Building a Neural Network Scoring Model with JMP and Microsoft SQL Server to Predict Aggregate Motor Vehicle Driver Alcohol Impairment

David Whitchurch, LSU CARTS, Baton Rouge, LA

ABSTRACT

According to a predictive model used by the National Highway Traffic Safety Administration, fatalities in crashes where at least one motor vehicle driver was alcohol-impaired (BAC >= 0.08%) accounted for 28% of all motor vehicle crash fatalities in 2017. While there has been a decline in this percentage since 2013, there is still much work to be done in reducing the number of fatalities in motor vehicle crashes involving one or more impaired drivers. Since many drivers in motor vehicle crashes are not tested for blood alcohol content, predictive models are used to estimate the number who are under the influence of alcohol at the time of a crash. Louisiana currently uses an “alcohol-involved” model, which predicts whether a driver involved in a motor vehicle crash had BAC of 0.02% or greater. While this model is useful for highway safety education and awareness efforts, a separate model is needed for predicting driver alcohol impairment (BAC of 0.08% or greater).

JMP Pro provides a collection of tools and helpful add-ins which make the process of creating and scoring a classification model very efficient. In this paper, I will explore how JMP Pro, in conjunction with Microsoft SQL Server, can be used for data exploration, model construction, model evaluation, and record scoring of a classification model created to provide aggregate prediction of driver impairment in Louisiana motor vehicle crashes.

INTRODUCTION

The purpose of the model described in this paper is to provide a valid and reliable means of accurately predicting driver alcohol impairment at an aggregate level. The intent of this paper is to give a high-level overview of how the model was developed and subsequently scored using JMP Pro and Microsoft SQL Server, which were used for data preparation, model selection, model construction, model parameter adjustment, model scoring, and model deployment.

DATA COLLECTION AND PREPARATION

Data used for fitting the initial models were obtained from Louisiana State University’s Center for Analytics and Research in Transportation Safety (CARTS) crash data warehouse. The data included 5915 motor vehicle driver crash data records from fatal crashes which occurred between 2010-2017 where driver BAC information had been recorded. Using data from less severe crashes is problematic since by law, all drivers in fatal crashes are required to be tested for blood alcohol content, while in less severe crashes the decision to test is made by the investigating officer. This creates an inherent selection bias in terms of which drivers are tested in non-fatal crashes. Therefore, only data from fatal crashes were used for modeling. These data were imported into a JMP data table from CARTS crash data warehouse using a custom SQL query.

VARIABLE SELECTION AND PREPARATION

This Initial variable selection for all models was informed by the current “alcohol-involved” model (BAC >=0.02%) used in production at CARTS, whose independent variables include crash time, day of week, officer suspected driver condition (e.g., “Normal”, “Drinking Alcohol – Impaired”, “Distracted or Inattentive”, etc.), single-vehicle crash, and driver restraint (seat belt) use. Assessment of variable importance was performed on 11 indicator variables containing this information, as well as 42 other potentially significant variables in order to assess predictor contributions to the model. The assessment was performed using the predictor screening platform in JMP, which uses a bootstrap forest algorithm to identify potential predictors for each response (BAC08 = ‘Y’, BAC08 = ‘N’). For each response, a bootstrap forest model using 500 decision trees was built. The column contributions to the bootstrap forest model for each predictor were then ranked from highest to lowest ("Overview of the Predictor Screening Platform", 2019).

To improve function of the predictor screening platform’s underlying bootstrap forest algorithm and reproduce the indicator variables in the CARTS current predicted alcohol model, 13 formula columns were created in JMP. The two columns included in the final model were calculated as follows:

- *Crash Time* values were grouped into four levels based on the hour in which the crash occurred:10pm-3am, 3am-6am, 6am-6pm, and 6pm-10pm. The new column was named “Crash Hour Range”.

• *Crash Day of Week* values were grouped into two levels, Monday-Thursday and Friday-Sunday to create a new binary column named “Weekend Crash”.

A total of 53 predictors were initially screened in JMP. The results of automated predictor screening for predictors used in the final model are displayed in Figure 1.

