MSIS 5633 Predictive Analytics Technologies

(Section 28453)

Term Project

Predicting Injury Severity with Risk Factors KNIME

Due Date

May 6, 2021

By:

Justine Gramling

Weston Parker

Joshua Parrack

Khyati Nema

Executive Summary

The goal that we set out to reach is one of the utmost importance; building predictive models to uncover the key underlying factors that lead to high injury severity and death in automotive accidents. All of the modeling that follows in this report was carried out in KNIME, with some additional table creation and data exploration done within Excel. The process that we followed throughout this report is the CRISP-DM model.

The data that was used in this report comes from the National Highway Traffic Safety Administration. It comes from their Crash Report Sampling System (2016-2019), and in particular we are using four data sets from that report; Vehicle, Accident, Person, and Distract. Initially, upon joining these four data files together, we had a total of 202 variables and 130,228 records. Significant work had to be done with the data before modeling could be carried out. Through exploration and looking at the surrounding literature we were able to filter the 202 variables down to 28 that would be used for the modeling. Through grouping/binning some of the variable responses and eliminating incomplete/unusable data we were able to get down to 15,939 records. We also chose to bin our target variable into Low Injury sustained and Severe Injury sustained as a part of this trimming down.

We were now ready to model and built seven different models for comparison. The models we used are: Decision Tree, Random Forest, Neural Network, Logistic Regression, Gradient Boosted Trees, Naive Bayes, and K-Nearest Neighbor. All of the models provided meaningful results, but we chose Gradient Boosted Trees as our most successful model due to its very high sensitivity along with a strong accuracy score. We also found meaning in exploring the splits of the decision tree to pull out some key underlying factors, such as wearing a seat belt.

This modeling was not carried out on behalf of any specific person or organization, but we believe that it still leads to deployable outputs. We do encourage further exploration and data analysis before making significant investments based on our findings, but this report lays the groundwork of areas that should be focused on both from the manufacturing side, as well as the public policy side. We are hopeful that this report, and big data and its analysis in general, can help mitigate the global public health issue of auto accidents that we are currently faced with and help shape a safer future.

Business Understanding

Car accidents and their related injuries and deaths have become a one of the most pressing public health concerns of the modern world. Over time the auto manufacturing industry has continued to strive for producing vehicles that are safer and safer for consumers to use to work towards minimizing this risk. Governments, at all levels, have also made a concerted effort to limit the health risks due to driving by restructuring the way that roads are both designed and maintained. On top of the physical changes, of both vehicles and roads, it is also important to take note of the increasing social focus that has been placed on driver safety with the idea that better dissemination of knowledge to drivers can also decrease the risks that they are faced with. These combined efforts have started to lead to incremental decreases of deaths and injuries in recent years,

> "For the second consecutive year, the U.S. experienced a small decline in roadway deaths... In 2019, an estimated 38,800 people lost their lives to car crashes – a 2% decline from 2018 (39,404 deaths) and a 4% decline from 2017 (40,231 deaths). About 4.4 million people were injured seriously enough to require medical attention in crashes last year – also a 2% decrease over 2018 figures," (National Safety Council).

While this does show the start of an encouraging trend, there is still a long way to go in reducing these numbers, especially considering that they still remain much higher than they were earlier in this decade. The way to continue decreasing these numbers is by leveraging the large data sets that are available with an increasing suite of tools that can be used to analyze them.

The underlying causes for traffic accidents and the severity of injuries caused by them are immensely complicated and very difficult for these groups to fully understand if they are only taking their piece of the picture into account. By being able to gain clarity on how these numerous factors are interwoven we can start to piece together the most important areas that need addressing to decrease the amount of accidents leading to severe injuries, be those changes technical, behavioral, infrastructural or otherwise. Our main goal of the rest of the report is to do just that. We are looking to build a variety of predictive models from the data that we have access to so that we can determine which factors lead to the greatest risk of injury severity, with the

ultimate goal of that knowledge being available to shape public and private action moving forward to decrease the risks faced by those on the road.

