1/15/2016

Classifying Fatal Automobile Accidents in the US, 2010-2013

Using SAS Enterprise Miner to Understand and Reduce Fatalities

Team Orange

ABSTRACT

We set out to model two of the leading causes of fatal automobile accidents in America – drunk driving and speeding. We created a decision tree for each of those causes that uses situational conditions such as time of day and type of road to classify historical accidents. Using this model a law enforcement or town official can decide if a speed trap or DUI checkpoint can be the most effective in a particular situation. This proof of concept can easily be applied to specific jurisdictions for customized solutions.

INTRODUCTION

Approximately 33,000 people die in automobile accidents in the US annually (calculated from 2010-2013). The purpose of this project is to determine which law enforcement methods can be the most effective to help prevent fatal accidents and ultimately save lives. To do this we analyzed data from the Fatal Accident Reporting System (FARS), which is available on the National Highway Traffic Safety Administration website (National Highway Traffic Safety Administration, 2015). This extensive data set contains reporting information of every fatal automobile accident in the United States of America since 1975.

The objective of this study is to model two of the leading causes of fatal accidents, drunk driving and speeding, to determine which one authorities should focus on under specific circumstances. Based on variables such as time of day, day of the week, and type of road, a law enforcement official can easily determine if a fatal accident is more likely to be caused by drunk driving or speeding. With this information in mind an official can make an informed decision as to what kind of enforcement measures are appropriate – DUI checkpoints or speed traps.

We present this as a proof of concept which can be applied to specific states or jurisdictions such as counties. It is likely that this model will look different for a small county with a college or university versus a densely-populated county due to shifting population trends. Weather conditions such as snow may be more important in states which are unaccustomed to freezing conditions versus states where drivers are equipped to drive on frozen roads.

METHODS

DATA SELECTION

Seven public data sets were available for analysis. After close examination of each data source we selected the FARS data because of the opportunity it presents to impact public safety. Accidents in the FARS system “involve a motor vehicle traveling on a traffic way customarily open to the public, and must have resulted in the death of a motorist or a non-motorist within 30 days of the crash” (National Highway Traffic Safety Administration, 2015). This data is made publicly available by the government for download and exploration.

In 2010, the FARS system was merged with the National Automotive Sampling System General Estimates System (NASS GES), and the data elements were standardized across the systems. We used the data from 2010-2013 to ensure that the variables are consistent for our analysis. Additionally, we focus on drunk driving and speeding which are highly influenced by policy changes and attitudinal shifts in the population. We decided that starting with a more recent sample (still robust at 121,226 observations) would provide the most relevant model going forward. Once we selected the data, we examined and cleaned the data in Base SAS 9.4.

CLEANING

Each year in the FARS database has a separate folder with several tables to describe the accidents. Across all years, each crash is identified by a unique Consecutive Number, which is a combination of the state code and accident number. We created a primary key variable which combines the Consecutive Number of each case with the year of the crash. This primary key allows us to merge crash data across all tables and append the yearly data for each file.

Our analysis included evaluation of 121,226 crashes and eight variables: *Drunk Drivers*, *Speeding Related, Restraint System/Helmet Use, Roadway Function Class, Route Signing, Crash Time (Hour), Crash Date (Day of Week),* and *Atmospheric Conditions*. A reference table of variable names and format is in Appendix Table 1.

We binned all variables, excluding *Crash Date (Day of Week)*, to create the following nominal variables used in the analysis: *Drunk, Speeding, Restraint, Road Category, Road Type, Crash Time,* and *Weather Category*. The binned values for these variables are in Appendix Table 1.

At each step in the data cleaning process we implemented multiple steps in the code to verify the proper transformation of the raw data. For example, after creating the primary key, Unique Id, we ran a proc freq to ensure that there is only one occurrence of each case in the main Accident data set.

ANALYSIS

After cleaning the data we exported it to SAS Enterprise Miner Workstation 13.2 for analysis. We used the Data Partition node to split the data into a training data set (80%, 96,978 observations) and a validation data set (20%, 24,248 observations) via simple random sampling. The three target variables – *Drunk*, *Speeding*, and *Restraint* – are not mutually exclusive, so we built separate models to predict each. A summary of the process flow in Enterprise Miner for each model is in Figure 1.

