MILLIMAN WHITE PAPER

Driving midday in January wearing your seatbelt

Liz Young
Kayla Spitzner

In 2018, we used the new SOLYS predictive modeling software to determine which factors have the strongest association with (a) accidents that involve an injury, and (b) the number of persons injured in an accident within the 2016 Crash Report Sampling System (CRSS) data of the National Highway Traffic Safety Administration. These objectives are associated with specific variables from the CRSS database, one representing the imputed maximum severity resulting from a crash, and another the imputed number of persons injured resulting from a crash, respectively. We filtered the data for all targets such that the CRSS variable representing maximum severity imputed in a crash indicated that an injury or a fatality had occurred. Our team in Denver designed and carried out an approach to answer these questions utilizing the SOLYS tool.

DATA PREPARATION

The CRSS database has a plethora of variables available for analysis. Many of these variables are categorical, with individual values often representing similar or comparable features relevant to the crash in question. An important part of our analysis was grouping and one-hot-encoding categorical variables. To group, we selected categorical features that were similar and considered them together. After grouping, one-hot-encoding entailed creating a binary variable for each feature, and removing the original variable. This process improves fits for some predictive models, and helps make clear what specific aspects of the variable were relevant when evaluating the importance of variables in our final fits. For instance, we are easily able to point out that whether a vehicle was rear-ended is important when predicting the variable Maximum Severity. Without this encoding, we might only be able to say more vaguely that point of impact was important. Grouping and one-hot-encoding enabled us to draw more explicit connections between features and crash outcomes.

We split the data into training and testing subsets, to facilitate validation of our models. We saved the complete data set for use in training our final selected models.

FITTING MODELS

We used our training data set to generate predictions from the models we trained. We then used the testing data set to validate these models. After selecting our final parameters and models, we refit to the complete data set. We used the feature importance values from these final models to inform our weighted importance tables. Model-fitting functionality is explored further in the "Working in SOLYS" section of this paper.

FEATURE IMPORTANCE

To come to our final importance rankings, we devised a method of combining the feature importance assigned by each model. We normalized the importance scores generated by each contributing model. Next, we assigned each normalized importance score a weight: the inverse of the test root-mean-squared error (RMSE) produced by that model. Finally, we calculated the inverse-RMSE weighted average of those normalized importance scores.

The results of our analysis are summarized below.

FIGURE 1: FEATURE IMPORTANCE IN ACCIDENTS THAT INVOLVE AN INJURY, TARGETING MAXIMUM SEVERITY RESULTING FROM CRASH

FEATURE	IMPORTANCE
Whether or not the injured was transported to a hospital	0.2133
Event was a collision with other MVIT	0.0980
Whether or not restraint equipment was used by injured	0.0721
Whether or not the vehicle was rear-ended	0.0364
Age of injured, at time of crash	0.0346
Whether or not the airbag was deployed	0.0301
Hour of day in which crash occurred	0.0288
Month in which crash occurred	0.0253
Whether or not the crash occurred in a junction area	0.0248
Whether or not the driver was charged with a violation	0.0237
Whether or not EMS Ground tranporation was used	0.0219

FIGURE 2: FEATURE IMPORTANCE IN ACCIDENTS THAT INVOLVE AN INJURY, TARGETING NUMBER OF PERSONS INJURED RESULTING FROM CRASH

FEATURE	IMPORTANCE
Number of people in MVIT involved in crash	0.5048
Whether or not the injured was transported to a hospital	0.1161
Number of MVIT involved in crash	0.0354
Whether or not the injured person was driving	0.0336
Whether or not the vehicle was a bus	0.0313
Number motor vehicles (incl. parked) involved in crash	0.0299
Hour of day in which crash occurred	0.0268
Month in which crash occurred	0.0209
Age of injured, at time of crash	0.0146

Findings

MAXIMUM SEVERITY

Figure 1 lists the importance of the top 11 features when targeting Maximum Severity. “Whether or not the injured was transported to a hospital” is most indicative of the severity of the injury, not surprisingly so. We assume fatalities occurring on the spot would be in this category. Also cracking the top 20 list were those specifically transported by emergency medical services (EMS) ground and EMS air. Intuitively, those brought in by helicopter were in grave need of medical care. Obviously, hospitalization would be a key question for a claims adjuster to ask when fronting reserves on an auto crash.