Data Understanding

The complete dataset is obtained from 4 SAS data files viz, Accident, Vehicle, Person and Distract which gives us the detailed car crashes that happened in the United States of America from the year 2016 till December 2020.

The accident file contains information related to crash, road conditions and environmental conditions. There are 54409 records and 50 columns. This file has data elements such as Number of Persons injured in crash, crash date and and time, weather and lighting condition, manner of collision, whether alcohol was involved in crash, location of car related to junction etc. Case Number is the primary key in this file.

The Vehicle file includes variables related to vehicles involved in the crash. This file has 96717records and 90 columns. It includes data elements such as Vehicle Body Type, Vehicle Model year, Number of people injured in Vehicle, Maximum injury severity in the Vehicle etc. In this file, there are 2 primary keys viz,Case Number and Vehicle Number.

The person file provided details related to age of the people involved in crash , seat position, injury severity and other similar data. Primary Keys are Case Number, Vehicle Number and Person Number. This file has 135410 records and 56 columns.

This Distraction file contains information about Driver's Distraction. It has 96751 records and 11 columns. In this file, Primary keys are Case Number and Vehicle Number.

The data of all the above 4 files were obtained either from the Police crash report or by interpreting the information provided in the crash report. The data is captured from the review of crash diagrams, Police Officer's written summary of the crash or with combinations of data elements on the report. We have excluded some records of data elements which were found missing or entered unknown or not reported as they were resulting in incomplete data for analysis.

Data Preparation

In order to read all the 4 SAS data files viz, Accident, Vehicle, Person and Distract, SAS7BDAT Reader node is used. Then the 4 SAS data files are merged into a single database using joiner nodes in Knime tool.

Accident and Vehicle file are Right Outer Joined on Case Number with the help of Joiner node. Then the merged data of Accident and Vehicle file are left outer joined with Person Data on Case Number and Vehicle Number with another Joiner Node. And again merged data of all the 3 files viz Accident, Vehicle and Person are left outer join with Distract data on Case Number and Vehicle Number. Final data after merging all the four files gave us 130228 records and 202 columns.

Column filter nodes are used to remove unwanted variables. We chose 28 important variables from 202 variables to predict target variable injury severity.

We chose variable SEAT_POS with its attribute value = 11 because we wanted to focus on Driver's Data. Driver factor is considered as important because it is major which contributes in car crashes and thus leading to injuries.

We used Rule-based row filter to filter unknown and not reported code from the variables such as HARM_EV, MAN_COLL, TYP_INT, LGT_COND, WEATHER, GVWR,BODY_TYP,MOD_YEAR, DEFORMED,TOWED, SPEEDREL, VSURCOND,P_CRASH2,AGE and AIR_BAG.

We categorized months into seasons such as Fall, Winter, Spring and Summer and days of the week into weekdays and weekends through Rule Engine Node in Knime.

Because of the many different types of Body Types, we decided to filter out the larger vehicles not typically associated with everyday driving. The vehicles filtered out were things like school buses, freight vans, and semi trucks. Essential commercial vehicles were filtered out. This left everything as small as a 2 door compact to a SUV to be included in our data.

Also, our target variable originally consisted of 10 different values. We needed to get this to a binary option. To do this we filtered out all values between 5 and 9. This left values 0 through 4 for our models. We chose 4 as our cutoff because values 0,1, and 2 were good descriptions of low severity injuries and values 3 and 4 were really the best and most reliable

indicators of a high severity injury. We did this using a rule engine node. Values 0,1, and 2 are labeled “low” and values 3 and 4 are labeled “high”.

Vehicle Age was also a variable we used for our models. This was a variable that had to be constructed. So, we took a math formula node and subtracted 2019 (the latest vehicle model in the data) from the model year of the vehicle involved in a crash (MOD_YEAR).