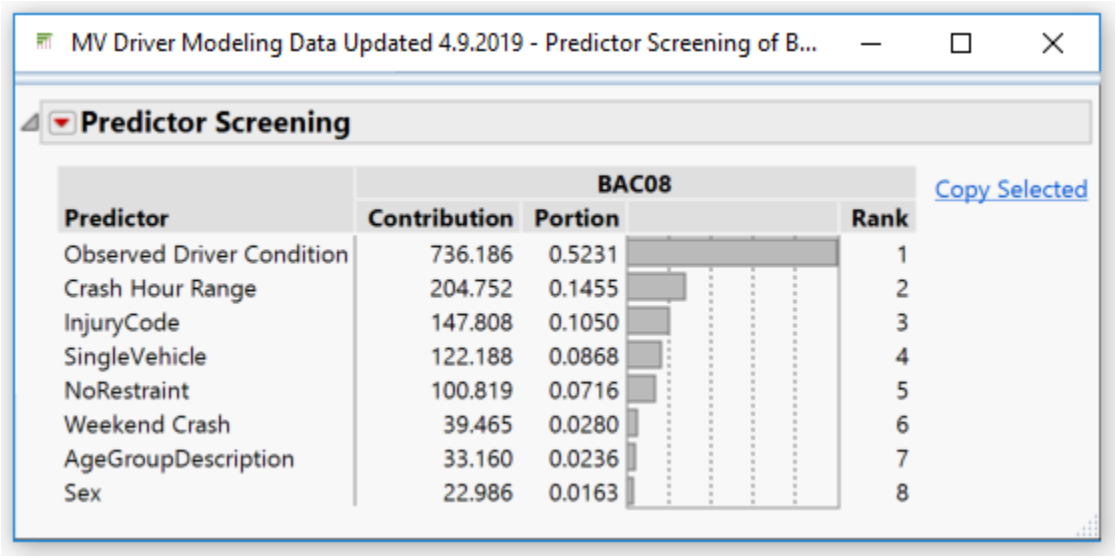

Figure 1 – JMP Predictor Screening Output

After screening, level reduction was also performed on the *Condition Code* column, which contained 14 levels. A new *Observed Driver Condition* column with 5 levels was created in order to increase the performance of the models during the model-fitting process. Finally, data were partitioned into a training set containing 60% of the observations (3490 records) and a validation set containing 40% of the observations (2327 records), stratified on the target variable (BAC08) using JMP’s “Stratified Random” option under the “Make Validation Column” menu.

MODELING

MODEL SELECTION

Models were fit with JMP using the following methods: Neural Network, Partition (Decision Tree), Bootstrap Forest, Boosted Tree, and Logistic Regression. For methods which required indicator variables, JMP automatically created these variables when the models were fit. JMP’s Model Comparison feature was used to compare model results. Figure 2 lists the results, which show the neural network model performed best in terms of Entropy RSquare, Generalized RSquare, RMSE, Mean Absolute Deviation, and AUC. The slight advantage in AUC was determined to be particularly desirable, since the cutoff value was to be adjusted for scoring purposes in order to minimize bias of the aggregate prediction results. For this reason, the neural network model was deemed to be the most appropriate for scoring and use in CARTS production environment.