The CRSS variable Sequence of Events, in general, describes the crash. This is a multilevel category that includes subcategories such as “Non-harmful events,” “Non-collision harmful events,” “Collision with object not fixed (other than vehicles),” and ”Collision with fixed object.” It is understandable that the next feature of importance then, “Event was a collision with other MVIT (motor vehicle in transit),” would be indicative of injury, more so than the other subcategories contained in this variable. In fact, no other Sequence of Events-based variable made the top 20.

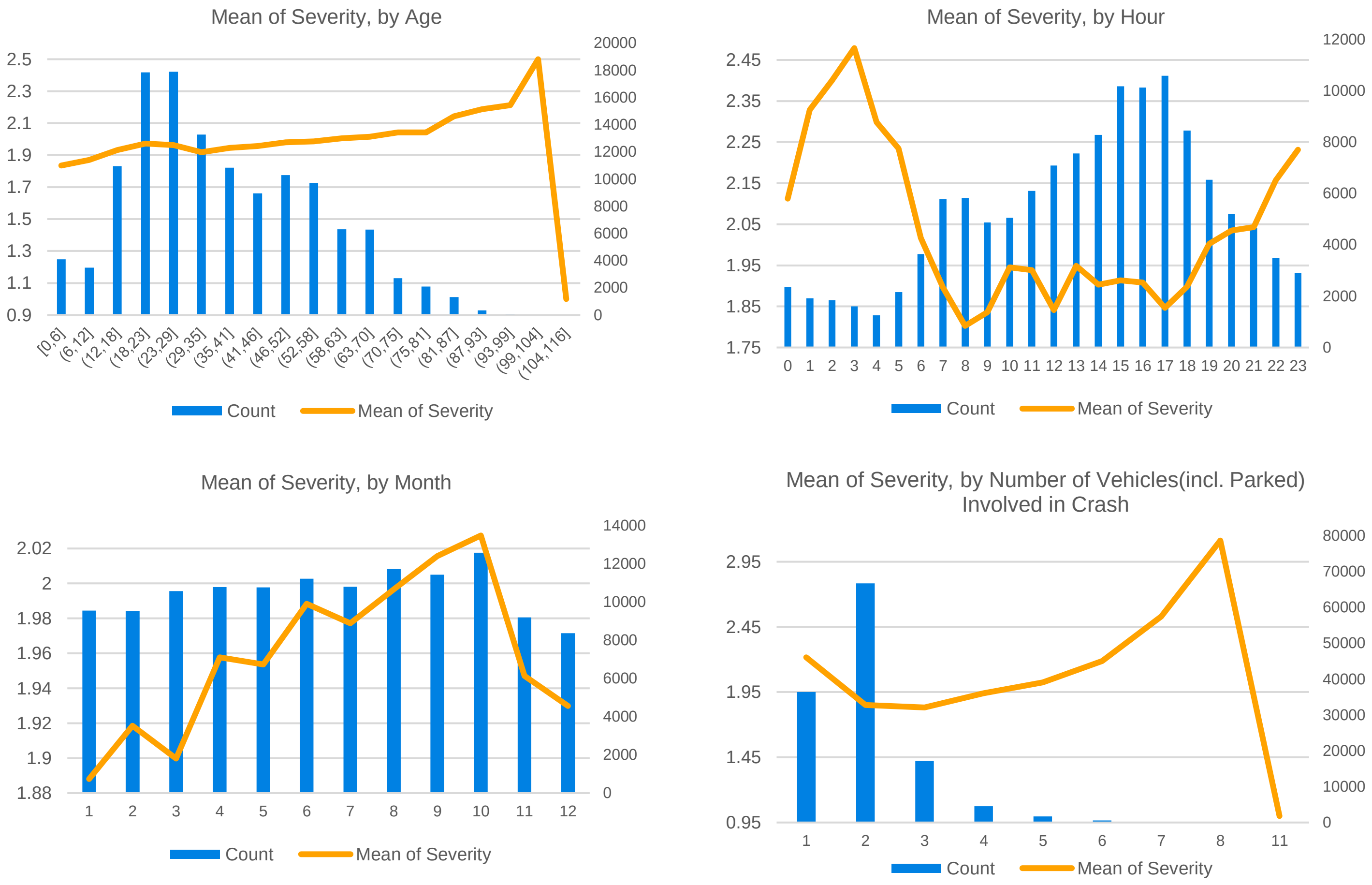
The next item is a reminder to wear those seat belts! CRSS variable Restraint Use describes any safety restraint equipment used, including a helmet for motorcyclists; the feature “Whether or not restraint equipment was used by injured” emerged as an important determinant of severity.