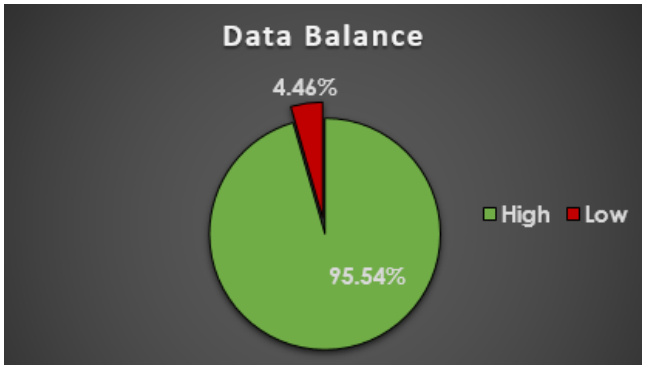
Once this was done, we used the column filter node to select the variables we wanted to keep for our models. In the end, we kept 28 variables.

After finalizing our variables, we ran these values through a normalizer node in order to get all the variables in the same scale and eliminate the problem of variables with higher nominal values displaying the most importance.

A color manager was utilized for a small bit of color and visual aspect added to the models. We used green to signify the low severity crashes and red to signify the high severity crashes.

Finally, within the data partitioning node, we seperated the data into 70% training and 30% validation data. We decided to stick to this rather “default” value because even with the multiple row filters that greatly reduced the amount of records in our data, we were left with plenty of data to properly assign into both training and validation datasets.

After all of this data prep, now came the difficult part where we had to make a decision on which direction to go. Because the data is imbalanced with 95.54% of values being “high” and 4.46% being “low”, we badly needed to decide which way to input these values into the models.



We really had two options: smote or equal size sampling. In the end, we chose equal size sampling because rather than oversampling the minority class many times, this node would randomly eliminate rows from the majority class until there was an equal number of "high" and "low" values going into the learning nodes. Also, frankly, when each of the other groups is likely to use the SMOTE node, why not lean the other direction?

S.No.	Variable	Variable Description
1	CASENUM	This data element is the unique case number assigned to each crash. It appears on each data file and was used to join all the 4 data files. It was removed for the model building purpose.
2	REGION	This data element identifies the region of the country where the crash occurred.
3	VE_TOTAL	This data element is the number of contact motor vehicles that the officer reported on the police crash report as a unit involved in the crash.
4	NUM_INJ	This data element records the number of persons injured in the crash and is derived by counting all persons with "Injury Severity" in the crash.
5	MONTH	This data element records the month in which the crash occurred.
6	DAY_WEEK	This data element records the day of the week on which the crash occurred.
7	HARM_EV	This data element describes the first injury or damage producing event of the crash.
8	MAN_COLL	This data element describes the orientation of two motor vehicles in-transport when they are involved in the "First Harmful Event" of a collision crash. If the "First Harmful Event" is not a collision between two motor vehicles in-transport, it is classified as such
9	TYP_INT	This data element identifies and allows separation of various intersection types.
10	LGT_COND	This data element records the type/level of light that existed at the time of the crash as indicated in the police crash report.
11	WEATHER	This data element records the prevailing atmospheric conditions that existed at the time of the crash as indicated in the police crash report.
12	RELJCT2_IM	This data element identifies the crash's location with respect to presence in or proximity to components typically in junction or interchange areas.
13	ALCHL_IM	This data element records alcohol use for drivers, pedestrians, cyclists and other types of non-motorists (except occupants of motor vehicles not in-transport) involved in the crash.
14	BODY_TYP	This data element identifies a classification of this vehicle based on its general body configuration, size, shape, doors, etc.
15	MOD_YEAR	This data element identifies the manufacturer's model year of this vehicle.
16	J_KNIFE	This data element identifies whether this vehicle experienced a jackknife anytime during the unstabilized situation.
17	GVWR	This data element identifies the gross vehicle weight rating of this vehicle if applicable.
18	DEFORMED	This data element records the amount of damage sustained by this vehicle as indicated on the police crash report based on an operational damage scale.
19	TOWED	This data element describes the mode by which this vehicle left the scene of the crash.
20	SPEEDREL	This data element records whether the driver's speed was related to the crash as indicated by law enforcement.
21	VSURCOND	This data element identifies the attribute that best represents the roadway surface condition prior to this vehicle's critical precrash event.
22	P_CRASH2	This data element identifies the attribute that best describes the critical event which made this crash imminent (i.e., something occurred which made the collision possible).
23	AGE	This data element identifies this person's age at the time of the crash, in years, with respect to their last birthday.
24	INJ_SEV	This data element describes the severity of the injury to this person in the crash using the KABCO scale.
25	SEAT_POS	This data element identifies the location of this person in or on the vehicle.
26	AIR_BAG	This data element records air bag availability and deployment for this person as reported in the police crash report.
27	SEX_IM	This data element identifies gender of the Person involved in the Crash
28	EJECT_IM	This data element describes the ejection status and the degree of ejection for this person, excluding motorcycle occupants.

