SOLYS

A collaborative analytics platform

Gene Dan, FCAS, MAAA, CSPA
Olivia Esterlis
Tony Huang, PhD
Ora Suslovich
Cindy Yang
Yoel Zuman

SOLYS is a predictive modeling platform, internal to Milliman, based on Apache Spark, a powerful, open-source, distributed computing system. As part of a company-wide, performance-testing initiative, we were tasked with using SOLYS to answer two questions regarding the Crash Report Sampling System (CRSS) of the National Highway Traffic Safety Administration (NHTSA):

1. **Which factors have the strongest association with accidents that involve an injury?**
2. **Which factors have the strongest association with the number of persons injured in an accident?**

Our team, composed of members from the Chicago Cyber Risk Solutions and New York Casualty practices, approached these questions by deploying a combination of generalized linear models and machine learning methods such as gradient boosting machines and random forests against the CRSS data set, hosted on our local SOLYS cluster.

Our efforts resulted in the engineering of over 5,000 variables and the selection of the two best models out of more than 50 candidate models. Although these two models have almost 50 variables apiece, we present the 20 most important variables from each model in the table in Figure 1.

We determined that generalized linear models were the best models for this project, on the grounds of variable reasonableness, model parsimony, and model practicality. Although machine learning methods sometimes offered superior predictive performance, we did not believe that this advantage outweighed the softer, more qualitative aspects of predictive modeling—such as model interpretability.

Our goal is to not only predict, but also to explain, inform, and persuade—and because of these human aspects, we selected generalized linear models (GLMs) for their strong predictive performance and mathematical elegance.

During the course of our analysis, we discovered that head-on collisions (clock point 12), the presence of pedestrians, motorcycles, and rollovers were major predictors of automobile accident injuries.

FIGURE 1: MOST IMPORTANT VARIABLES IDENTIFIED (EXCL. COEFFICIENTS)

VARIABLE RANK	MODEL 1: ACCIDENTS INVOLVING INJURY	MODEL 2: NUMBER OF PERSONS INJURED
1	At least one motorcycle involved	Number of vehicles hit at clock point 12
2	Number of vehicles with disabling damage	At least one vehicle with a front airbag deployed
3	At least one vehicle with a front airbag deployed	Number of pedestrians
4	At least one rollover or overturn occurred	Number of rollover or overturns involved
5	Number of pedestrians	At least one passenger in transit
6	At least one pedestrian or pedacyclist was not in a school zone	At least one vehicle with no airbag deployed
7	At least one pre-event object or animal involved	At least one pre-event object or animal involved
8	Number of vehicles hit at clock point 12	Imputed number of females involved
9	At least one female was involved	Number of vehicles with disabling damage
10	Number of vehicles traveling between 1 and 20 miles per hour	Accident not at an intersection
11	Number of pre-event backing actions	Number of motorists involved
12	At least one non-motorist crossing roadway	Number of vehicles with minor damage
13	Number of persons who did not use a restraint	At least one person sitting on the second seat, left side
14	Imputed total model age of vehicles	At least one person was in a front seat other than left, middle, or right
15	Number of vehicles with minor damage	Number of vehicles traveling between 1 and 20 miles per hour
16	Number of pedestrians or pedacyclists not at a crosswalk	Number of people aged between 41 and 60
17	At least one passenger in transit	Imputed total model age of vehicles
18	At least one person took an alcohol blood test	At least one person used no restraint
19	At least one vehicle with a combination of airbags deployed	Median number of occupants
20	Number of blacked-out drivers prior to critical event	At least one two-way divided unprotected median involved

In addition to identifying variable importance, GLMs also provide coefficient magnitudes—that is, whether a variable positively or negatively contributes to accidents involving injury or the number of persons injured. For example, you may wonder why slow-moving vehicles (number of vehicles traveling between 1 mph and 20 mph) was identified as an important variable. It was due to the fact that this variable was shown to negatively contribute to the likelihood of an accident resulting in an injury, as seen in the table in Figure 7 below. In order to properly stratify and rank the likelihood of an accident involving an injury, we must identify not only the variables that are associated with the most severe accidents, but also mitigating factors. Practical applications of such a model include triaging first responders and setting case reserves for insurance claims.

On the other hand, while the machine learning methods were able to identify variable importance, they could not identify the extent to which those variables were either positively or negatively associated with accidents involving injury or the number of persons injured. This was a major consideration in our selection of GLMs over the machine learning methods.