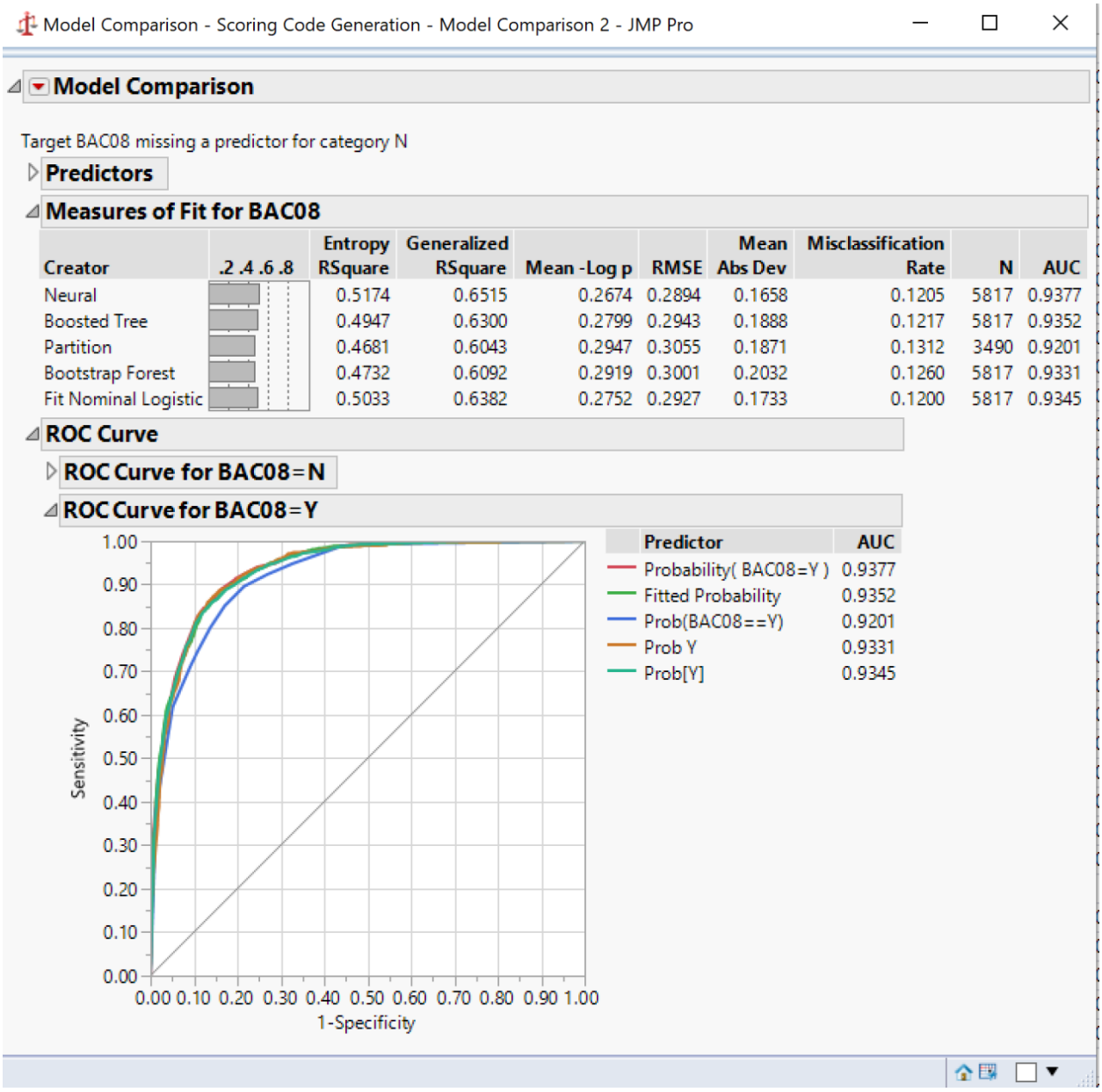

Figure 2: JMP Pro Model Comparison Results

NEURAL NETWORK MODEL DESCRIPTION

Neural networks are models are based on a biological model of activity in the brain and mimic the way that human experts learn. Their main strength is high predictive performance, with a structure capable of capturing very complex relationships between predictors and a response without these relationships being explicitly defined as interaction terms (Shmueli, Bruce, Stephens, & Patel, 2017, pp. 245–246). Another strength of neural networks is their high tolerance for noisy data. The main weakness of neural networks is the inability to provide insight into the structure of the relationship between predictors and responses. Another significant weakness is the tendency of neural networks to overfit to the training data. There is always a danger that if the network only sees cases in a certain range, the predictions outside that range can be invalid (Shmueli, Bruce, Stephens, & Patel, 2017, p. 264). Specific steps described later in the report were taken in order to evaluate and reduce this possibility (see “Model Parameter Settings and Variables”).

MODEL ASSESSMENT

This Results for the neural network model are listed in Figure 3. Using a 0.5 cutoff, the predictive accuracy of the model on the validation data significantly outperformed the accuracy of prediction using the naïve rule (88.27% vs. 76%). Of more importance, the AUC of model results on the validation data was 0.9399. Model results for training data were very similar.