Modeling

Once the data was explored and cleaned, it was then put into the seven different models which are decision tree, random forest, neural network, logistic regression, gradient boosted trees, naive bayes, and k nearest neighbor.

Starting with the **decision tree**, the levels are split and chosen based off of the most important variables in the data set. In the decision tree learner node, we used the gini index as the quality measure and set the tree to "no pruning". Figure E shows the first three levels of the tree and the variables the tree chose to root its decisions from. The first variable the tree chose is considered the most significant variable in the data set. This variable is the number of injured people. and whether or not it is less than or equal to 0.0455 or greater than 0.0455. This variable shows to be quite important as 100% of the crashes with less than .0455 number of injured result in a low crash severity. The next level of the tree and second most important variable is how deformed the vehicle was. If this variable was higher than 0.5556 then about 83% of wrecks resulted in high severity. The decision tree continues to base rules by the variables' level of significance, which creates more branches, therefore creating a more pure tree.

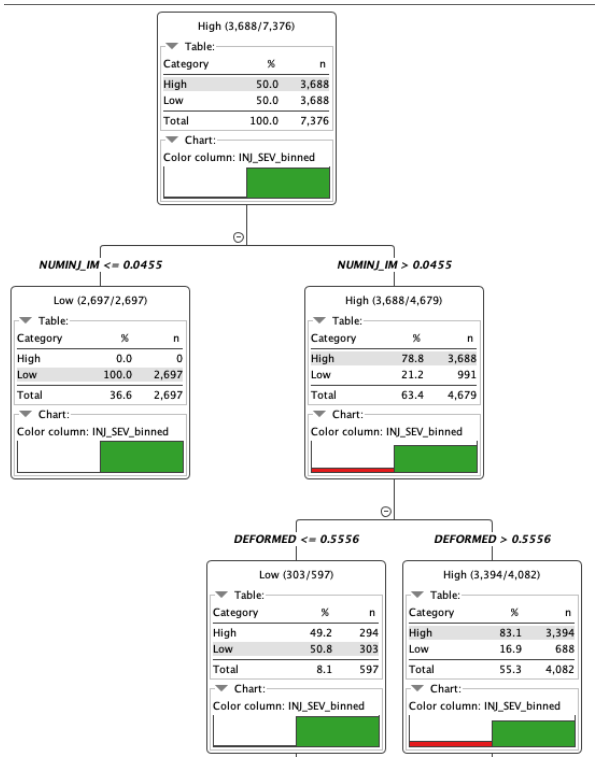


Figure E: First three levels of the decision tree. The first three levels display the top three important variables in the data. As the tree creates new branches, the purity increases, which is shown by the red and green distribution (red being high severity and green being low severity).

To determine the efficiency of the prediction, the ROC node and scorer nodes were connected to the model. This is also true for the remaining models. The results of the scorer node is later explained in the *Evaluation* section. The decision tree ROC value is 85.9%. This is a good value to begin comparing the other models to. Figure F displays the decision tree ROC curve.