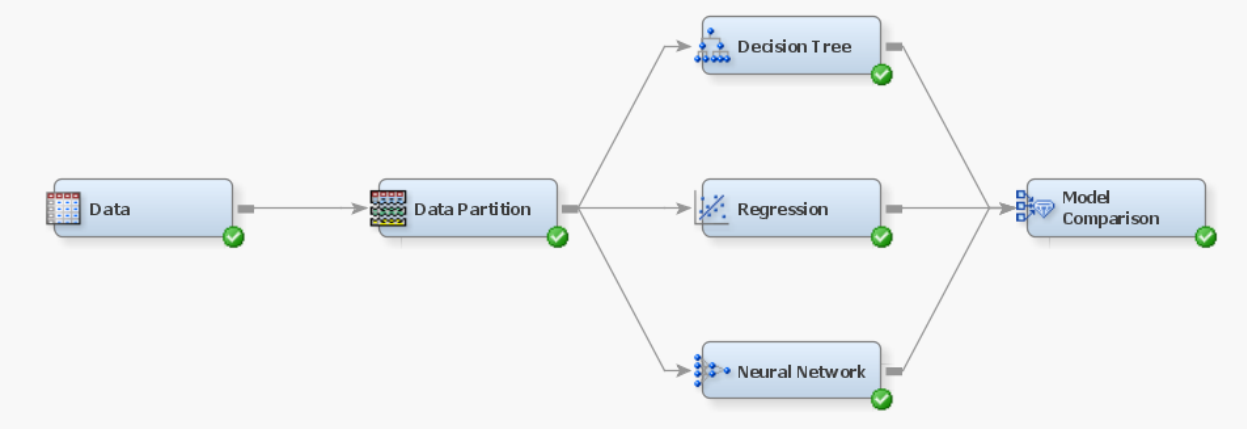


Figure 1- Enterprise Miner Diagram

We compared three models; decision trees, logistic regressions, and neural networks for each target variable. All of the input variables to the models are nominal. Each decision tree uses Chi-square probability logworth as the splitting criterion. We reduced the required splitting significance level to 0.1 (from the default of 0.2) to account for the size of the data set, and we increased the maximum number of splits from 2 to 4 for model flexibility. The default leaf size of 5 resulted in small leaves, so we increased the minimum leaf size to 1,000 observations in the training data set.

We identified the logistic regression input variables via backward selection with a required stay significance of 0.0002 (Raftery, 1995) to preserve model hierarchy and allow for interaction terms. We kept the default settings in Enterprise Miner for the neural networks. We compared the models using the validation data set misclassification rates and the area under the ROC curves (AUC), which are in Table 1.

	Decision Tree		Logistic Regression		Neural Network	
	Misc. Rate	AUC	Misc. Rate	AUC	Misc. Rate	AUC
Drunk	0.266	0.732	0.267	0.749	0.266	0.749
Speeding	0.300	0.571	0.297	0.616	0.298	0.617
Restraint	0.414	0.601	0.412	0.613	0.412	0.612

Table 1- Validation misclassification rates and AUC for all target variables and models

We are able to predict if speeding or drinking are factors in fatal accidents based on the situational inputs in our models. The misclassification rates for the three types of models were incredibly similar; the biggest difference in accuracy for any of the models is 0.003 for the speeding models. At an alpha of 0.0008 (Raftery, 1995), the differences in the AUCs are statistically different for the *Speeding* tree ($P<0.0001$), but not for the *Drunk Driving* tree ($P=0.019$). While the AUC is lowest for decision trees, we contend that the gain in interpretability of decision trees outweighs the small loss in precision. The intended users of these models are local police, town, and state officials who likely do not have training in statistics, so the interpretability of decision trees is a priority.

Given the relatively high misclassification rates for the *Restraint* models, we determined that the inputs we selected are not adequate to predict whether or not seat belts were used in fatal accidents. We concluded that seatbelt use in fatal accidents is independent of the variables in our model. The final results of our analysis are limited to the drunk driving and speeding models.