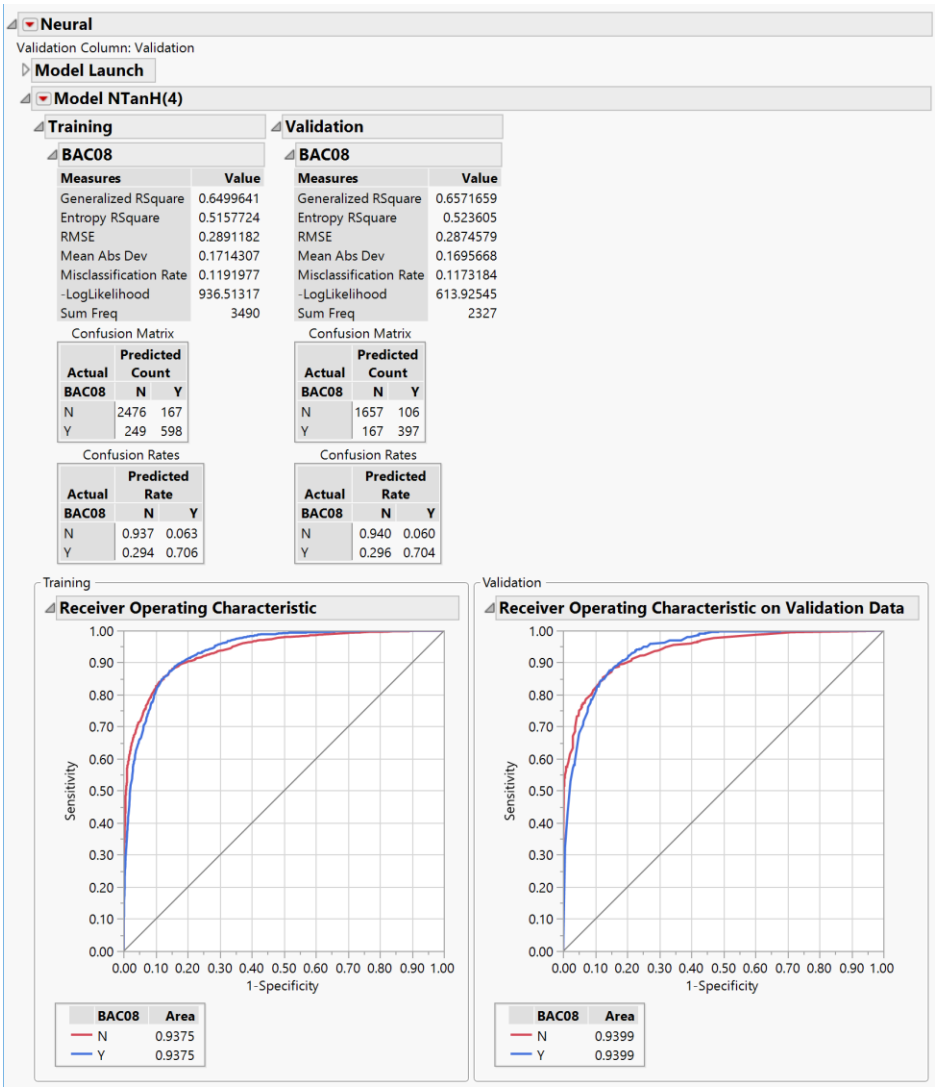

Figure 3 – Neural Network Model Classification Results in JMP

MODEL PARAMETER SETTINGS AND VARIABLES

Model parameters were adjusted according to the following goals to achieve optimized model specifications:

- Reduce the possibility of overfitting.
- Provide a parsimonious model.

The primary neural network parameters in JMP allow the assignment of one or two hidden layers, the activation function, and the number of nodes per layer. A hyperbolic tangent activation function was selected due to the binary classifier (BAC08=Y, BAC08=N). From there, an iterative process was used in which the initial model was refit using 1-2 hidden network layers and varying numbers of nodes per layer. The final parameter settings included one hidden

layer with four nodes. This specification provided the highest degree of model accuracy and similar predictive results on the validation data over multiple model runs.

FINAL MODEL SPECIFICATION

The final model used the hyperbolic tangent function and 8 variables, and 1 hidden layer with 4 nodes. Figure 4 shows a diagram of the final model structure.

