Paper 1891-2014

Using SAS® Enterprise Miner to predict the Injury Risk involved in Car Accidents

Prateek Khare, Oklahoma State University; Vandana Reddy, Oklahoma State University;
Goutam Chakraborty, Oklahoma State University

ABSTRACT

There are yearly 2.35 million road accident cases recorded in the U.S. Among them, 37,000 are considered fatal. Road crashes cost USD 230.6 billion per year, or an average of USD 820 per person. Our efforts are to identify the important factors that lead to vehicle collisions and to predict the injury risk involved in them. Data was collected from National Automotive Sampling System (NASS), containing 20,247 cases with 19 variables. Input variables describe the factors involved in an accident such as Height, Age, Weight, Gender, Vehicle model year, Speed limit, Energy absorption in Collision & Deformation location, etc. The target variable is nominal showing levels of injury. Missing values in interval variables were imputed using mean and class variables using the count method. Multivariate analysis suggests high correlation between tire footprint and wheelbase (Corr=0.97, P<0.0001) and original weight of car and curb weight of car (Corr=0.79, P<0.0001). Variables having high kurtosis values were transformed using range standardization. Variables were sorted using variable importance using decision tree analysis. Models such as multiple regression, polynomial regression, neural network, and decision tree were applied in the dataset to identify the factors that are most significant in predicting the injury risk. MLP neural network came out to be the best model to predict injury risk index, with the least Average Squared Error of 0.086 in validation dataset.

INTRODUCTION

The data is collected from National Automotive Sampling System (NASS), the following data is of Fatality Analysis of Accidents. It has attributes and factors relating to accidents.

SUMMARY OF ANALYSIS

The data is collected from National Automotive Sampling System (NASS), the following data is of Fatality Analysis of Accidents. It has attributes and factors relating to accidents.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Ids_DATA	20247
TRAIN	EMWS1.Part_TRAIN	14173
VALIDATE	EMWS1.Part_VALIDATE	6074

Table 1. Data Partition Summary

- Dataset:
  Input- 2 ID, 14 Interval, 6 Nominal variables
- Target- 1 Nominal variables

Variable Name	Long Name	Units/States	Comment
GV_CURBWGT	Vehicle Curb Weight	kg	
GV_DVLAT	Lateral Component of Delta V	km/h	
GV_DVLONG	Longitudinal Component of Delta V	km/h	
GV_ENERGY	Energy Absorption	J	
GV_FOOTPRINT	Vehicle Footprint	m^2	calculated as WHEELBAS x ORIGAVTW
GV_LANES	Number of Lanes	count	
GV_MODELYR	Vehicle Model Year	year	
GV_OTVEHWGT	Weight Of The Other Vehicle	kg	
GV_SPLIMIT	SpeedLimit	mph	converted into U.S. customary units
GV_WGTCDTR	Truck Weight Code	missing = Passenger Vehicle	
		6,000 and less	
		6,001 - 10,000	
OA_AGE	Age of Occupant	years	
OA_BAGDEPLY	Air Bag System Deployed	Nondeployed	
		Bag Deployed	
OA_HEIGHT	Height of Occupant	cm	
OA_MAIS	Maximum Known Occupant AIS	Not Injured	AIS Probability of Death
		Minor Injury	0%
		Moderate Injury	1-2%
		Serious Injury	8-10%
		Severe Injury	5-50%
		Critical Injury	5-50%
		Maximum Injury	100% (Unsurvivable)
		Unknown	Missing Value
OA_MANUSE	Manual Belt System Use	Used	
		Not Used	
OA_SEX	Occupant's Sex	Male	
		Female	
OA_WEIGHT	Occupant's Weight	kg	
VE_GAD1	Deformation Location (Highest)	Left	
		Front	
		Rear	
		Right	
VE_PDOF_TR	Clock Direction for Principal Direction of Force (Highest)	Degrees	Transformed variable, rotated 135 degrees counterclockwise

Table 2. Variable Description

Variable Summary

Role	Measurement Level	Frequency Count
ID	INTERVAL	1
ID	NOMINAL	1
INPUT	INTERVAL	14
INPUT	NOMINAL	6
TARGET	NOMINAL	1

Table 3. Variable Summary

- GV_ENERGY showed high kurtosis values of +36.97. (Variable transformation node) was used to transform the distribution to make it normal.
- Other values are observed and recorded for better understanding of dataset.
- Missing values in Interval variables were imputed by mean values using (Impute Node).

