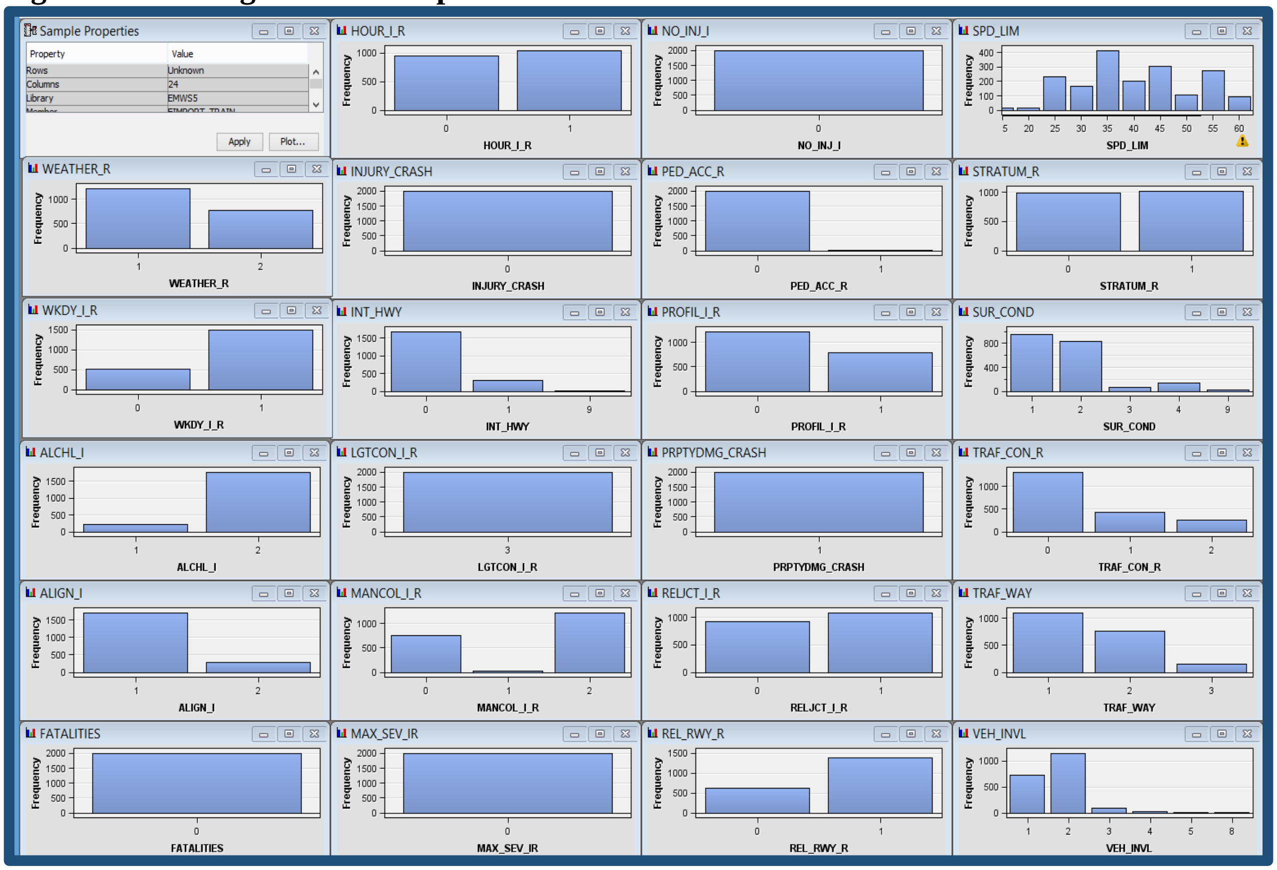


Figure 4A: Fit Statistics from the Model Selection node in SAS Enterpriser Miner.