RESULTS & RECOMMENDATIONS

We created two final decision trees with identical input variables but different target variables – drunk driving and speeding. The observations used in the trees are aggregated at the national level, and the proportions and sample sizes reported in each node are for the validation data set. The sample size proportions remain relatively constant between the training and validation data. Dark shades of blue represent leaves with higher proportions of drunk drivers or speeders. In both trees (Figure 2 and Figure 3) the first split is on the binned variable *Time of Day*. The values of the bins are in Table 2. We used a categorical weather variable as an input to each model, however that variable did not end up in either tree. Our interpretation is that weather may be specific to an individual jurisdiction; snow may not have the same impact on fatal accidents in the North as in the South. Larger versions of the *Drunk* and *Speeding* trees are in Appendix Figure 1 and Appendix Figure 2, respectively.

Bin Value	Bin Label
10 PM - 4 AM	Bar Time
4 AM - 10 AM	Morning Commute
10 AM - 4 PM	Mid-day Drive
4 PM - 10 PM	Evening Commute

Table 2- Time of Day variable binned values

The decision tree for drunk driving, in Figure 2, demonstrates that fatal accidents involving drunk drivers occur throughout the day, but they are least common during the Mid-day Drive hours. During Bar Time (10 PM - 4 AM) the rate of drunk fatal accidents is nearly twice the average observed in the overall sample, at 58%. Once we follow subsequent splits from the Bar Time node, we see that Saturdays and Sundays during Bar Time has one of the highest rates of drunk related fatal accidents, with a lift of 2 for that node. This is not surprising given that most bars close after midnight, so bar patrons who drive home on Friday and Saturday nights after midnight are technically driving on Saturdays and Sundays, respectively. That split aligns with our preconceptions about drunk driving; it mostly happens on the weekend. However, our tree uncovers a pattern in the data that is not quite as obvious. The highest rate of fatal accidents related to drinking and driving is from 10 PM to 4 AM on Mondays, Tuesdays, and Wednesdays on County roads. The detection of this hidden pattern is one of the most powerful results a data mining technique, such as a decision tree, can have.

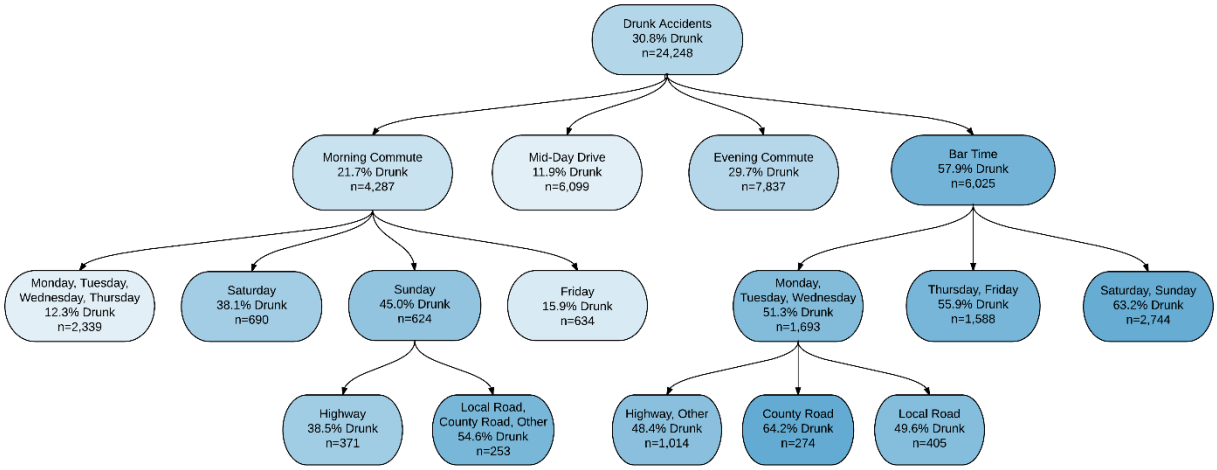


Figure 2- Decision tree with Drunk Driving as target variable

The first split in the *Speeding* decision tree, Figure 3, indicates that fatal accidents during Bar Time are the most likely to involve speeding versus all other times. It is interesting to note that fatalities occurring on highways are least likely to have speeding listed as a contributing factor, while those on Local or County roads are more likely to involve speeding. Local roads in Rural areas are particularly prone to speeding-related fatalities, with speed contributing to a majority of fatal accidents (53%). Surprisingly, the type of road is not significant in classifying whether or not fatal accidents involve speeding for any other times of day (outside of 10 PM - 4 AM).