The CRSS data set

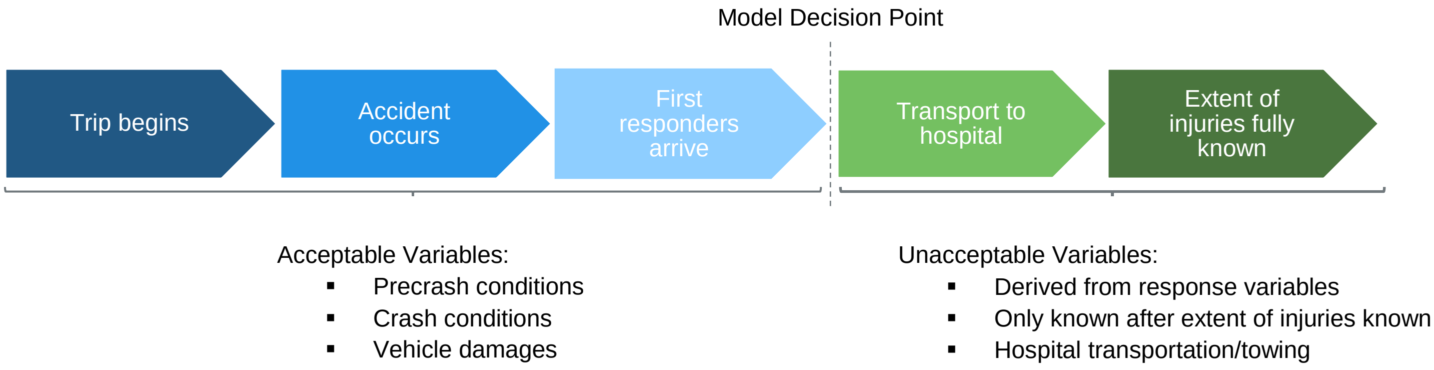
The Crash Reporting Sampling System (CRSS) contains information on police-reported automobile crashes, including vehicle, personnel, and other circumstances related to accidents.

The table in Figure 2 summarizes the 22 data files used in the analysis. Via the SOLYS Jupyter notebook environment, we performed feature engineering—that is, created new variables from combinations of existing variables—to expand the original 504 variables contained in the data set into 5,173 variables for modeling.

FIGURE 2: SUMMARY OF CRSS DATA ELEMENTS

SOURCE FILE	ORIGINAL VARIABLE COUNT	ENGINEERED VARIABLE COUNT
Accidents	51	211
Vehicles	87	1,913
People	61	486
Parked Vehicles	50	519
Pedestrians	31	569
Crash Events	15	243
Vehicle Events	17	243
Vehicle Events (continued)	13	197
Damage	11	33
Distractions	11	47
Driver Impairments	11	29
Vehicle Factors	11	43
Maneuvers	11	21
Violations	11	175
Visuals	11	39
Circumstances	12	47
Non-motorist Impairments	12	27
Non-motorist Actions	12	31
Safety Equipment	12	17
Accident (Auxiliary)	26	83
Vehicle (Auxiliary)	9	41
Person (Auxiliary)	19	159
TOTAL	**504**	**5,173**

FIGURE 3: MODELING SCENARIO



Modeling scenario

Operational failure arises when models fail to consider the practical and human aspects of the scenario at hand. Even talented modelers may inadvertently include information within the model that will only be available after the model makes its decision. Models that appear to be highly predictive in a test environment oftentimes fail in production, leading to costly mistakes.

Therefore, we discussed the need to balance predictive accuracy and practicality. We made the assumption that the model decision point would occur shortly after the arrival of first responders, but before the towing of vehicles and transporting of victims to the hospital. We eliminated all variables that occur after the decision point from consideration, illustrated in Figure 3.

Variable selection

With over 5,000 variables under consideration, it was necessary for the team to use automated selection algorithms within SOLYS to determine what variables would go into the models. A combination of elastic net and tree models was iteratively deployed to rank the variables by importance, with the top 50 considered for each model.

We incorporated human judgment into the process by evaluating the variables for reasonableness and removed those deemed undesirable upon each iteration. This iterative process is depicted in Figure 4.