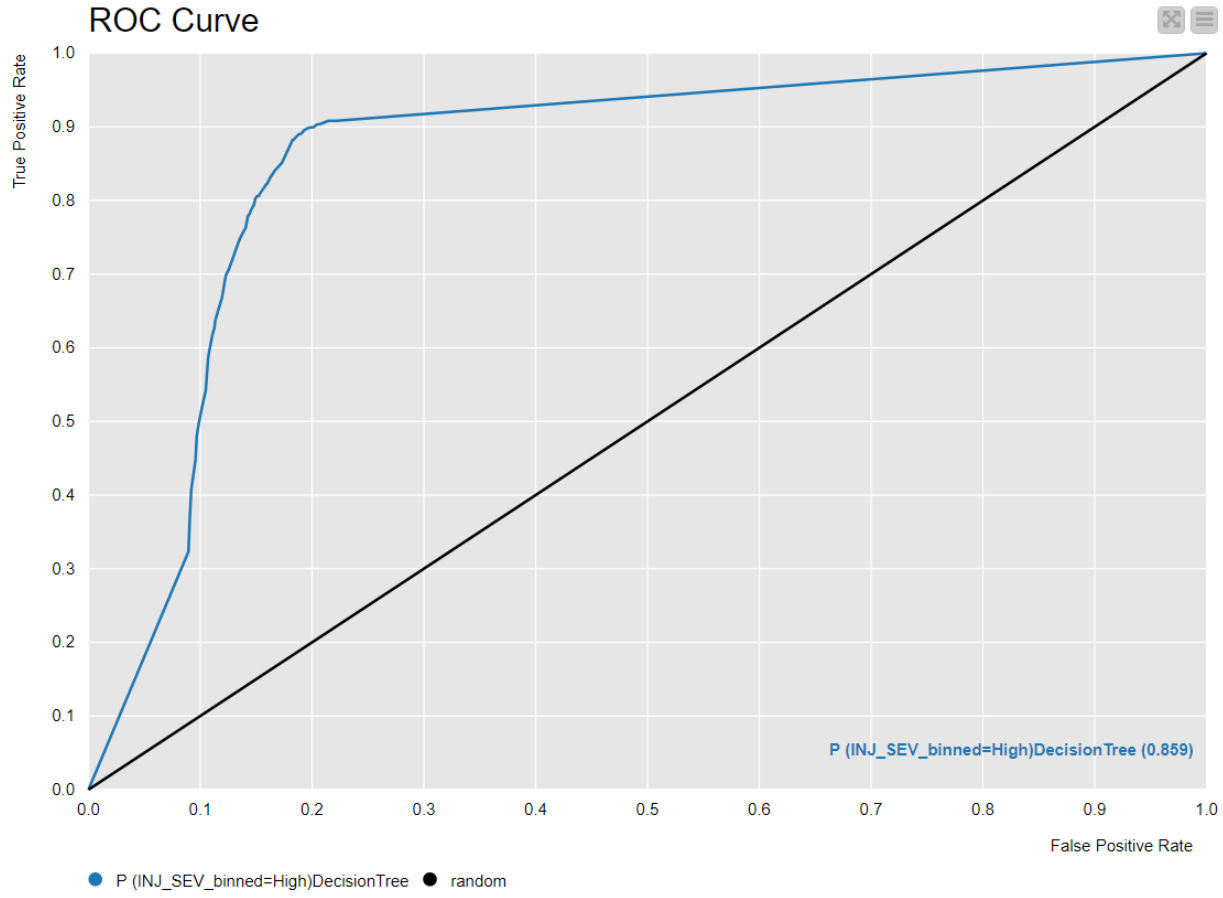


***Figure F**: ROC curve for decision tree. The ROC value is 85.9%.*

Next, a **random forest** was built in addition to the decision tree. The random forest algorithm uses an ensemble of decision trees and builds one, effective model with this. The random forest ROC value is 94.1%. This is a great enhancement from the previous model,

decision tree. This improvement comes from the algorithm's design of using many decision trees to make an overall, better optimized model. Figure G shows the curve and value of the random forest.

