MILLIMAN WHITE PAPER

Predictive analytics uncovers most influential factors behind car accidents

William Torres
Tim Vosicky
Jill Rosenblum

We all know wearing a seatbelt and obeying speed limits decrease our chances of being injured in a car accident, but the question is how much?

The National Highway Traffic Safety Administration (NHTSA) has collected crash data through its Crash Report Sampling System (CRSS) since the early 1970s. It draws from a nationally representative sample selected from the estimated 5 million to 6 million police-reported annual crashes, involving all types of motor vehicles, pedestrians, and cyclists, ranging from property damage only to crashes that involve fatalities.

Our goal is to use state-of-the-art machine learning algorithms and the information gathered by the CRSS to uncover patterns behind the data—to determine the most influential factors that lead to severe injuries and fatalities.

Using a predictive analytic approach

Using Milliman's predictive modeling tool, SOLYS, and the data set provided by the NHTSA, we have created a model to determine which factors have the strongest associations with accidents that involve an injury and the number of individuals injured in an accident.

SOLYS allows us to manipulate large amounts of data and build predictive models to uncover patterns within the data. Models can be built using innovative algorithms (Random Forests, Gradient Boosting Trees, etc.) that are widely used in Kaggle competitions. They allow us to predict outcomes such as whether an individual involved in an accident is likely to be injured or not, given a specific set of circumstances, and furthermore the potential severity of such injury. We can also determine how these factors or circumstances impacted predictions.

Selecting variables to be used from the data set is a key step of the process. Some variables can be considered to be related to the outcome itself rather than being factors or characteristics of the observation. Examples of this type of variable are airbag deployment, passenger ejection from the vehicle, or final damage sustained by the vehicle. Notice that these variables would increase the accuracy of any model because they are highly correlated to severe accidents. Rather, the goal should be to focus on factors that represent potential risk characteristics *prior* to the accident.

SOLYS features two modalities, the user interface (UI) and the notebook. The UI provides a user-friendly environment that allows users with limited data science background to use a "point and click" approach to predictive modeling, while maintaining most of the options that R and Python offer. For a more "expert mode" version, SOLYS offers an environment built in the Jupyter notebook that allows for creation and sharing of documents that contain live code, equations, visualization, and narrative text, accessible from a web browser.

Models

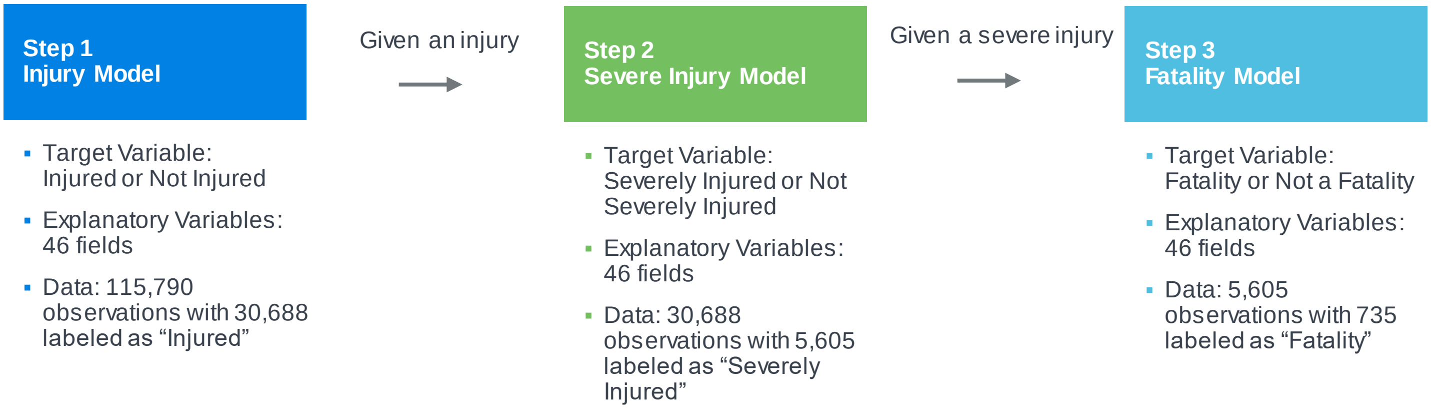
Tree-based methods such as Gradient Boosting Trees have proven to be very powerful techniques in building supervised models. They have become the de facto choice of tree ensemble models, used widely in Kaggle competitions. Implementations such as Random Forests and XGBoost are the leading choices when building models to predict a binary outcome given a structured data set.