”Whether or not the vehicle was rear-ended” was also an important feature, derived from a CRSS variable indicating impact point from another vehicle. Other impact points between cars were far less important.

“Age of injured, at time of crash” rounded out the top five. With this being a numerical variable, a little more exploration was needed within SOLYS to see if any specific age had a stronger correlation than others. Figure 3 exhibits the cross-analyses charts of Maximum Severity with a few numerical variables. It shows that older persons are more susceptible to injury, especially ages 75 and up. You can also clearly see where the spikes are in “Month in which crash occurred” (September, October), and “Hour of day in which crash occurred” (early morning hours). One could surmise that this is due to things like early snowfall mixed with unprepared drivers and tires, icy roads, or sun glare, as well as sleepy drivers who should have pulled over for the night.

“Number of motor vehicles (incl. parked) involved in crash” is also represented in this figure. It is interesting to see that the average Maximum Severity decreases as we go from one vehicle being involved in an accident to two. For accidents involving three or more vehicles, the Maximum Severity increases until a steep drop from eight vehicles to 11, but the paucity of data at this level should give pause to drawing any conclusions from this abrupt drop.

FIGURE 3: MEAN OF MAXIMUM SEVERITY, BY VARIOUS IMPORTANT FEATURES



Mean of Severity, by Age

Mean of Severity, by Hour

Mean of Severity, by Month

Mean of Severity, by Number of Vehicles(incl. Parked) Involved in Crash

NUMBER OF PERSONS INJURED

Figure 2 above lists the importance of the top nine features when targeting Number of Persons Injured. As you could guess, “Number of people in MVIT involved in crash” has the highest influence on the number injured.

As with Maximum Severity, whether transportation to the hospital was used was of high importance, coming in second-highest.