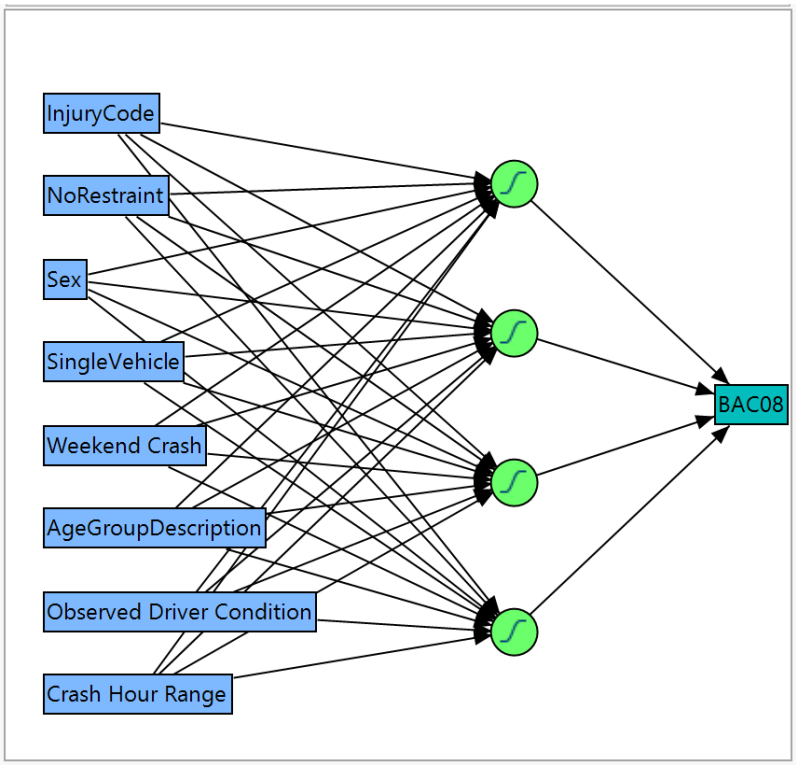

Figure 4 – Diagram of Final Model Structure in JMP

CUTOFF SELECTION

At a 0.5 cutoff, training and validation confusion matrices (Figure 3) indicated a high number of false negatives in comparison to false positives. A standard measure to indicate this relationship is bias, which is calculated as:

$$\text{Bias} = \frac{False\ (+) - False\ (-)}{Total\ Number\ Predicted}$$

Since the model was to be used for aggregate prediction, a major goal was to minimize the absolute value of bias. A bias of zero indicates an equal number of false positives and false negatives. To determine the optimal cutoff, 20 confusion matrix sets were generated using a JMP add-in (Murphrey, 2014) and results were examined. A final cutoff level of 0.44 was selected in order to both minimize bias and help ensure that the overall prediction would slightly favor false positive results over false negative results. Classification matrices based on the 0.44 cutoff for training and validation data, along with bias and other associated results, are shown in Figure 5.

Validation		Predicted	
		N	Y
Actual	N	1622	141
	Y	142	422

Bias	-0.00043
Accuracy	0.878384

Bias Abs Difference	0.000143
Bias Average	-0.00036

Training		Predicted	
		N	Y
Actual	N	2437	206
	Y	207	640

Bias	-0.00029
Accuracy	0.881662

	Training Data	Validation Data
Accuracy	88.17%	87.84%
Error	11.83%	12.16%
Sensitivity	75.56%	74.82%
Specificity	92.21%	92.00%
False Positive	7.79%	8.00%
False Negative	24.44%	25.18%
Bias	-0.03%	-0.04%

Model Variables
Observed Driver Condition
Crash Hour Range
Single Vehicle
Injury Code
No Restraint
Weekend Crash
Age Group Description
Sex

Figure 5 – Final Model Classification Results: 0.44 Cutoff

SCORING AND EVALUATION

MODEL SCORING AND ASSESSMENT

Scoring code for use in SQL Server was created in JMP and subsequently modified in SQL Server for use in CARTS data warehouse ETL processes. This code was applied to Louisiana motor vehicle driver crash data records from 2005 – present. The following columns were created in the data warehouse:

- ProbBAC08Y – probability of driver BAC >= 0.08% (range 0-1)
- PredBAC08 – prediction of driver BAC >= 0.08% (binary 0,1)
- DriverProbBAC08 – (applied to drivers and occupants) probability of driver BAC >= 0.08% (range 0-1)
- DriverPredBAC08 – (applied to drivers and occupants) prediction of driver BAC >= 0.08% (binary 0,1)
- ProbBAC08YCrash – for a single crash, maximum individual driver probability of

BAC >= 0.08% (range 0-1)

- PredBAC08Crash - prediction of whether at least one driver in crash had BAC >= 0.08% (binary 0,1)

For drivers without BAC results, the neural network model formula was applied to calculate results. For drivers with BAC results, the following calculations were used:

- ProbBAC08Y was set to 1 if BAC08 =" Y", 0 otherwise.
- PredBAC08 was set to "Y" if BAC08 =" Y", "N" otherwise.

Data including the new calculated columns were used to create visualizations to further evaluate the fit and accuracy of the model over all available years of the data. Figure 6 shows the number of motor vehicle drivers who were predicted to have used alcohol in fatal crashes. “Predicted BAC08” results shown are based on scoring which includes the neural network model results on driver observations without BAC results as well as actual BAC results where available. The narrowing gap shown between “Predicted BAC08” and “Actual BAC08” over time is primarily due to a marked improvement on collection and reporting of actual BAC results.