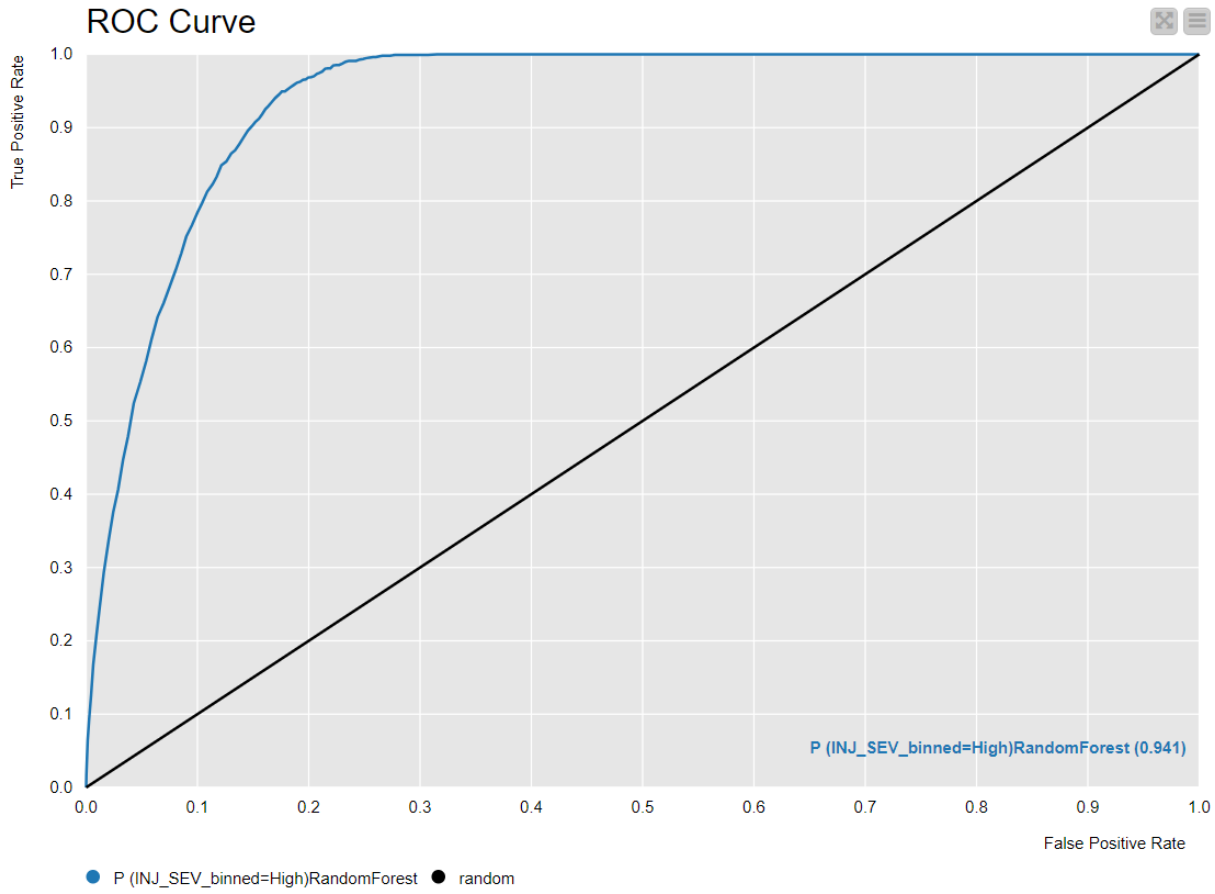


Figure G: *Random forest ROC curve. The ROC value is 94.1%.*

Following, a **neural network model** was built. The neural network model requires that the variables are all numeric, which means that use of a one to many node is included before modeling. The one to many node turns all categorical variables into numeric by using a 0 or 1. Doing so prepares the categorical input variables for the neural network. The neural network algorithm operates by taking the input variables and making links between them, which then link to a final predicting output. The ROC value for the neural network is 93.3%. This model is performing better than the decision tree, but worse than the random forest, in regard to the ROC metric value. Figure H shows these results.