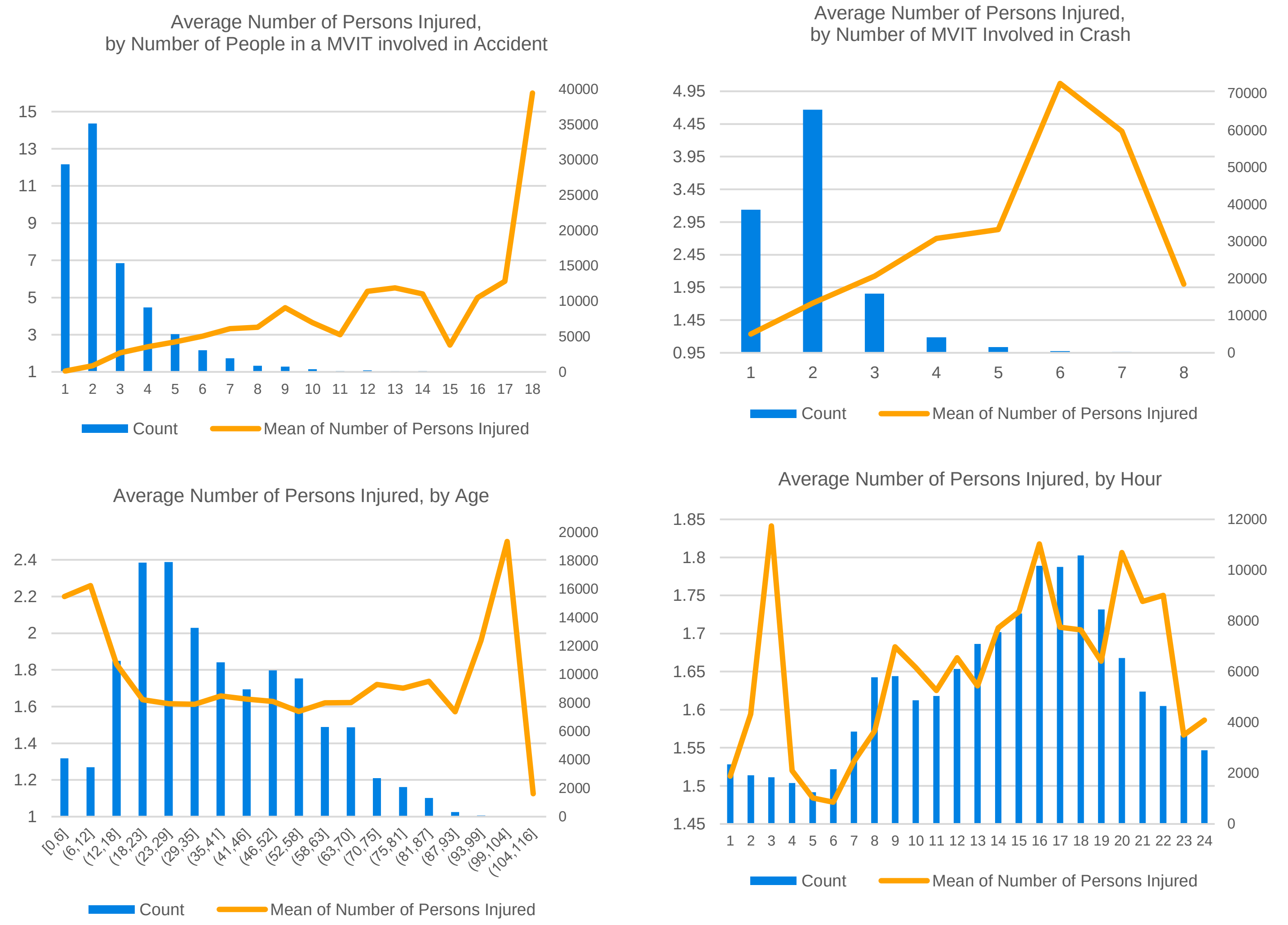
In Figure 4, the numerical variables are graphed alongside the target variable Number of Persons Injured, similarly to Figure 3 (for Maximum Severity). The third-highest explanatory variable, “Number of MVIT involved in crash,” is also shown in Figure 4.

Generally, more vehicles in an accident corresponds to more people involved.

There were similar spikes to Number of Persons Injured as Maximum Severity for “Hour of day in which crash occurred” and “Age of injured, at time of crash,” namely, early morning and older occupants, respectively. “Age of injured, at time of crash” also shows higher areas for children (ages 18 and under), which follows from the fact that a minor in a car necessitates at least two people being present in the car (as minors cannot drive). “Age of injured, at time of crash” and “Hour of day in which crash occurred” were numbers seven and nine in the importance list.

Two variables of importance for Number of Persons Injured not previously mentioned for Maximum Severity, are ranked fourth and fifth: “Whether or not the injured person was driving,” and “Whether or not the vehicle was a bus.” Naturally, a bus will hold more passengers subject to injury than, say, a motorcycle.