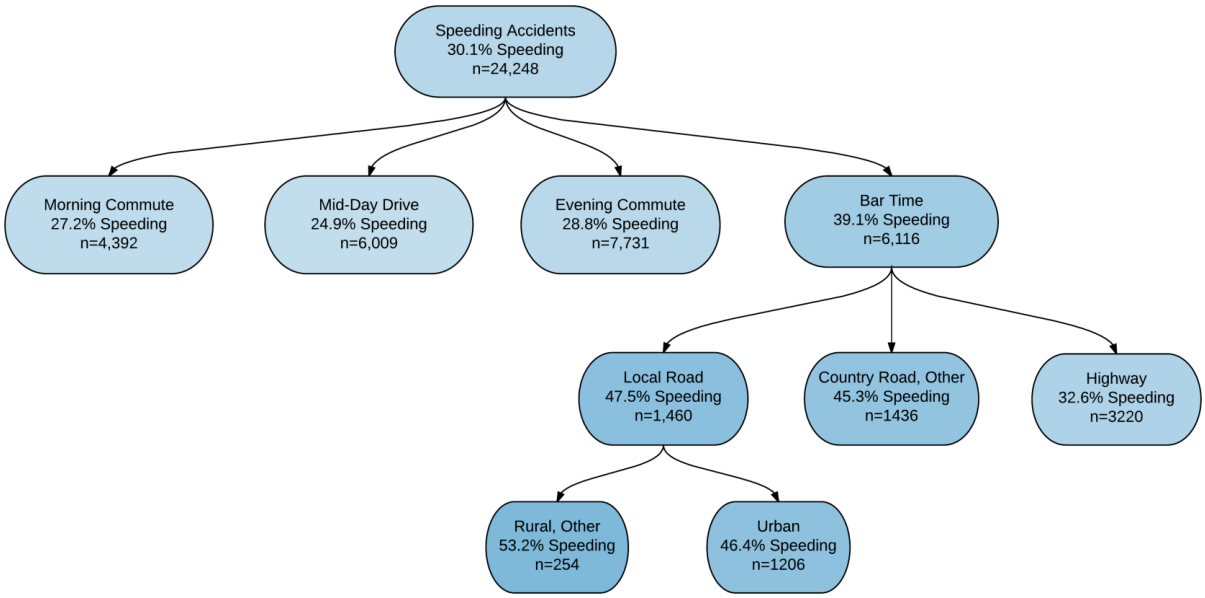


Figure 3- Decision tree with Speeding as target variable

Both the *Speeding* and *Drunk Driving* decision trees use the same inputs, and as such they are easy to use concurrently. By inputting day of week, time of day, road category and road type, probabilities can be obtained from both models for whether fatal car accidents are likely to involve speeding or drunk driving. For instance, a local police department considering a location for a DUI checkpoint to prevent fatal accidents during the workweek would be advised to use a County Road between 10 PM and 4 AM. However, during the weekend the type of road is insignificant; therefore, the checkpoint could be placed on any road. A police department trying to prevent fatalities involving speeding would know, using the same inputs, that speeding is most likely to contribute to fatal accidents on Local and County Roads during Bar Time.

CONCLUSION

We created two decision trees to model how situational conditions relate to drunk driving and speeding, two of the main causes of automobile accidents which result in fatalities. In this proof of concept we successfully demonstrated how a law enforcement official may capitalize on data-driven decision making to maximize the efficiency of their efforts and minimize preventable fatalities.

The decision trees are highly interpretable, making them an excellent choice for a non-technical user. Future studies may be location specific to appropriately address the unique conditions in each state, or even county. We expected the weather variable to stay in the model, however its absence from the model allows for longer-term planning which does not require advance knowledge of exact weather conditions. We are developing an application which enables the end user to select values of the input variables (i.e. time, day, etc.), and receive an automatic output indicating the likelihood that a fatal accident will involve drinking or speeding. Our goal is to make these models easy to use so that they may be deployed in as many jurisdictions as possible.