FIGURE 4: VARIABLE SELECTION PROCESS

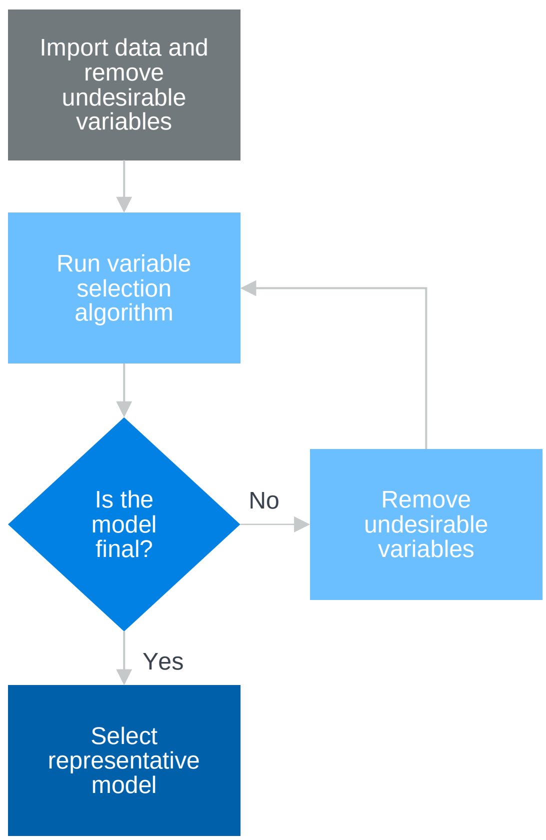
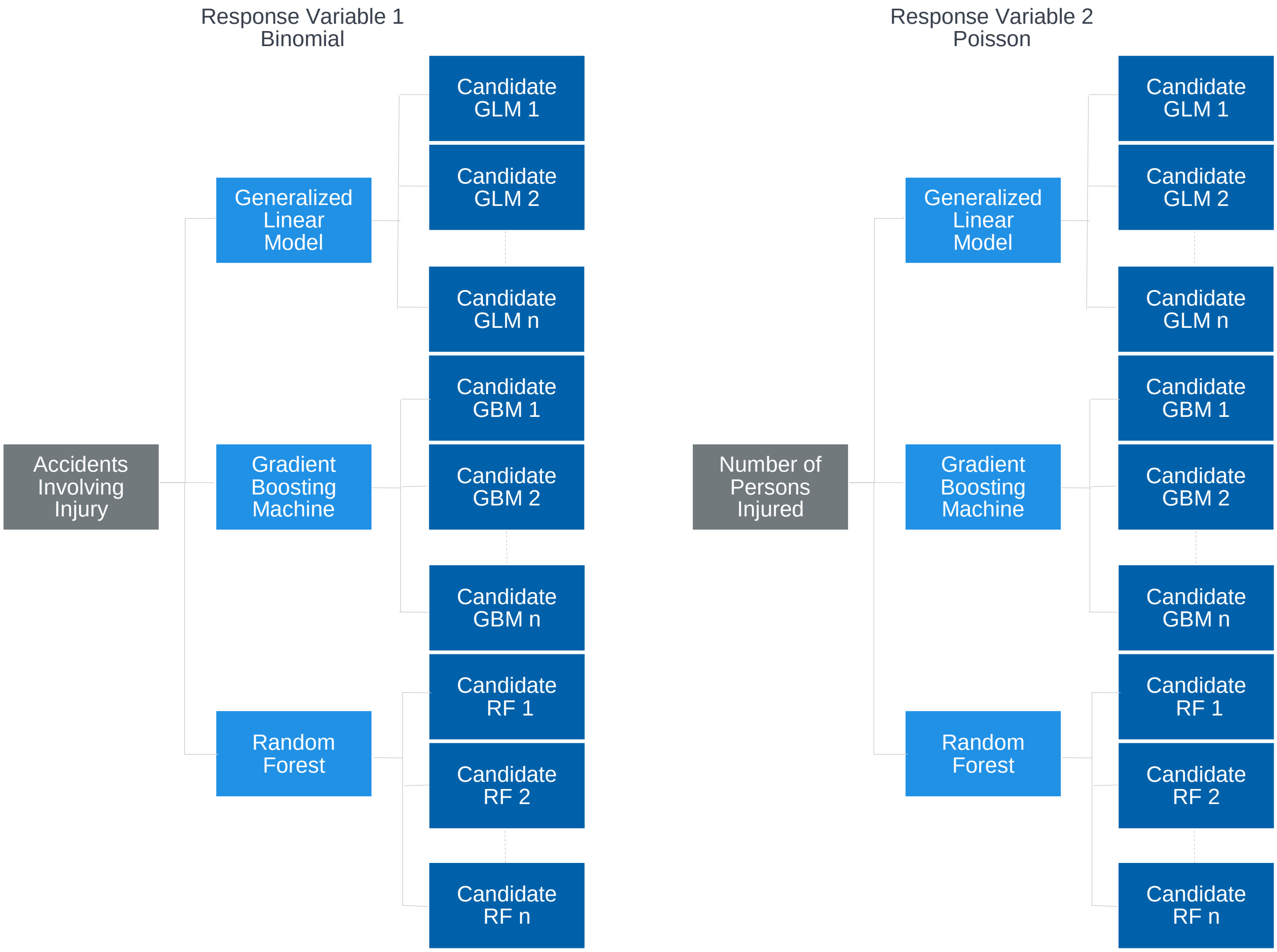