FIGURE 4: MEAN OF NUMBER OF PERSONS INJURED, BY VARIOUS IMPORTANT FEATURES



Average Number of Persons Injured, by Number of People in a MVIT involved in Accident

Average Number of Persons Injured, by Number of MVIT Involved in Crash

Average Number of Persons Injured, by Age

Average Number of Persons Injured, by Hour

WORKING IN SOLYS

SOLYS is predictive modeling software with an elegant graphical user interface developed by Milliman's Paris-based Casualty practice. SOLYS enables the user to perform complex predictive modeling projects from start-to-finish, without requiring a PhD in data science, or senior-developer-level expertise in Python, R, or any other programming environment typically required to perform these tasks. The interface is visually organized in a way that follows the standard flow of a predictive modeling project, and relevant analyses and summaries are automatically generated (and easily exported) along the way.

The data science workflow is neatly organized into the following five overarching stages that are navigated alongside the main display of the tool:

1. Project
2. Data Management
3. Algorithms
4. Visualizations
5. Dashboard Graph

Below, we describe the process and tasks performed in each stage of our analysis.

Project

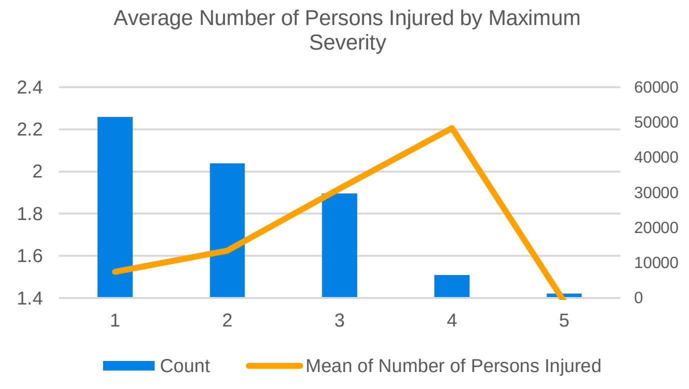
The “Project” stage enables users to create new projects and view existing projects. Additionally, they can each add or remove team members from specific projects, as well as edit their names and descriptions. We used this stage to create different projects for each of the different models we tuned, so that things would stay overtly organized with a number of different models in one or several general purpose projects

Data Management

In the “Data Management” stage, we managed our data sources in the “Data Source” step. Once data was uploaded to the appropriate location, this is where we brought it into our project. Next, we used the “Data Frame” step to read in our comma-separated values files. After reading in, we were able to peek at the data we imported, as well as specify parameters for reading in the data (such as what kind of delimiter was used in the file). The “Data Frame” step also allows the user to load data from SQL Server and Postgre SQL servers after providing server addresses, ports, database names, table names, usernames, and passwords.

In the “Exploration” step, we used the Summary Generator to perform Mono and Cross analyses on our data frames, such as those displayed in Figure 5, helping us get a better picture of the data we were modeling. This step also allows for data exploration using SQL queries.

FIGURE 5: AVERAGE NUMBER OF PERSONS INJURED, BY MAXIMUM SEVERITY



In the “Preparation” step, the “Join” tool enables the user to join data sets (choosing from multiple join types), or stack them. Joining data sets was very important for our task, given that the CRSS data set is comprised of 19 different, overlapping, keyed data sets. “Preparation, Drop” is where we dropped variables deemed redundant or likely to leak data. ”Preparation, Add” is where we added variables, such as our one-hot-encoded features.

At the “Sampling” step of Data Management, we split our master data set (comprised of all joined 19 data sets) into a Training and a Test data set, using the conventional 70/30 allocation. SOLYS allows the user to name the new data frames, and automatically appends the names with “_train” or “_test,” making for convenient reference later in the process. SOLYS also allows the user to specify a “seed,” which allows for replication between projects, if one wishes to use identical training/test splits. We chose the recommended "42" as our seed, given its status as the Answer to the Ultimate Question of Life, the Universe, and Everything (*Hitchhiker’s Guide to the Galaxy* by Douglas Adams).