REFERENCES

National Highway Traffic Safety Administration. (2015). Fatal Analysis Reporting System: Detailing the Factors Behind Traffic Fatalities on our Roads.

Raftery, A. E. (1995). Bayesian Model Selection in Socia Research. *Sociological Methodology, 25*, 111-163.

APPENDIX

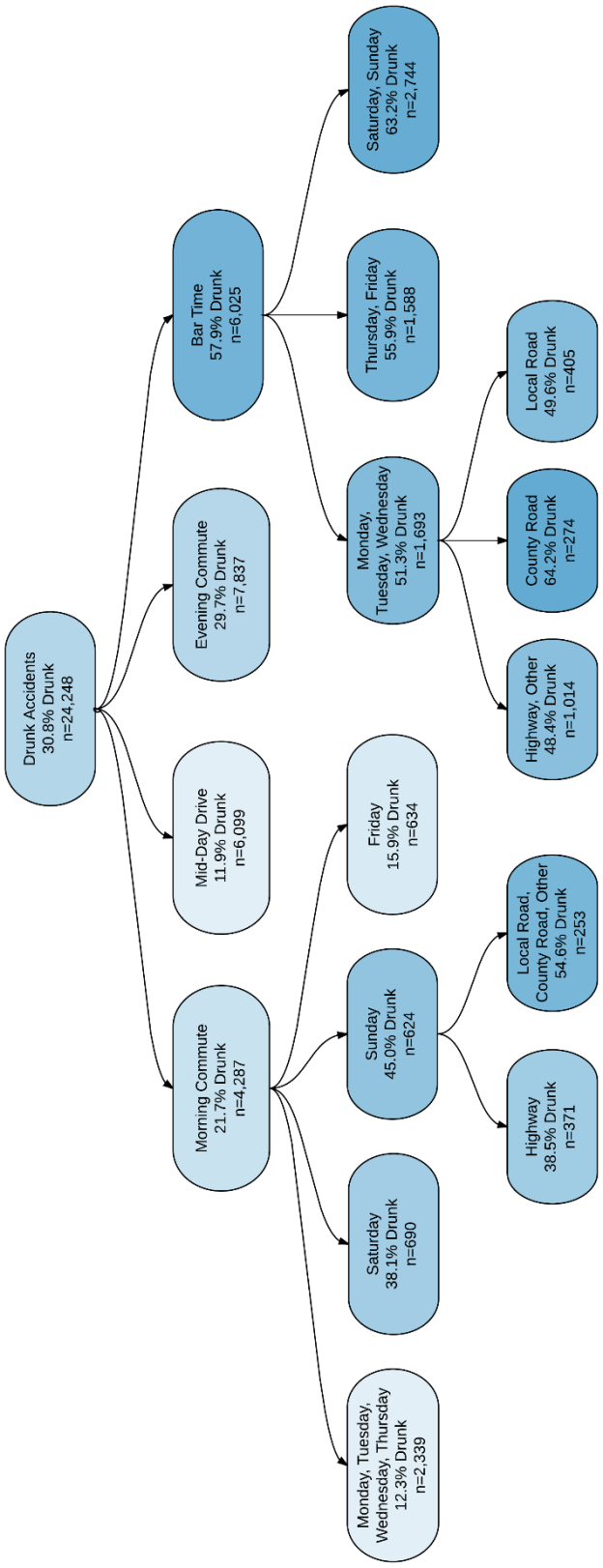
Original Variable	Binned Variable	Role	Type	Level
Drunk Drivers	**Drunk**	**Target**	**Binary**	
0	No Drunk Drivers Involved	Target	Binary	0
>=1	At least One Drunk Driver Involved	Target	Binary	1
Restraint System/Helmet Use	**Restraint**	**Target**	**Binary**	
Not Applicable	No Restraint/Unknown	Target	Binary	0
Child Restraint Type Unknown	No Restraint/Unknown	Target	Binary	0
None Used-Motor Vehicle Occupant	No Restraint/Unknown	Target	Binary	0
Other Helmet	No Restraint/Unknown	Target	Binary	0
No Helmet	No Restraint/Unknown	Target	Binary	0
Other	No Restraint/Unknown	Target	Binary	0
Not Reported	No Restraint/Unknown	Target	Binary	0
Unknown	No Restraint/Unknown	Target	Binary	0
Helmet, Unknown if DOT Compliant	No Restraint/Unknown	Target	Binary	0
Unknown if Helmet Worn	No Restraint/Unknown	Target	Binary	0
Shoulder Belt Only Used	Restraint Used	Target	Binary	1
Lap Belt Only Used	Restraint Used	Target	Binary	1
Shoulder and Lap Belt Used	Restraint Used	Target	Binary	1
DOT-Compliant Motorcycle Helmet	Restraint Used	Target	Binary	1
Restraint Used-Type Unknown	Restraint Used	Target	Binary	1
Child Restraint System-Forward Facing	Restraint Used	Target	Binary	1
Child Restraint System-Rear Facing	Restraint Used	Target	Binary	1
Booster Seat	Restraint Used	Target	Binary	1
Speeding Related	**Speeding**	**Target**	**Binary**	
No	No Speeding or Unknown	Target	Binary	0
No Driver Present/Unknown	No Speeding or Unknown	Target	Binary	0
Unknown	No Speeding or Unknown	Target	Binary	0
Yes, Racing	Yes at Least One Driver Speeding	Target	Binary	1
Yes, Exceeding Speed Limit	Yes at Least One Driver Speeding	Target	Binary	1