For this particular problem, a multistep model ensemble was used, following a conditional probability scheme. For Step 1, the model predicts the probability of an individual involved being injured or not, as a binary outcome. Given that the individual involved is injured, Step 2 predicts the probability of a severe injury as opposed to a non-severe injury, again a binomial response. Finally, for Step 3, given that the individual has had a severe injury, the model now looks at the probability of the injury being fatal. This process effectively converts a four-level multinomial problem into a three-step binomial model.

An alternative approach would be to build a multiclass classification model, but one can argue that a multistep (rather than a multiclass) model ensemble is advantageous for the following reasons:

- Models will focus on the response with a data set already split. This “forced” initial split will help guide the algorithms to pick variables most influential in each scenario. For example: car model age does not play a key role when predicting whether a driver will experience an injury, but given that there is an injury it becomes significant when determining the severity, as newer vehicles tend to have more advanced safety features.
- The approach allows for different interpretations of variables. For example: the use of restraint equipment is paramount for every individual involved in an accident, but this could mean wearing a helmet for a motorcyclist versus wearing a seatbelt for a vehicle passenger.
- Parameters can be tuned to create a better fit in each scenario.
- Results can be isolated, allowing us to measure the influence of the variables in each outcome.
- A multistep approach complies with the goal of predicting a hierarchical response (No Injury > Injury > Severe Injury > Fatality).

FIGURE 1: 3-STEP MODEL APPROACH



Step 1
Injury Model

- Target Variable: Injured or Not Injured
- Explanatory Variables: 46 fields
- Data: 115,790 observations with 30,688 labeled as “Injured”

Given an injury →

Step 2
Severe Injury Model

- Target Variable: Severely Injured or Not Severely Injured
- Explanatory Variables: 46 fields
- Data: 30,688 observations with 5,605 labeled as “Severely Injured”

Given a severe injury →

Step 3
Fatality Model

- Target Variable: Fatality or Not a Fatality
- Explanatory Variables: 46 fields
- Data: 5,605 observations with 735 labeled as “Fatality”

Using the results of each model, a final prediction can be estimated following a logic similar to a decision tree, as illustrated in Figure 6 below.

Motorist Injury model

This model predicts the likelihood of injury, regardless of how severe, for an individual inside a vehicle at the time of the accident.

The results in this model indicate that front-to-front collisions represent the most dangerous situations leading to injuries in accidents. A single variable, “non-collision harmful events,” was created, composed of various individual event types such as rollovers, explosions, pavement surface irregularities, etc. The presence of these event types in the injury model, but their absence from the severe and fatality models, suggests that these events lead to injuries but play no role in its severity.

Another interesting result was motorist age. Although it is intuitive that younger individuals tend to be more resilient to injuries, it is surprising that all models give a strong signal to this variable with a direct proportional relationship (i.e., as an individual gets older, they are more likely to suffer an injury). Another important fact about motorist age is that this variable has little to no correlation to other variables, which increases its value as a result.

FIGURE 2: INJURY MODEL (STEP 1): TOP INFLUENTIAL VARIABLES

Variable	Variable Gain (Global)	Mean Log-odd Contribution (Local)
Manner of Collision	**11%**	
Not collision with MV in transport		-0.073
Front-to-Rear		0.022
Front-to-Front		0.510
Angle		0.243
Sideswipe, Same Direction		-0.740
Sideswipe, Opposite Direction		-0.560
Rear-to-Side		-0.629
Rear-to-Rear		-0.678
Other		-0.472
Non-Collision Harmful Events	**9%**	
0 events		-0.057
> 1		1.271
Crash Type	**6%**	
Changing Traffic		0.056
Intersecting Paths		0.119
Other		-0.147
Same Traffic, Same Direction		-0.049
Same Traffic, Opposite Direction		0.376
Single Driver		-0.034
Motorist Age (years)	**6%**	
0-9		-0.354
10-19		-0.137
20-29		-0.090
30-39		-0.010
40-49		0.100
50-59		0.124
60-69		0.157
+70		0.240
Restraint (seatbelt)	**5%**	
No seatbelt used		1.187
Seatbelt used		-0.044

Motorist Severe Injury model

The Motorist Severe Injury model predicts the likelihood of a severe injury for an individual inside a vehicle, given that the individual has sustained an injury at the time of the accident.