Algorithms

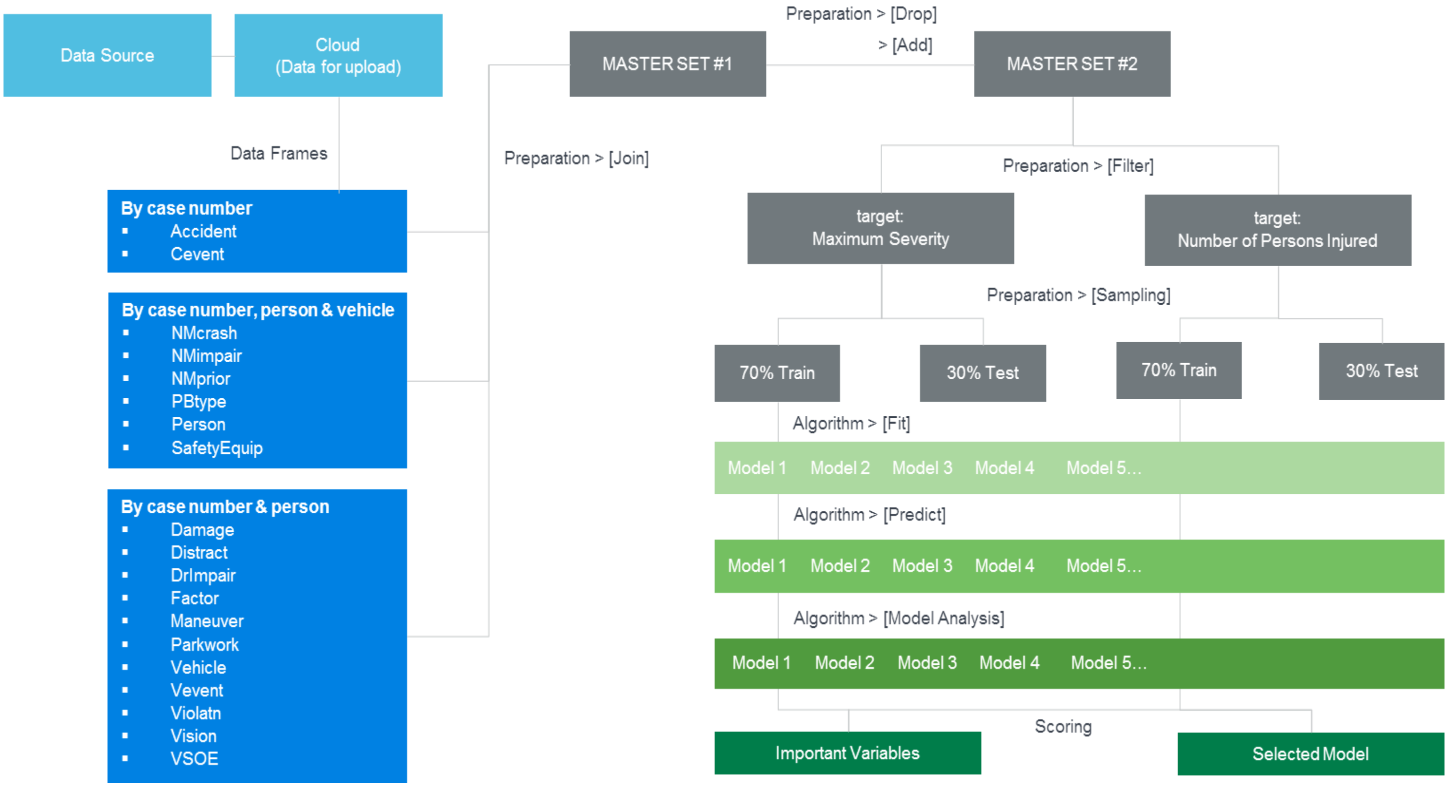
The “Algorithms” stage is where most of our work happened. In the “Model Management” step, we created new algorithms, and were able to view algorithms that already existed in the project. The “Fit” step allows the user to parameterize and fit the models using a choice of Spark in Python (recommended for big data), or R (recommended for small data). We found the R implementation suitable for our modeling needs. After selecting our algorithm and the data frame to work on, we were able to specify which model to use, as well as its parameters. SOLYS offers a variety of models to choose from: classification and regression trees (CART), generalized linear model (GLM), bootstrap aggregating (bagging), random forest, gradient boosting, and K-means clustering. We found bagging models to have the lowest error, based on the inverse RMSE in this situation. See Figure 6 for a comparison of model errors among our selected final models.