Yes, Too Fast for Conditions	Yes at Least One Driver Speeding	Target	Binary	1
Yes, Specifics Unknown	Yes at Least One Driver Speeding	Target	Binary	1
Atmospheric Conditions	**WeatherCat**	**Input**	**Nominal**	
Fog, Smog, Smoke	Air Quality	Input	Nominal	1
Severe Crosswinds	Air Quality	Input	Nominal	1
Blowing Sand, Soil, Dirt	Air Quality	Input	Nominal	1
Clear	Clear	Input	Nominal	2
Cloudy	Cloudy	Input	Nominal	3
Blowing Snow	Freezing	Input	Nominal	4
Freezing Rain or Drizzle	Freezing	Input	Nominal	4
Sleet or Hail	Freezing	Input	Nominal	4
Snow	Freezing	Input	Nominal	4
No Additional Atmospheric Conditions	Other	Input	Nominal	5
Other	Other	Input	Nominal	5
Not Reported	Other	Input	Nominal	5
Unknown	Other	Input	Nominal	5
Rain	Rainy	Input	Nominal	6
Crash Time(Hour)	**Hour**	**Input**	**Nominal**	
10pm-4am	Bar Time	Input	Nominal	1
4pm-10pm	Evening Commute	Input	Nominal	2
4am-10am	Morning Commute	Input	Nominal	3
10am-4pm	Mid-day Drive	Input	Nominal	4
Roadway Function Class	**RoadCat**	**Input**	**Nominal**	
Unknown	Other	Input	Nominal	1
Rural-Principal Arterial-Interstate	Rural	Input	Nominal	2
Rural-Principal Arterial-Other	Rural	Input	Nominal	2
Rural-Minor Arterial	Rural	Input	Nominal	2
Rural-Major Collector	Rural	Input	Nominal	2
Rural-Minor Collector	Rural	Input	Nominal	2
Rural-Local Road or Street	Rural	Input	Nominal	2
Rural-Unknown Rural	Rural	Input	Nominal	2
Urban-Principal Arterial-Interstate	Urban	Input	Nominal	3
Urban-Principal Arterial-Freeways	Urban	Input	Nominal	3
Urban-Other Principal Arterial	Urban	Input	Nominal	3
Urban-Minor Arterial	Urban	Input	Nominal	3
Urban-Collector	Urban	Input	Nominal	3
Urban-Local Road or Street	Urban	Input	Nominal	3
Urban-Unknown Urban	Urban	Input	Nominal	3
Route Signing	**RoadType**	**Input**	**Nominal**	
County Road	County Road	Input	Nominal	1