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Misclassification Rate	Train: Average Squared Error	Selection Criterion: Valid: Misclassification Rate ▲	Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion	Train: Roc Index	Valid: Roc Index	Test: Roc Index
Y	Reg4	Reg4	D_Forward ...	MAX_SEV_IR	0.396374	0.231465	0.391888	0.391888	29862.39	30015.16	0.641	0.64	0.647
	CUT	Reg9	Cutoff Step...	MAX_SEV_IR	0.396374	0.231465	0.391888	0.391888	29862.39	30015.16	0.641	0.64	0.647
	Reg6	Reg6	H_Polynom...	MAX_SEV_IR	0.391187	0.229193	0.392943	0.392943	29704.21	30114.29	0.651	0.65	0.656
	Reg5	Reg5	E_Stepwis...	MAX_SEV_IR	0.397376	0.231619	0.39323	0.39323	29875.28	30020.01	0.64	0.639	0.646
	Reg3	Reg3	C_Backwor...	MAX_SEV_IR	0.397333	0.230804	0.396203	0.396203	29829.85	30111.28	0.645	0.643	0.65
	Reg	Reg	A_Regressi...	MAX_SEV_IR	0.395807	0.231282	0.396586	0.396586	29869.78	30127.08	0.642	0.641	0.646
	Reg2	Reg2	B_Default ...	MAX_SEV_IR	0.395981	0.230346	0.396778	0.396778	29813.15	30207.15	0.647	0.644	0.651
	Reg8	Reg8	G_Var. Sel. ...	MAX_SEV_IR	0.397507	0.232375	0.398217	0.398217	29942.48	30071.13	0.636	0.638	0.641

Figure 5A: ROC Chart Results from Model Comparison Node.

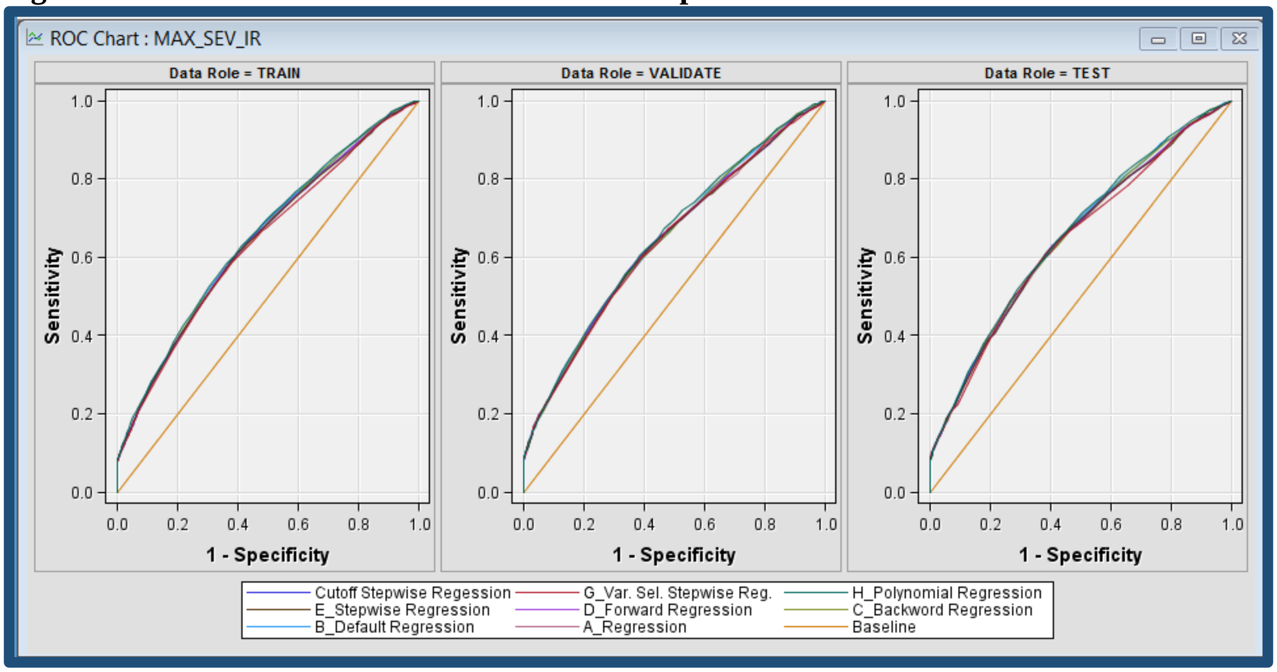


Figure 6A: Type 3 Analysis Effects for G_Variable Selection Regression.