FIGURE 6: MODEL PERFORMANCE

TARGET	MODEL	TRAIN RMSE	TEST RMSE	INVERSE OF TEST RMSE	INVERSE TEST RMSE WEIGHT
Maximum Severity	CART	0.8088	0.8197	1.2200	24%
	Bagging	0.6511	0.6768	1.4775	30%
	Gradient Boosting	0.8407	0.8432	1.1860	24%
	Random Forest	0.8950	0.8950	1.1173	22%
Number of Persons Injured	CART	0.7675	0.7859	1.2724	26%
	Bagging	0.6195	0.6443	1.5522	31%
	Gradient Boosting	0.8415	0.8334	1.1998	24%
	Random Forest	1.0432	1.0640	0.9398	19%

After the model fit concludes, SOLYS generates a number of informative plots and tables, which include: feature importance, iterative error, and finalized fit parameter information. At the "Predict" step, we selected our training data frame and the algorithm to evaluate, then named our prediction data frame that would result. This predicted data frame is nearly identical to the one specified at this stage, with an added column for the predicted variable. In the subsequent "Model Analysis" step, we were able to specify whether to analyze as a regression or a classification algorithm, and SOLYS generated informative figures about the error of the model. We used the "Predict" and "Model Analysis" steps to get RMSE for our initial tuning-stage models, and after satisfied with those, chose our final models, then predicted again, except using the holdout test data, to use the (inverse of) the resulting RMSE as its weight when we would combine the feature importance at the end of the process. Figure 7 summarizes our general approach to modeling.

FIGURE 7: MODELING WORKFLOW



Visualizations

The Visualizations stage enables the user to upload data to be visualized graphically, which can be viewed in the subsequent Dashboard Graph stage. In this stage, we uploaded the top features shown at the start of this article, along with our importance metric, to visualize their relative importance. See Figures 8 and 9 below:

FIGURE 8: IMPORTANCE OF FEATURES – MAXIMUM SEVERITY

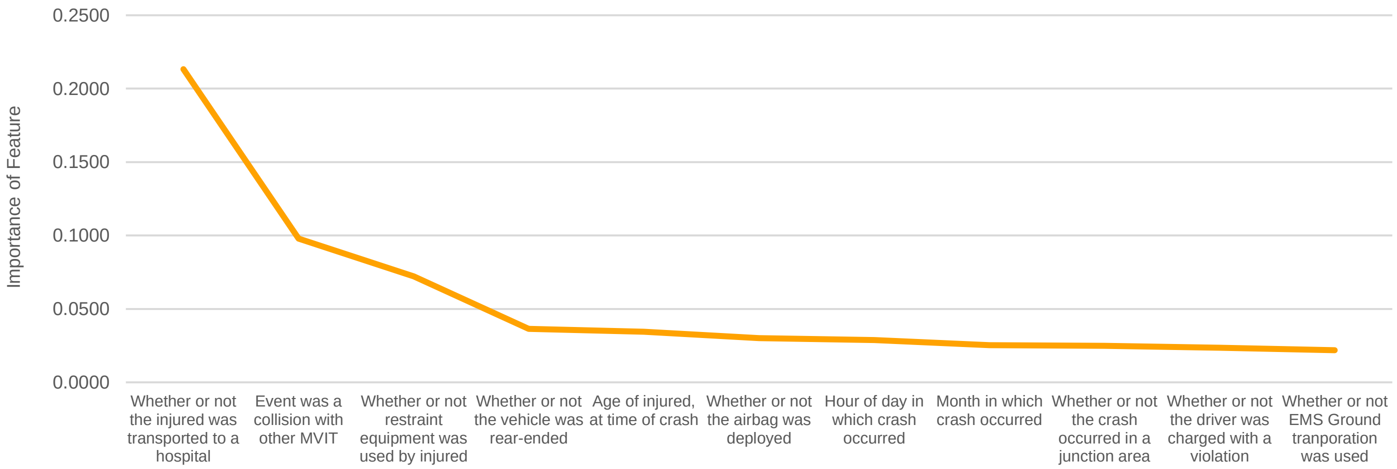
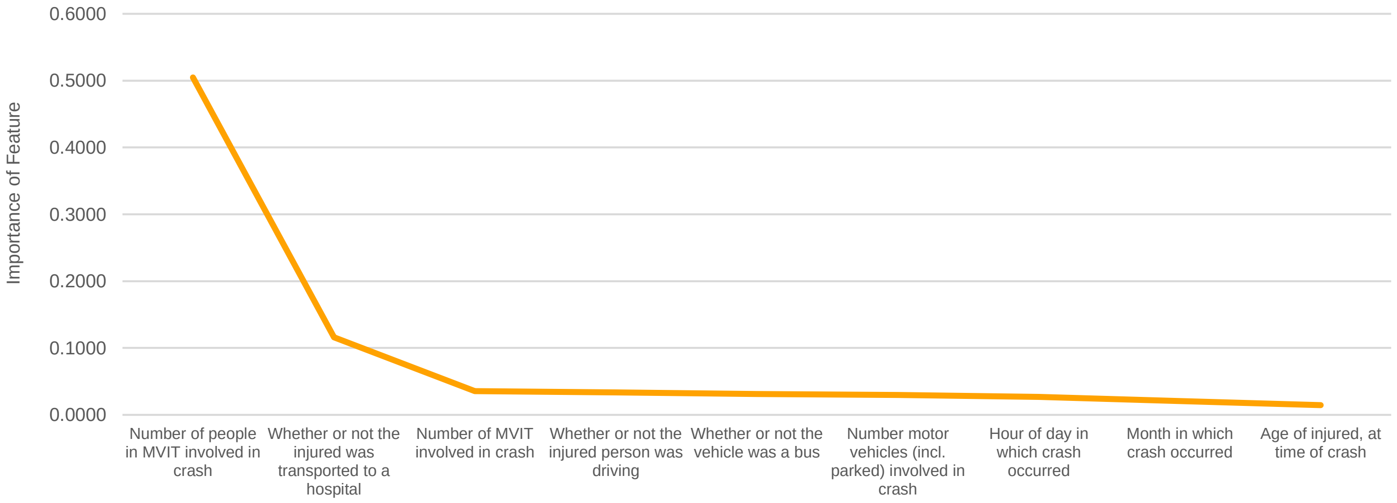


FIGURE 9: IMPORTANCE OF FEATURES – NUMBER OF PERSONS INJURED



Dashboard Graph

The "Dashboard Graph" stage allows the user to stack and load visualizations created in the "Visualization" stage. This is a crucial, but perhaps overlooked, portion of the machine learning process. Being able to go from initial exploration to graphing widgets usable in the boardroom in the same tool provides tremendous value. Additionally, this stage allows the analyst to easily reference important visualizations while they are working in other parts of the SOLYS system.

Closing

SOLYS gave us an all-inclusive, easy-to-use, powerful suite of predictive modeling tools. It enabled us to complete every step of a predictive modeling task: loading and preparing data, modeling, presentation, and everything in between. Along the way we were able to generate new insights about what factors influence injury in automobile crashes. So if you want to be safest when driving, be sure to only get on the road midday in January, and of course, wear your seat belt!

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
Liz Young
liz.young@milliman.com
Liz's pronouns: They/Them/Theirs

Kayla Spitzner
kayla.spitzner@milliman.com
Kayla's pronouns: She/Her/Hers