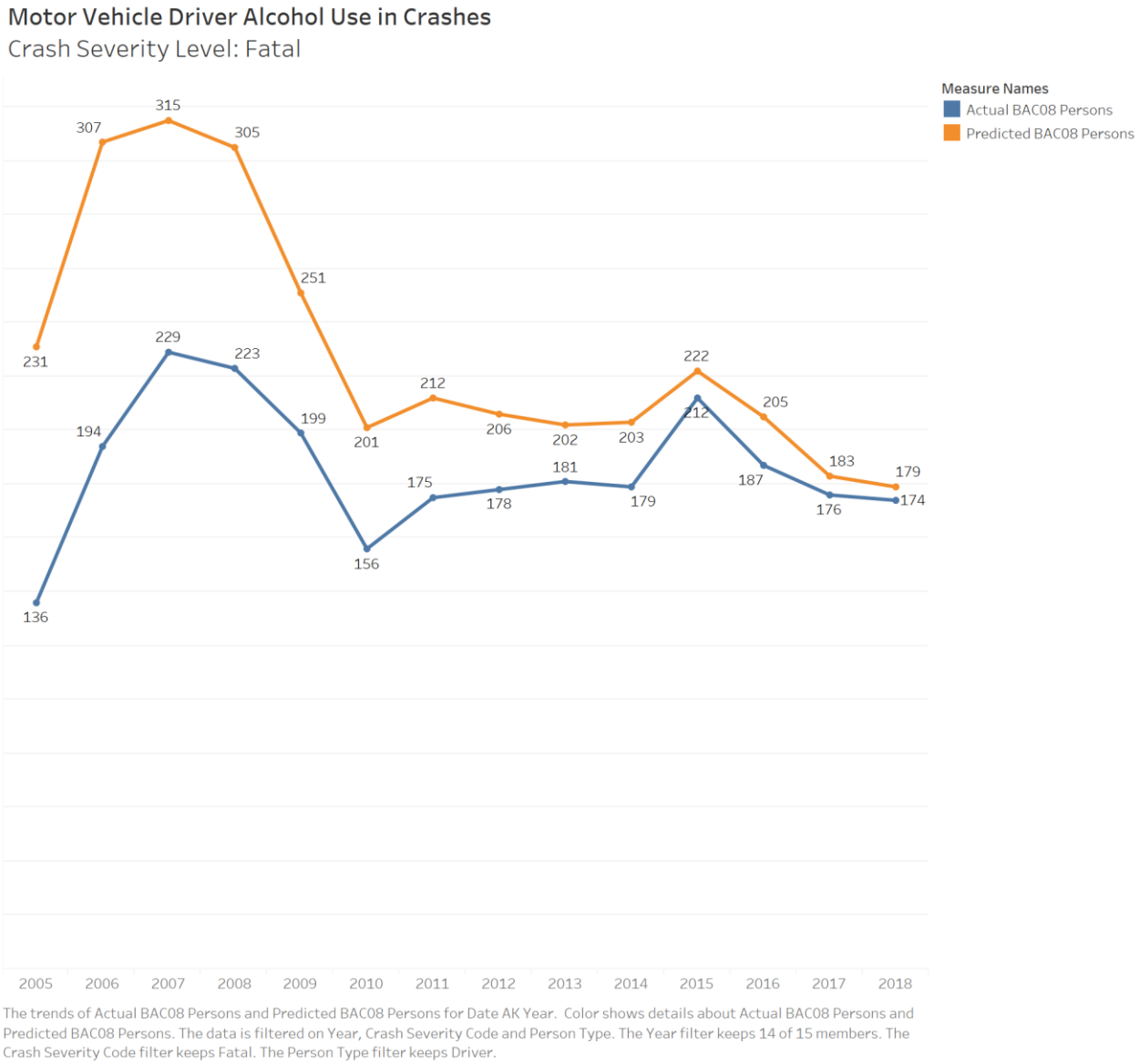

Figure 6 – MV Driver Alcohol Use in Crashes: Actual vs. Predicted

Figure 7 shows the number of fatal injuries in crashes where there was predicted or measured alcohol use by one or more motor vehicle drivers involved in the crash.