Output

```
The selected model, based on the misclassification rate for the validation data, is the model trained in Step 7. It consists of the following effects:

Intercept  PED_ACC_R  REL_RWY_R  REP_SPD_LIM  REP_VEH_INVL  STRATUM_R  TI_MANCOL_I_R2  TI_SUR_COND1


     Likelihood Ratio Test for Global Null Hypothesis: BETA=0

     -2 Log Likelihood          Likelihood
  Intercept     Intercept &        Ratio
       Only      Covariates     Chi-Square       DF     Pr > ChiSq

  31804.758       29910.482      1894.2761       15         <.0001


            Type 3 Analysis of Effects

                                  Wald
Effect              DF    Chi-Square    Pr > ChiSq

PED_ACC_R            1      265.2783        <.0001
REL_RWY_R            1       20.6190        <.0001
REP_SPD_LIM          7       69.5311        <.0001
REP_VEH_INVL         3      245.8305        <.0001
STRATUM_R            1      257.5413        <.0001
TI_MANCOL_I_R2       1       54.8149        <.0001
TI_SUR_COND1         1       29.5016        <.0001
```

Figure 7A: Chi-Square Plot from Summary Statistics.

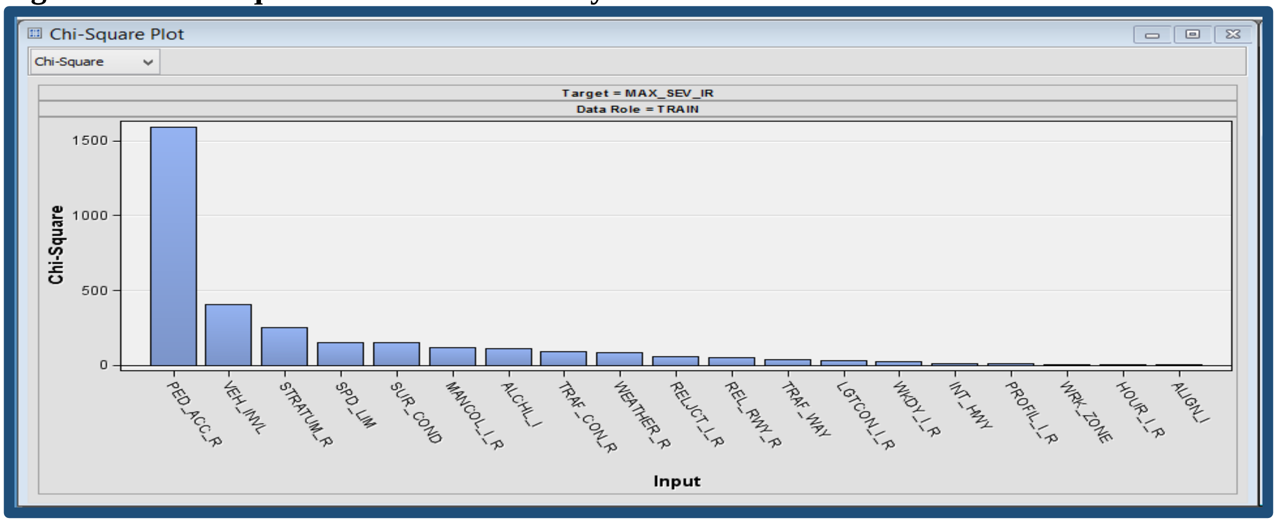


Figure 8A: Gradient Boosting Decision Tree Results.