Class Variable Summary Statistics

Variable	Label	Type	Number of Levels	Missing
GV_SPLIMIT		N	16	149
GV_WGTCDTR		C	3	0
OA_MAIS		N	7	0
VE_GAD1		C	4	510

Table 4. Class Variable Summary

Interval Variable Summary Statistics

Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
GV_CURBWGT		30	13406	670.00	4250.00	1617.49	391.597	1.07286	1.5425
GV_DVLAT		4060	9376	-114.00	118.00	0.18	12.907	-0.22654	4.4263
GV_DVLONG		4060	9376	-145.00	84.00	-14.65	17.256	0.38927	3.5164
GV_ENERGY		4060	9376	4.00	9852.00	498.13	628.508	4.63584	36.9760
GV_FOOTPRINT		156	13280	2.47	7.67	4.36	0.640	1.68880	3.2613
GV_MODELYR		0	13436	2000.00	2012.00	2003.63	2.770	0.51647	-0.6115
GV_OTVEHWGT		1362	12074	680.00	4540.00	1627.81	409.744	1.09564	2.3364
OA_AGE		9	13427	0.00	97.00	40.15	17.378	0.66782	-0.3092
OA_HEIGHT		1497	11939	60.00	216.00	170.80	10.644	-0.77336	7.3553
OA_MANUSE		261	13175	0.00	1.00	0.89	0.319	-2.41991	3.8565
OA_WEIGHT		1434	12002	31.00	150.00	78.87	19.794	0.86038	0.8768
VE_ORIGAVTW		154	13282	105.00	185.00	154.77	7.611	0.68378	0.2960
VE_PDOF_TR		1278	12158	5.00	355.00	152.30	67.391	1.19424	1.2676
VE_WHEELBAS		4	13432	141.00	438.00	281.08	28.688	1.84458	4.0701

Table 5. Interval Variable Summary

- Missing values in Class variables are being imputed using count method.
- Dataset Preparation:

Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
GV_CURBWGT	MEAN	IMP_GV_CURBWGT	1617.7405552	INPUT	INTERVAL		39
GV_DVLAT	MEAN	IMP_GV_DVLAT	0.1132372805	INPUT	INTERVAL		5877
GV_DVLONG	MEAN	IMP_GV_DVLONG	-14.74583521	INPUT	INTERVAL		5877
GV_ENERGY	MEAN	IMP_GV_ENERGY	501.52221222	INPUT	INTERVAL		5877
GV_FOOTPRINT	MEAN	IMP_GV_FOOTPRINT	4.3644293941	INPUT	INTERVAL		223
GV_LANES	COUNT	IMP_GV_LANES	2	INPUT	NOMINAL		3
GV_OTVEHWGT	MEAN	IMP_GV_OTVEHWGT	1629.9529371	INPUT	INTERVAL		1992
GV_SPLIMIT	COUNT	IMP_GV_SPLIMIT	35	INPUT	NOMINAL		221
OA_AGE	MEAN	IMP_OA_AGE	40.225024754	INPUT	INTERVAL		14
OA_HEIGHT	MEAN	IMP_OA_HEIGHT	170.83519635	INPUT	INTERVAL		2116
OA_MANUSE	MEAN	IMP_OA_MANUSE	0.8844191731	INPUT	INTERVAL		385
OA_SEX	COUNT	IMP_OA_SEX	Male	INPUT	NOMINAL		201
OA_WEIGHT	MEAN	IMP_OA_WEIGHT	78.80441989	INPUT	INTERVAL		2008
VE_GAD1	COUNT	IMP_VE_GAD1	Front	INPUT	NOMINAL		771
VE_ORIGAVTW	MEAN	IMP_VE_ORIGAVTW	154.76432786	INPUT	INTERVAL		219
VE_PDOF_TR	MEAN	IMP_VE_PDOF_TR	152.20975835	INPUT	INTERVAL		1864
VE_WHEELBAS	MEAN	IMP_VE_WHEELBAS	281.1159677	INPUT	INTERVAL		8