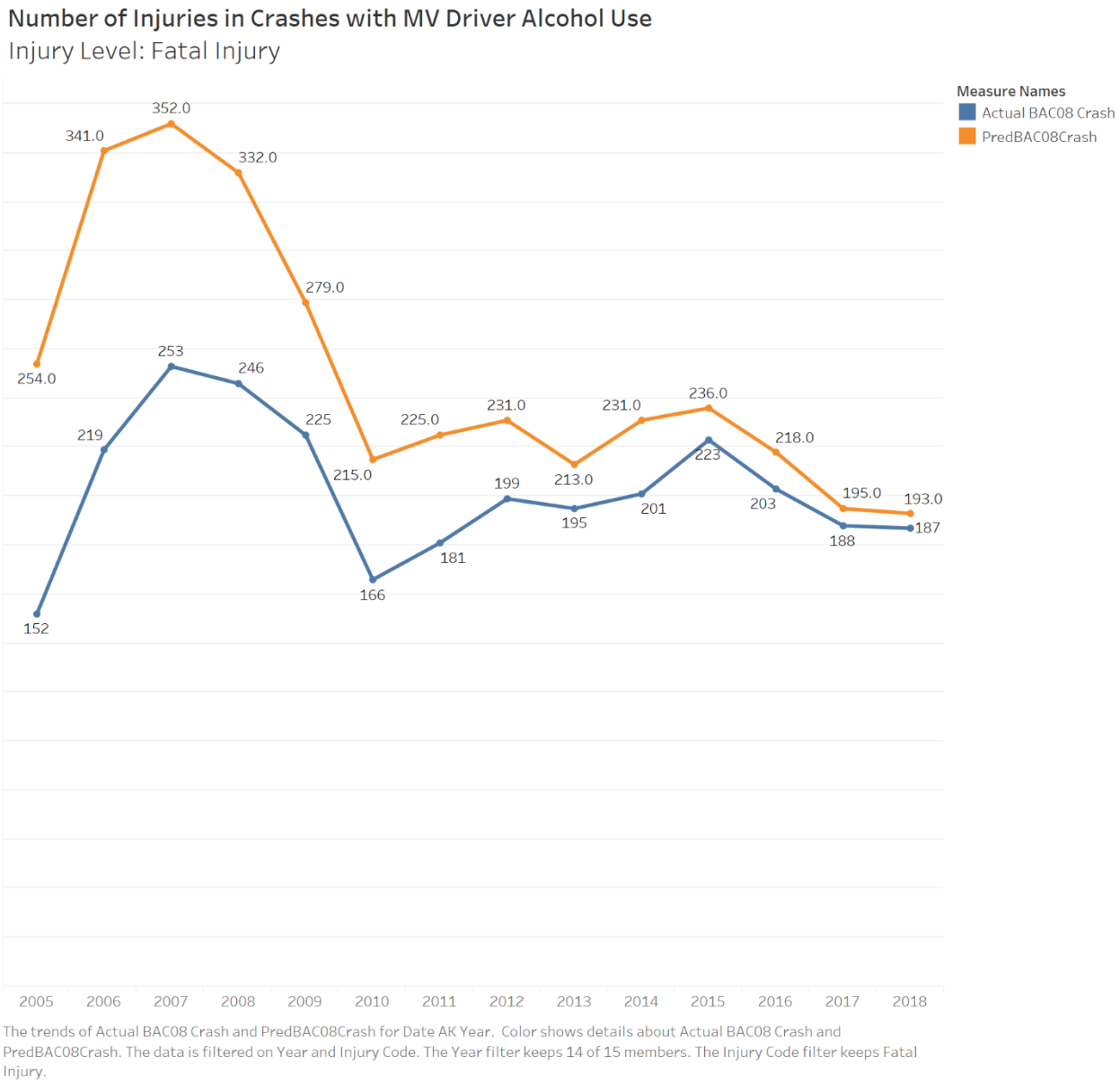

Figure 7 – Number of Injuries in Crashes with MV Driver Alcohol Use: Actual vs. Predicted

Figure 8 shows the percentage of fatalities in crashes where there was predicted or measured alcohol impairment of one or more motor vehicle drivers. It includes results from the most recent NHTSA model (National Highway Traffic Safety Administration, 2017).