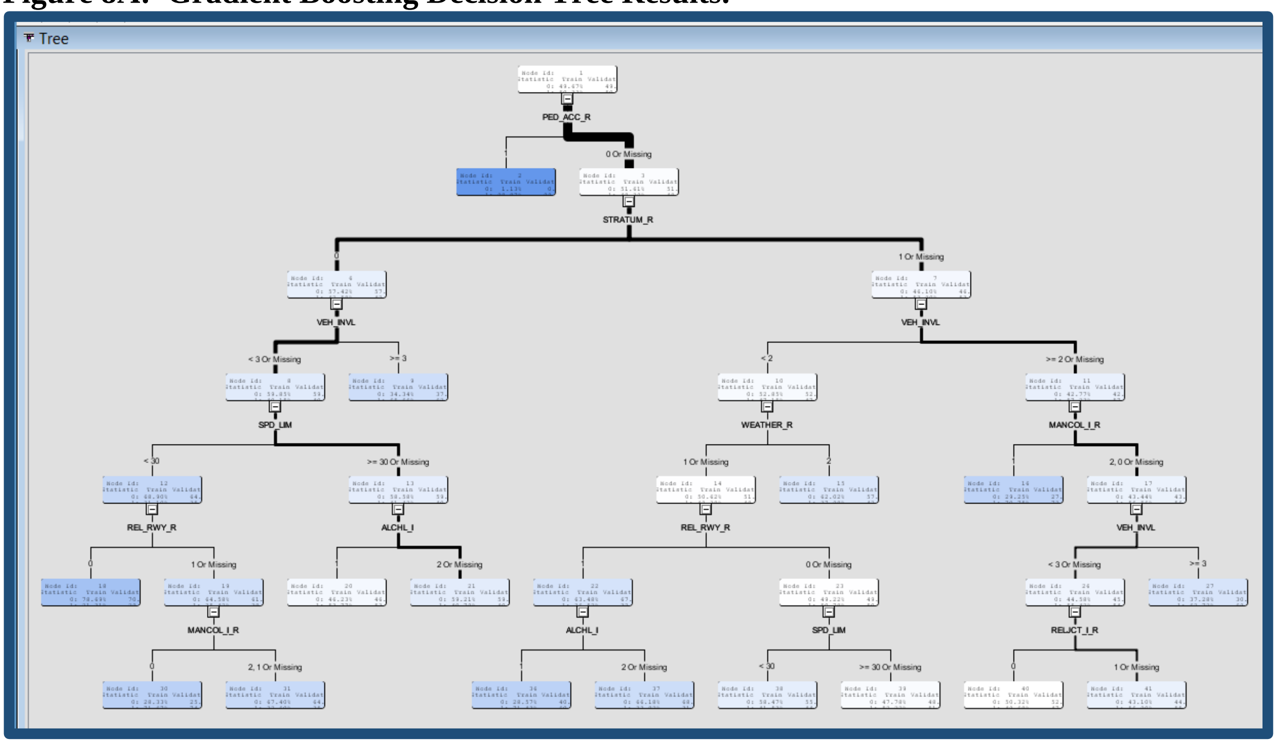