FIGURE 5: MODELS CONSIDERED



Models considered

For each response variable, we considered three types of models in SOLYS:

1. Generalized linear model (GLM)
2. Gradient boosting machine (GBM)
3. Random forest (RF)

Generalized linear models are commonly used in insurance applications. Their widespread acceptance by non-actuarial professionals, such as underwriters and claims adjusters, made them a natural choice to consider.

The machine learning models—GBMs and RFs—are gaining popularity among data scientists because they often produce superior predictions to GLMs. However, this improved accuracy comes at the expense of parsimony and transparency.

For each algorithm (GLM, GBM, RF), we selected one model out of a pool of candidate models, based on variable reasonableness. We then scored these models against each other to make a final selection.

Scoring and validation

To test the predictions, we used SOLYS to perform cross-validation—an iterative procedure that scores models on a portion of the data set that is not used for model fitting. This procedure generates a set of fit statistics for evaluating predictive performance:

1. Area under the curve (AUC)
2. Log loss
3. Root mean squared error (RMSE)
4. Mean squared error (MSE)
5. Root mean squared logarithmic error (RMSLE)
6. Mean absolute error (MAE)

We desire to maximize the AUC, while minimizing the other statistics.

Figure 6 summarizes the fit statistics for each response variable (accidents involving injury, number of people injured) and algorithm considered. The categories of statistics differ between the two response variables (i.e., no AUC for number of people injured) due to the distributions modeled (binomial for accidents involving injury, Poisson for number of people injured).

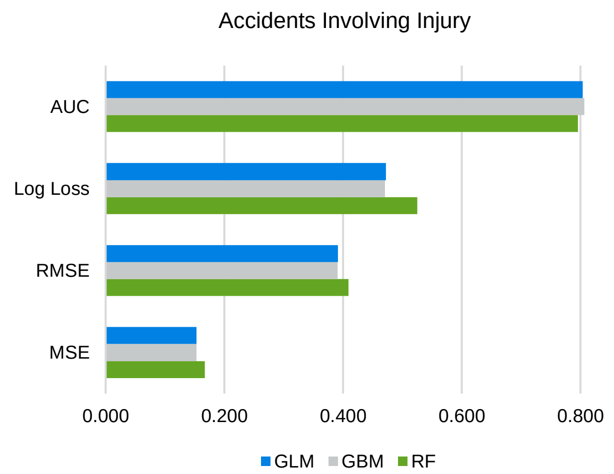
The results show that GLM outperformed RF on all metrics. GBM barely outperformed GLM for the first response variable, while mostly underperforming on the second.

We determined that the machine learning methods, GBM and RF, did not improve predictions enough to warrant choosing them over the elegant parsimony of GLM. We therefore chose GLM as our final model for both response variables.

Model results

The tables in Figures 7 and 8 show the selected GLM models for both response variables. Variables were ranked by the absolute value of standardized coefficients (derived from standardized parameters), but we display only the unstandardized coefficients for clarity and ease of reproducibility.

FIGURE 6: CROSS-VALIDATION HOLDOUT STATISTICS



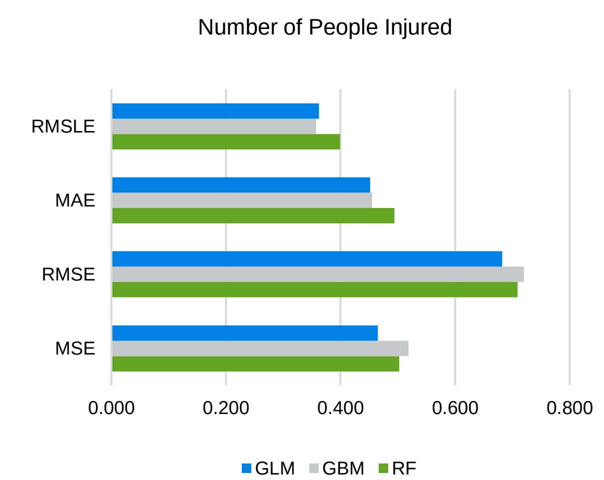


FIGURE 7: MODEL RESULTS: ACCIDENTS INVOLVING INJURY (LOG-LINK)