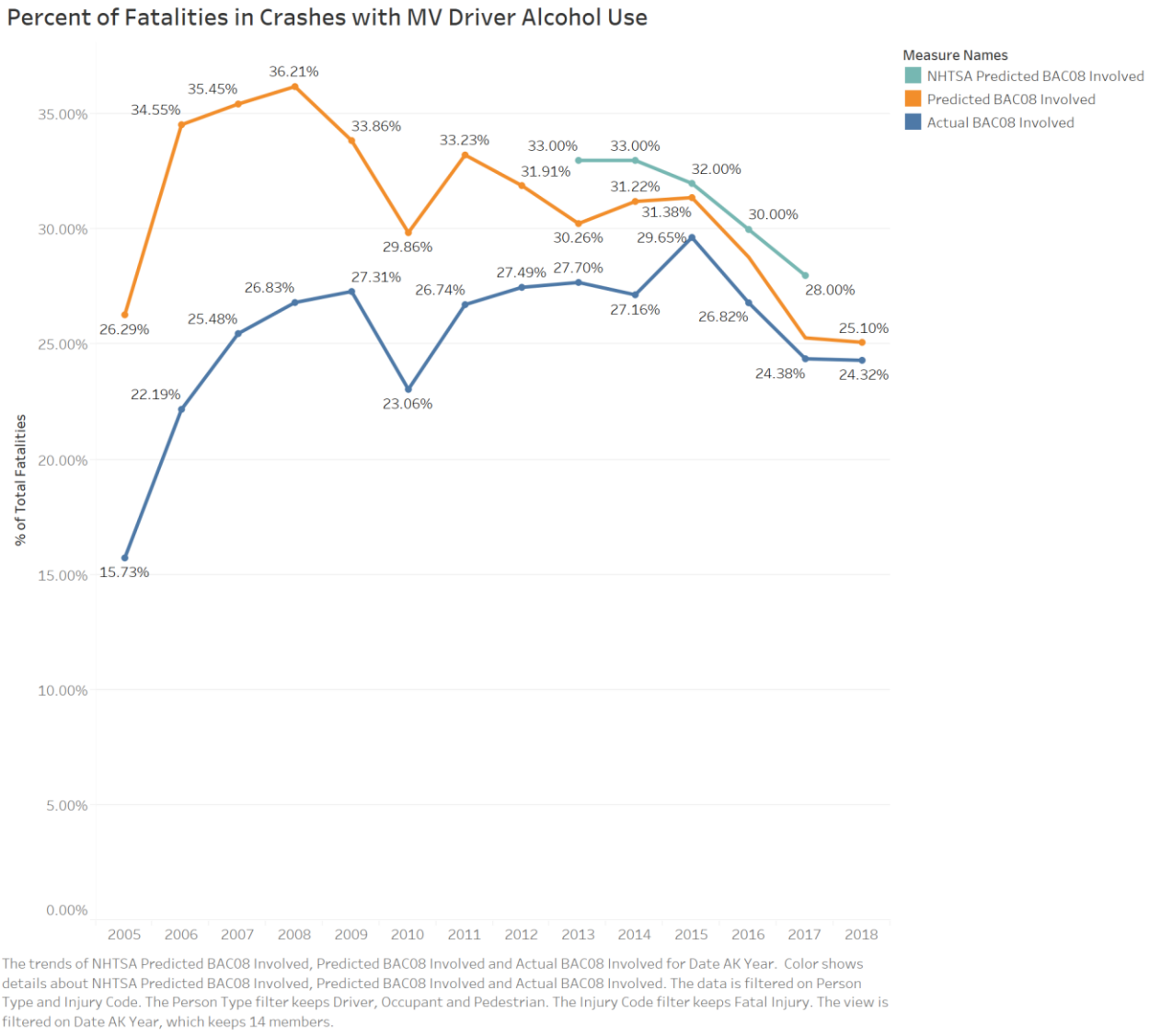

Figure 8 – Percent of Fatalities in Crashes with MV Driver Alcohol Use: Actual vs. Predicted

CONCLUSION

The preliminary results using 2018 crash data indicate the neural network model is performing as expected for aggregate prediction of motor vehicle driver alcohol-impairment in Louisiana. The modeling methods and features included with JMP Pro were extremely valuable in making several portions of the overall process much less time-consuming and much more efficient. As with any machine learning model planned for use in a production environment, ongoing evaluation of model performance is critical. In addition, more work needs to be done to uncover the structure of relationships between predictors within the model, as well as to identify other possibly useful predictors. In doing so, it may be possible to provide highway safety stakeholders with a more robust model and additional analytical tools to help them be more effective and/or efficient with their important efforts.

REFERENCES

- Overview of the Predictor Screening Platform. (2019, January 2). Retrieved April 15, 2019, from https://www.jmp.com/support/help/14-2/overview-of-the-predictor-screening-platform.shtml
- Shmueli, G., Bruce, P. C., Stephens, M. L., & Patel, N. R. (2017). Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro. Hoboken, New Jersey, United States: Wiley.
- Murphrey, W. (2014, September 30). Alternate Cut-off Confusion Matrix Add-In. Retrieved April 15, 2019, from https://community.jmp.com/t5/JMP-Add-Ins/Alternate-Cut-off-Confusion-Matrix-Add-In/ta-p/22357
- National Highway Traffic Safety Administration. (n.d.). Traffic Safety Facts Annual Report Tables. Retrieved April 15, 2019, from https://cdan.nhtsa.gov/STSI.htm

RECOMMENDED READING

Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro, ISBN 9781118877432

CONTACT INFORMATION

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

David Whitchurch
Louisiana State University Center for Analytics and Research in Transportation Safety
179 South Quad Drive
Baton Rouge, LA 70803
1.225.578.0366
dwhitc1@lsu.edu
http://carts.lsu.edu