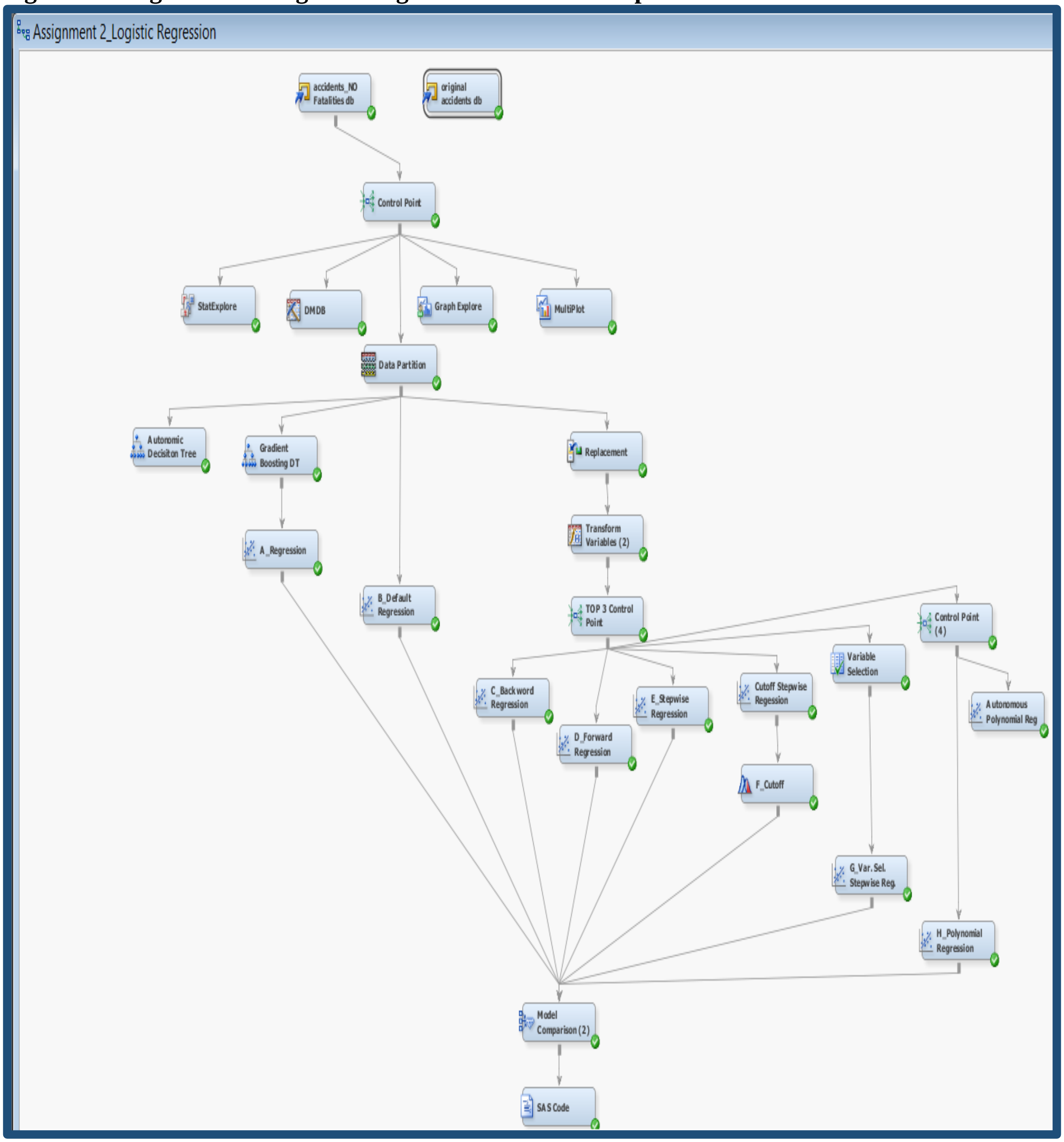