Display 1. Imputation summary SAS® Enterprise Miner

Best Transformation R Square

Original Variable	Computed Variable	Formula	RSquare
IMP_GV_ENERGY	CNTR_IMP_GV_ENERGY	(IMP_GV_ENERGY - 501.52221222)	0.113824
IMP_GV_ENERGY	STD_IMP_GV_ENERGY	(IMP_GV_ENERGY - 501.52221222) / 531.64432069	0.113824
IMP_GV_ENERGY	IMP_GV_ENERGY		0.113824
IMP_GV_ENERGY	RANGE_IMP_GV_ENERGY	(IMP_GV_ENERGY - 4) / (9852-4)	0.113824
IMP_GV_ENERGY	SQRT_IMP_GV_ENERGY	Sqrt(IMP_GV_ENERGY + 1)	0.111754
IMP_GV_ENERGY	OPT_IMP_GV_ENERGY	Optimal Binning(4)	0.100124
IMP_GV_ENERGY	LOG_IMP_GV_ENERGY	log(IMP_GV_ENERGY + 1)	0.083517
IMP_GV_ENERGY	LG10_IMP_GV_ENERGY	log10(IMP_GV_ENERGY + 1)	0.083517
IMP_GV_ENERGY	SQR_IMP_GV_ENERGY	(IMP_GV_ENERGY + 1)**2	0.06189
IMP_GV_ENERGY	INV_IMP_GV_ENERGY	1 / (IMP_GV_ENERGY + 1)	0.023306
IMP_GV_ENERGY	EXP_IMP_GV_ENERGY	exp(IMP_GV_ENERGY / 98.52)	0.002352

Display 2. Transformation R-Square summary shows variable transformation SAS® Enterprise Miner

Variable Importance

Obs	NAME	LABEL	NRULES	IMPORTANCE
1	IMP_GV_ENERGY	Imputed GV_ENERGY	4	1.0000
2	IMP_GV_DVLONG	Imputed GV_DVLONG	6	0.8908
3	IMP_GV_DVLAT	Imputed GV_DVLAT	7	0.5799
4	IMP_OA_MANUSE	Imputed OA_MANUSE	2	0.5580
5	IMP_OA_AGE	Imputed OA_AGE	4	0.3288
6	IMP_GV_SPLIMIT	Imputed GV_SPLIMIT	2	0.2837
7	IMP_GV_FOOTPRINT	Imputed GV_FOOTPRINT	3	0.2200
8	GV_WGTCDTR		2	0.1821
9	IMP_GV_OTVEHWGT	Imputed GV_OTVEHWGT	2	0.1645
10	IMP_VE_GAD1	Imputed VE_GAD1	2	0.1645
11	IMP_VE_ORIGAVTW	Imputed VE_ORIGAVTW	1	0.1191
12	IMP_VE_WHEELBAS	Imputed VE_WHEELBAS	1	0.0735

Table 6. Variable selection using Decision tree

RESULT & ANALYSIS

- Different models are used to predict the target GV_MAIS i.e. Injury risk index.
- The dataset was partitioned as 70% Training data and 30% validation data using Data Partition node
- Imputation Node was used to impute missing data with Mean & Count
- Highly skewed data distribution was transformed using R-Square
- Variable importance node was used to identify the most important variable that affects the model.

MODELS USED TO TRAIN AND VALIDATE ARE:

- Multiple Regression
- Polynomial Regression
- Multi-layer Perception Neural Network
- Radial Equal width Neural Network
- Decision Tree