Possibly the most interesting result in this step is the introduction of vehicle model age (at the date of the accident), which suggests that newer vehicles do not necessarily prevent injuries, but are better at reducing their severity. Traveling speed is also introduced, with a proportional relationship that seems intuitive. The faster a vehicle is traveling, the more likely it is that an injury will be severe in the case of an accident.

FIGURE 3: SEVERE INJURY MODEL (STEP 2): TOP INFLUENTIAL VARIABLES

Variable	Variable Gain (Global)	Mean Log-odd Contribution (Local)
Crash Type	**17%**	
Changing Traffic		-0.123
Intersecting Paths		-0.104
Other		-0.212
Same Traffic, Same Direction		-0.319
Same Traffic, Opposite Direction		0.602
Single Driver		0.438
Restraint (seatbelt)	**13%**	
No seatbelt used		1.106
Seatbelt used		-0.133
Motorist Age (years)	**8%**	
0-9		-0.446
10-19		-0.295
20-29		-0.113
30-39		-0.033
40-49		0.069
50-59		0.109
60-69		0.150
+70		0.450
Vehicle Model Age (years)	**6%**	
0-4		-0.221
5-9		0.018
10-14		0.116
15-19		0.192
+20		0.339
Traveling Speed (mph)	**5%**	
0-9		-0.230
10-19		-0.026
20-29		-0.043
30-39		-0.003
40-49		0.212
50-59		0.276
60-69		0.392
+70		0.419

Motorist Fatality model

Finally, the Motorist Fatality model predicts the likelihood of a fatality for an individual inside a vehicle, given that the individual has sustained a severe injury at the time of the accident.

The use of seatbelts and motorist age now play key roles in this model, a similar result from the previous steps, but now marking a difference between life and death.

An unexpected result from this step is the nonlinearity of traveling speed (see Figure 4). This is because many fatalities result from accidents where the vehicle was impacted from the side. Typically, vehicles impacted from the side were traveling with a speed less than 20 mph. Interestingly, side airbags are a security feature that was mandated in 1998.

FIGURE 4: FATALITY MODEL (STEP 3): TOP INFLUENTIAL VARIABLES

Variable	Variable Gain (Global)	Mean Log-odd Contribution (Local)
Restraint (seatbelt)	**14%**	
No seatbelt used		0.637
Seatbelt used		-0.257
Motorist Age (years)	**13%**	
0-9		-0.267
10-19		-0.269
20-29		-0.232
30-39		-0.204
40-49		-0.111
50-59		0.124
60-69		0.195
+70		0.700
Crash Type	**9%**	
Changing Traffic		-0.338
Intersecting Paths		-0.200
Other		-0.218
Same Traffic, Same Direction		-0.285
Same Traffic, Opposite Directior		0.337
Single Driver		0.171
Traveling Speed (mph)	**6%**	
0-9		0.016
10-19		-0.004
20-29		-0.190
30-39		-0.244
40-49		-0.214
50-59		-0.186
60-69		0.087
+70		0.359
Vehicle Model Age (Years)	**6%**	
0-4		-0.100
5-9		-0.012
10-14		0.008
15-19		0.043
+20		0.033

Modeling outcomes

The overall results give us a good accuracy at the personal level, with 71.2% accuracy when predicting the degree of injury (the sum of the bolded values in the table in Figure 5). The model overestimates 13.7% of the injuries and underestimates 15.1% of the observations (see Figure 5).

FIGURE 5: FINAL RESULTS (INDIVIDUAL LEVEL)

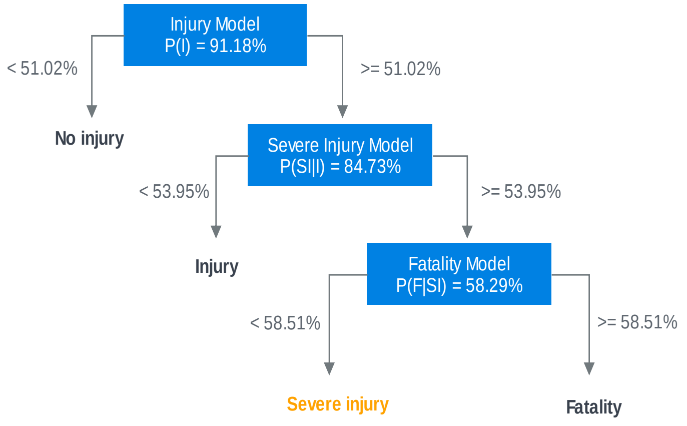
	Actual Values				
Prediction	Not Injured	Minor Injury	Severe Injury	Fatality	Total
Not Injured	56.8%	10.3%	1.0%	0.1%	68.1%
Minor Injury	9.6%	11.9%	1.9%	0.2%	23.6%
Severe Injury	1.8%	2.6%	2.1%	0.3%	6.7%
Fatality	0.2%	0.6%	0.4%	0.5%	1.6%
Total	68.3%	25.3%	5.4%	2.0%	100.0%

Predictions and variable importance

In making a prediction, understanding the influence of each variable is fundamental. There are different ways to approach a solution for this problem, although multiple packages provide explainers for tree ensemble models and their focus is to break down the log-odd contribution that each variable had on the prediction.