Figure 9A: Eight Model Logistic Diagram from SAS Enterprise Miner.



Figures 10A and 11A are the results from the Original Accidents Database with three levels for the Target Variable which includes Fatalities. The **PED_ACC, STRATUM and REP_VEH_INVOLV** are still the top three parameters as shown in the previous Type 3 Analysis. Please note that the Original Accidents database was hooked up to the same nodes as the non-fatal database. The G_Variable Selection Model was used to generate these figures.

Figure 10A: Score Rankings and Target Level for the Nominal Accident Database.

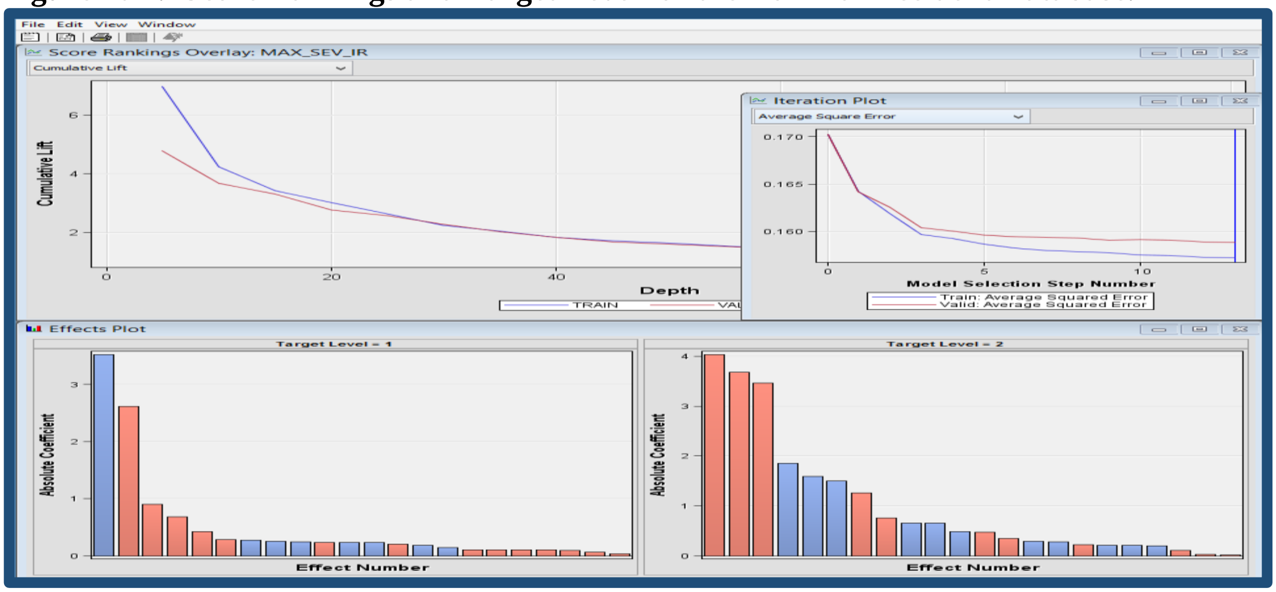
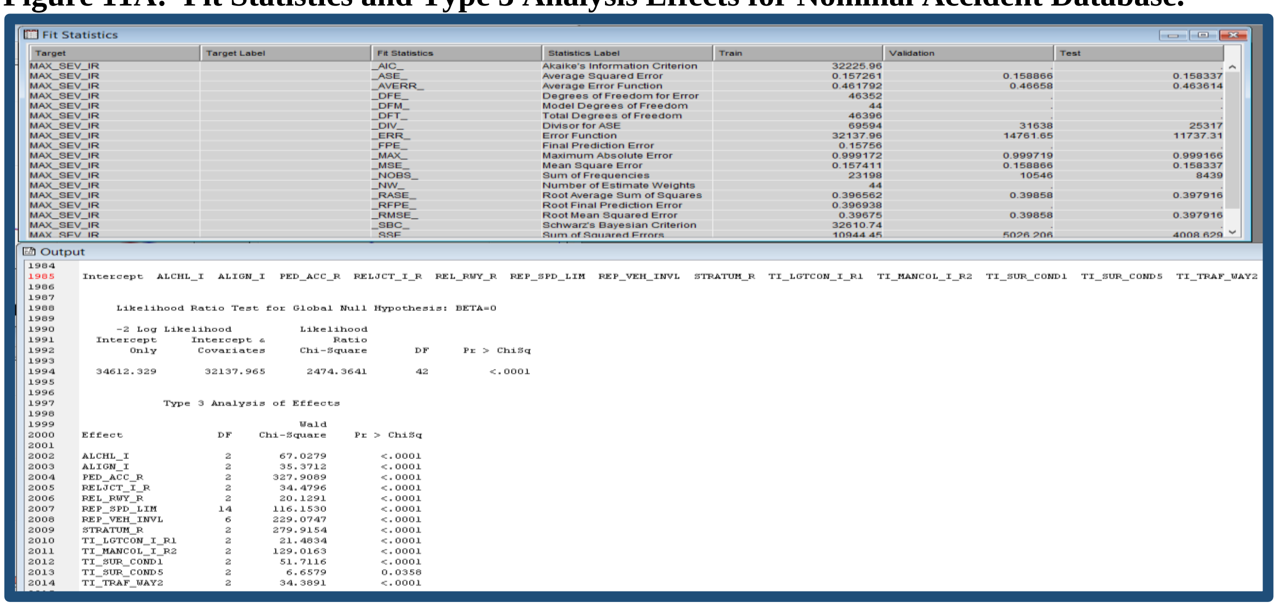