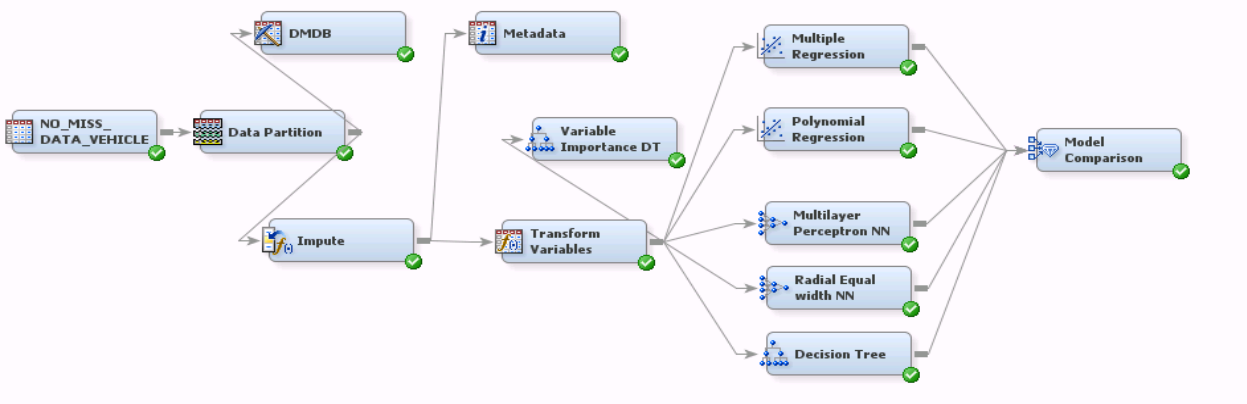


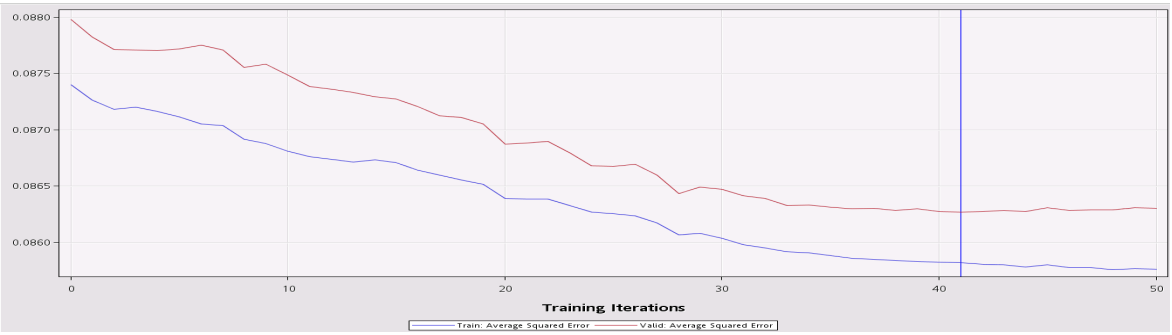
Diagram 1. Model Diagram (Nodes)

- Using Model comparison Node, Multi linear Perception Neural Network came out to be best model to predict Injury risk index, with least Average Square Error in validation dataset.