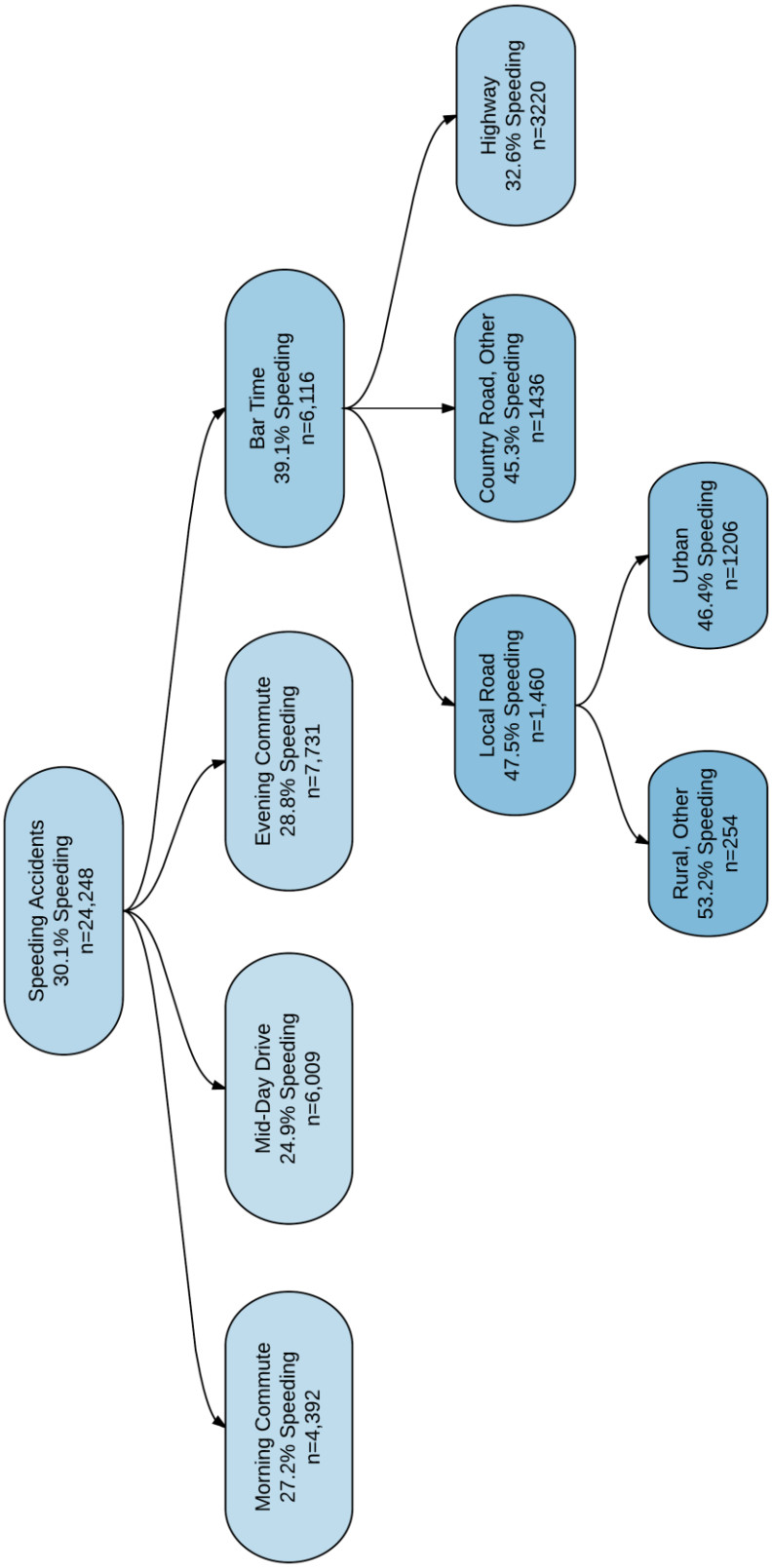
Interstate	Highway	Input	Nominal	2
U.S. Highway	Highway	Input	Nominal	2
State Highway	Highway	Input	Nominal	2
Local Street-Township	Local	Input	Nominal	3
Local Street-Municipality	Local	Input	Nominal	3
Local Street-Frontage Road	Local	Input	Nominal	3
Other	Other	Input	Nominal	4
Unknown	Other	Input	Nominal	4

Appendix Table 1- Variable Reference Table

Variables in green represent targets, variables in gold represent inputs.



Appendix Figure 1- Decision tree for drunk driving (large format)



Appendix Figure 2- Decision tree for speeding (large format)