Figure 11A: Fit Statistics and Type 3 Analysis Effects for Nominal Accident Database.



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
MAX_SEV_IR		_AIC_	Akaike's Information Criterion	32225.96	.	.
MAX_SEV_IR		_ASE_	Average Squared Error	0.157261	0.158866	0.158337
MAX_SEV_IR		_AVERR_	Average Error Function	0.461792	0.46658	0.463614
MAX_SEV_IR		_DFE_	Degrees of Freedom for Error	46352	.	.
MAX_SEV_IR		_DFM_	Model Degrees of Freedom	44	.	.
MAX_SEV_IR		_DFT_	Total Degrees of Freedom	46396	.	.
MAX_SEV_IR		_DIV_	Divisor for ASE	69594	31638	25317
MAX_SEV_IR		_ERR_	Error Function	32137.96	14761.65	11737.31
MAX_SEV_IR		_FPE_	Final Prediction Error	0.15756	.	.
MAX_SEV_IR		_MAX_	Maximum Absolute Error	0.999172	0.999719	0.999166
MAX_SEV_IR		_MSE_	Mean Square Error	0.157411	0.158866	0.158337
MAX_SEV_IR		_NOBS_	Sum of Frequencies	23198	10546	8439
MAX_SEV_IR		_NW_	Number of Estimate Weights	44	.	.
MAX_SEV_IR		_RASE_	Root Average Sum of Squares	0.396562	0.39858	0.397916
MAX_SEV_IR		_RFPE_	Root Final Prediction Error	0.396938	.	.
MAX_SEV_IR		_RMSE_	Root Mean Squared Error	0.39675	0.39858	0.397916
MAX_SEV_IR		_SBC_	Schwarz's Bayesian Criterion	32610.74	.	.
MAX_SEV_IR		_SSE_	Sum of Squared Errors	10944.45	5026.206	4008.629

Output

```
Intercept  ALCHL_I  ALIGN_I  PED_ACC_R  RELJCT_I_R  REL_RWY_R  REP_SPD_LIM  REP_VEH_INVL  STRATUM_R  TI_LGTCON_I_R1  TI_MANCOL_I_R2  TI_SUR_COND1  TI_SUR_COND5  TI_TRAF_WAY2


     Likelihood Ratio Test for Global Null Hypothesis: BETA=0

     -2 Log Likelihood          Likelihood
   Intercept     Intercept &         Ratio
        Only      Covariates    Chi-Square       DF     Pr > ChiSq

   34612.329       32137.965     2474.3641       42         <.0001


              Type 3 Analysis of Effects

                                 Wald
Effect               DF    Chi-Square    Pr > ChiSq

ALCHL_I               2       67.0279        <.0001
ALIGN_I               2       35.3712        <.0001
PED_ACC_R             2      327.9089        <.0001
RELJCT_I_R            2       34.4796        <.0001
REL_RWY_R             2       20.1291        <.0001
REP_SPD_LIM          14      116.1530        <.0001
REP_VEH_INVL          6      229.0747        <.0001
STRATUM_R             2      279.9154        <.0001
TI_LGTCON_I_R1        2       21.4834        <.0001
TI_MANCOL_I_R2        2      129.0163        <.0001
TI_SUR_COND1          2       51.7116        <.0001
TI_SUR_COND5          2        6.6579        0.0358
TI_TRAF_WAY2          2       34.3891        <.0001
```