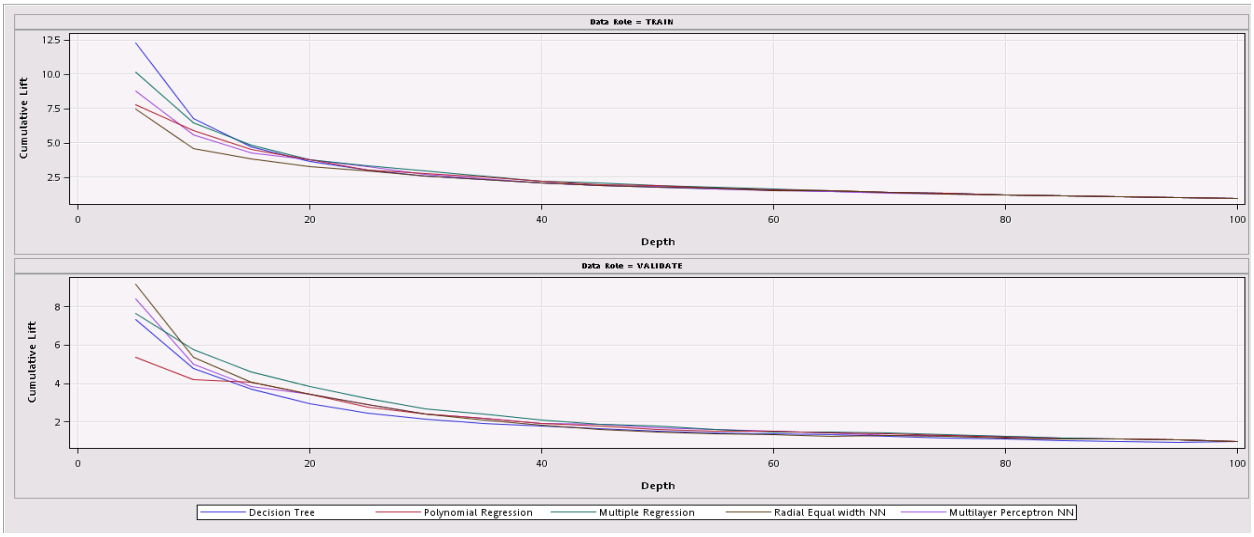
Fit Statistics

Selected Model	Model Node	Model Description	Target Variable	Selection Criterion: Valid: Average Squared Error	Train: Average Squared Error	Valid: Average Squared Error
Y	Neural	Multilayer Perceptron NN	OA_MAIS	0.086269	0.085818	0.086269
	Reg	Multiple Regression	OA_MAIS	0.087189	0.086547	0.087189
	Neural2	Radial Equal width NN	OA_MAIS	0.087366	0.08705	0.087366
	Reg2	Polynomial Regression	OA_MAIS	0.087398	0.086544	0.087398
	Tree	Decision Tree	OA_MAIS	0.087584	0.087122	0.087584

Display 3. Fit Statistics SAS® Enterprise Miner



Display 4. Average Square Error plot SAS® Enterprise Miner



Display 5. Cumulative Lift plot SAS® Enterprise Miner

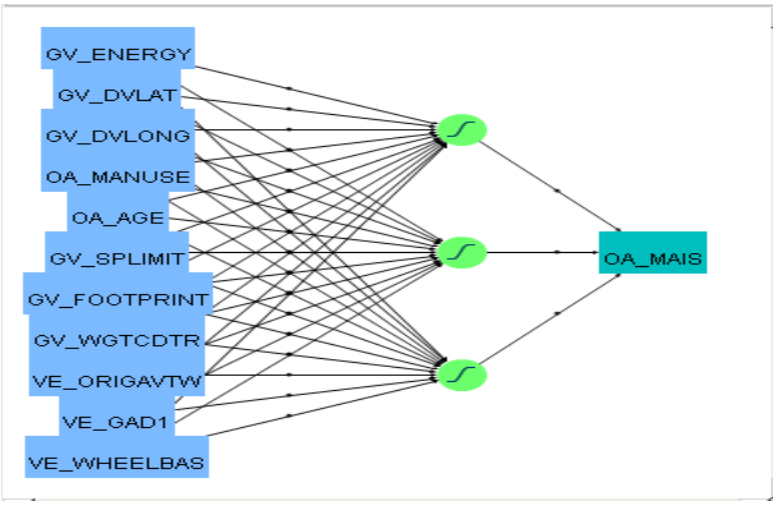


Diagram 2. Neural Network

CONCLUSION

- Various Regression models, Neural network models, Decision trees have been built and trained with numerous simulations to identify the important factors leading to car accidents in the United States
- MLP Neural Network model come out as the best model with the least average square error of 0.08626
- The factors are related to both the individual risk of the person involved in the accident and the technical and performance features of the car
- the car, lateral and longitudinal component of V Delta have been identified as the most important factors
- This analysis will help both the car manufactures & the people to prevent fatal car accidents

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Prateek Khare
Organization: Oklahoma State University
Address: 127 North Duck Street
City, State ZIP: Stillwater, OK 74075
Work Phone: 4057621514
Email: Prateek.khare@okstate.edu

Name: Vandana Reddy
Organization: Oklahoma State University
Address: 127 North Duck Street
City, State ZIP: Stillwater, OK 74075
Work Phone: 2192969394
Email: Vandana.reddy@okstate.edu

Name: Goutam Chakraborty
Organization: Oklahoma State University
Address: 420A Business Building
City, State ZIP: Stillwater, OK 74075
Work Phone: 4057447644
Email: Goutam.Chakraborty@okstate.edu