IMPORTANCE RANK	DESCRIPTION	COEFFICIENT
-	Intercept	-1.811
1	At least one motorcycle involved	3.033
2	Number of vehicles with disabling damage	0.541
3	At least one vehicle with a front airbag deployed	0.887
4	At least one rollover or overturn occurred	1.063
5	Number of pedestrians	1.339
6	At least one pedestrian or pedacyclist was not in a school zone	1.389
7	At least one pre-event object or animal involved	-0.844
8	Number of vehicles hit at clock point 12	0.298
9	At least one female was involved	0.385
10	Number of vehicles traveling between 1 and 20 miles per hour	-0.317
11	Number of pre-event backing actions	-0.973
12	At least one non-motorist crossing roadway	1.363
13	Number of persons who did not use a restraint	0.591
14	Imputed total model age of vehicles	0.014
15	Number of vehicles with minor damage	-0.174
16	Number of pedestrians or pedacyclists not at a crosswalk	1.123
17	At least one passenger in transit	0.269
18	At least one person took an alcohol blood test	1.277
19	At least one vehicle with a combination of airbags deployed	0.688
20	Number of blacked-out drivers prior to critical event	1.418
21	Number of pre-event actions going over the lane on the right side	-0.442
22	At least one vehicle with no airbag deployed	-0.256
23	Minimum age of pedestrians or cyclists involved	0.015
24	Number of people aged between 41 and 60	0.116
25	At least one pre-event braking action	0.330
26	Number of events that involved collision with a standing tree	0.377
27	Four-way intersection involved	0.165
28	At least one person was not ejected	-0.711
29	Number of events that involved reentering a roadway	0.390
30	Number of persons not in motor vehicles in transit	0.355
31	Number of failure to require restraint violations	1.242
32	Number of motorists involved	0.046
33	At least one vehicle was hit at clock point 9	0.183
34	At least one vehicle with side airbag deployed	0.484
35	Number of pre-event actions going straight	0.092
36	Number of persons with no misuse of restraint	0.032
37	Number of vehicles hit at clock point 3	0.113
38	At least one person was in a front seat other than left, middle, or right	0.067
39	At least one pedacyclist or pedestrian was male	0.256
40	Number of vehicles hit at top	0.088
41	At least one other vehicle encroaching from crossing street across path	0.024

FIGURE 8: MODEL RESULTS: NUMBER OF PEOPLE INJURED (LOG-LINK)

IMPORTANCE RANK	DESCRIPTION	COEFFICIENT
-	Intercept	-1.323
1	Number of vehicles hit at clock point 12	0.408
2	At least one vehicle with a front airbag deployed	0.608
3	Number of pedestrians	1.977
4	Number of rollover or overturns involved	1.062
5	At least one passenger in transit	0.397
6	At least one vehicle with no airbag deployed	-0.517
7	At least one pre-event object or animal involved	-0.848
8	Imputed number of females involved	0.170
9	Number of vehicles with disabling damage	0.188
10	Accident not at an intersection	-0.259
11	Number of motorists involved	0.095
12	Number of vehicles with minor damage	-0.223
13	At least one person sitting on the second seat, left side	-0.410
14	At least one person was in a front seat other than left, middle, or right	-0.194
15	Number of vehicles traveling between 1 and 20 miles per hour	-0.165
16	Number of people aged between 41 and 60	0.113
17	Imputed total model age of vehicles	0.008
18	At least one person used no restraint	0.307
19	Median number of occupants	0.065
20	At least one two-way divided unprotected median involved	-0.133
21	At least one vehicle with a combination of airbags deployed	0.301
22	Number of motorcycles involved	0.666
23	At least one person was not ejected	-0.680
24	At least one vehicle was hit at the top	-0.140
25	Number of people in an enclosed passenger or cargo area	0.880
26	At least one person took an alcohol blood test	0.171
27	At least one other vehicle encroaching from crossing street across path	0.110
28	At least one pedestrian or pedacyclist was not in a school zone	-0.193
29	At least one pre-event pedestrian, pedacyclist, or motorist involved	0.162
30	Number of moving license and registration violations	0.046
31	At least one vehicle with side airbag deployed	0.129
32	Regulatory sign other than stop, yield, or school zone	0.152
33	Imputed number of totally ejected people	-0.095
34	Number of violations for failure to require restraint use	-0.070
35	At least one vehicle was hit at clock point 9	-0.004

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

Gene Dan
gene.dan@milliman.com

Olivia Esterlis
olivia.esterlis@milliman.com

Ora Suslovich
ora.suslovich@milliman.com

Cindy Yang
cindy.yang@milliman.com

Yoel Zuman
yoel.zuman@milliman.com