For a particular observation in the data, we have a 46-year-old single driver who departed from the roadway and crashed with a culvert. Figure 6 shows the logic followed after the results from the models, and a threshold value is selected in each step to determine the final outcome. Notice that in Figure 6 the threshold values from the first two steps are surpassed, but not for the third step, making the final prediction “Severe Injury” instead of “Fatality.”

FIGURE 6: TURNING RESULTS INTO PREDICTIONS



In order to measure the local importance on each prediction, we estimate the logarithmic contribution each variable has to the final odds. Each variable contribution to the final result will depend on the particular value in the observation. Following the same example as before, Figure 7 shows the contribution from the most influential variables as well as the local value for the severe injury model.

The final prediction for this observation matches the actual reported injury ("Severe Injury"). According to the model it has been determined that its being a single car accident, along with the area of impact and the traveling speed, were the most aggravating factors that contributed to the prediction of its being a severe injury rather than just an injury, whereas using a seatbelt was a mitigating factor.

FIGURE 7: LOCAL VARIABLE INFLUENCE ON A SINGLE PREDICTION FROM SEVERE INJURY MODEL

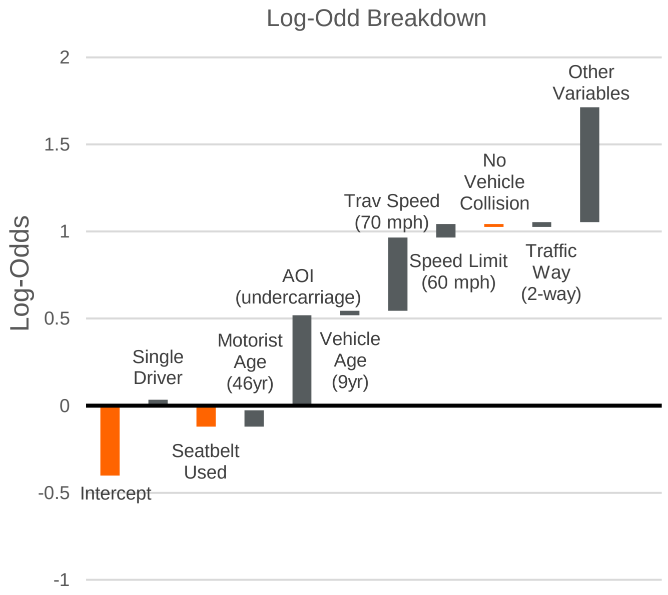


FIGURE 8: RESULTS FOR SINGLE PREDICTION FROM SEVERE INJURY MODEL

Variable	Value	Local Importance (Log-Odd)
Accident Type	Single Driver	+0.435
Restraint Used	Seatbelt	-0.154
Age	46 yr	+0.093
Area of Impact	Undercarriage	+0.546
Vehicle Age	9 yr	+0.026
Travelling Speed	70 mph	+0.420
Speed Limit	60 mph	+0.078
Manner of Collision	No Vehicle Collision	-0.016
Traffic Way	Two-way	+0.027
Other Variables		+0.660
Intercept		-0.401
	Total Log-Odd	1.714
Probability of Severe Injury		84.73%

Conclusion

Despite their prediction power, machine learning techniques are often regarded as black boxes due to the complexity of the algorithms. This “loss” of interpretability is perhaps one of the limitations that make its usage and implementation difficult in certain industries, such as the insurance and financial industries.

By breaking down the prediction into contributions from each variable included in the model, we are capable of not only predicting with accuracy, but also gaining an additional insight that allows for understanding the causes and taking actions that could help prevent unwanted outcomes.

Milliman

Milliman is among the world’s largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

William Torres
william.torres@milliman.com

Tim Vosicky
tim.vosicky@milliman.com

Jill Rosenblum
jill.rosenblum@milliman.com

© 2019 Milliman, Inc. All Rights Reserved. The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.