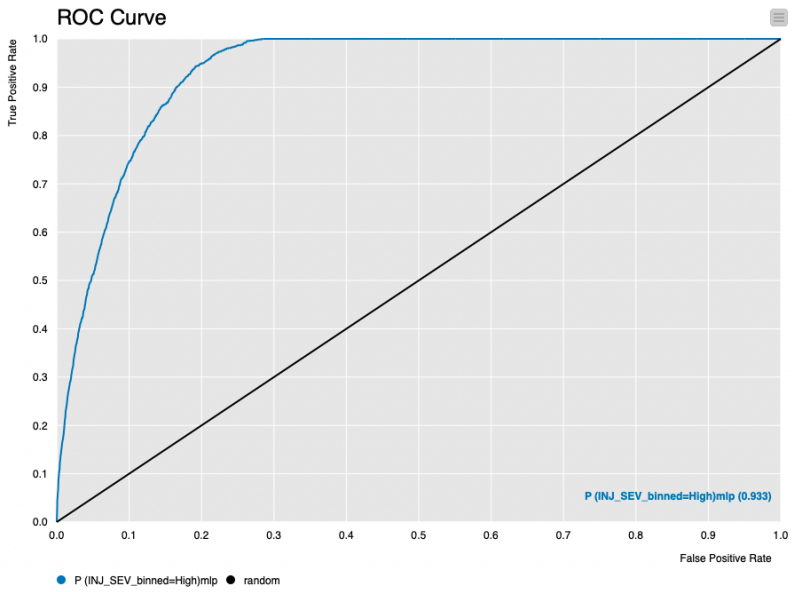


***Figure H**: The ROC value results for the neural network model. The ROC value is 93.3%.*

After, a **logistic regression** model was used. Logistic regression is considered to be one of the more basic models as it uses a mathematical function for explaining the input and target variables. The ROC value output for logistic regression is 90.5%. While this ROC value outperforms the decision tree, it is still worse than the random forest and neural network model. Figure I gives a visual of the curve.

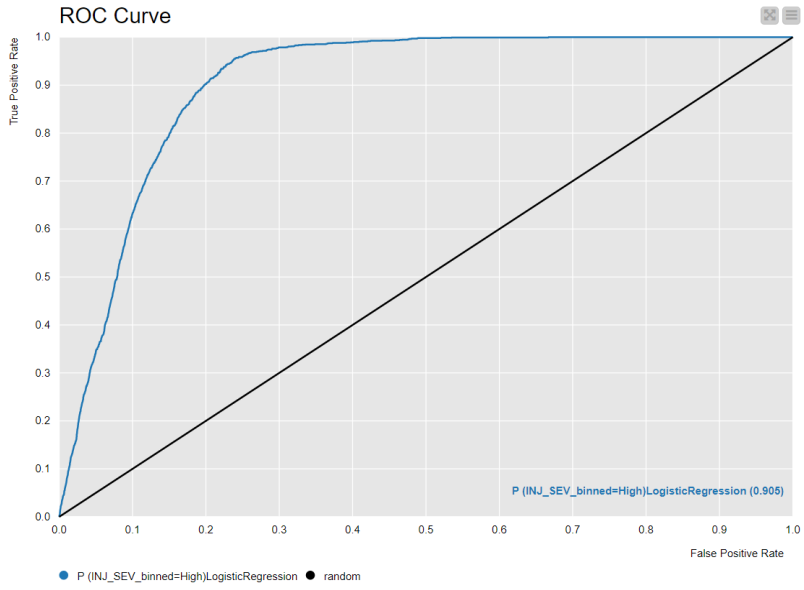


***Figure I**: Results of the Logistic Regression ROC value. The value is 90.5%.*

Then, a **gradient boosted trees model** was created. Gradient boosted trees is a machine learning model that uses foreknowledge that the next model, when combined with previous models, reduces prediction error. The model is optimized on reducing error. This model has several different ways the user can optimize it. We listed the limit on the number of levels at 4, the number of models produced by this node at 100, and the learning rate at .2. Setting the learning rate this low was to avoid overfitting as the equal size sampling node limited the number of records entering the learning nodes. The gradient boosted trees ROC value is 94.2%. This makes it the best model thus far in terms of ROC value. Figure J shows the ROC curve given for the model.

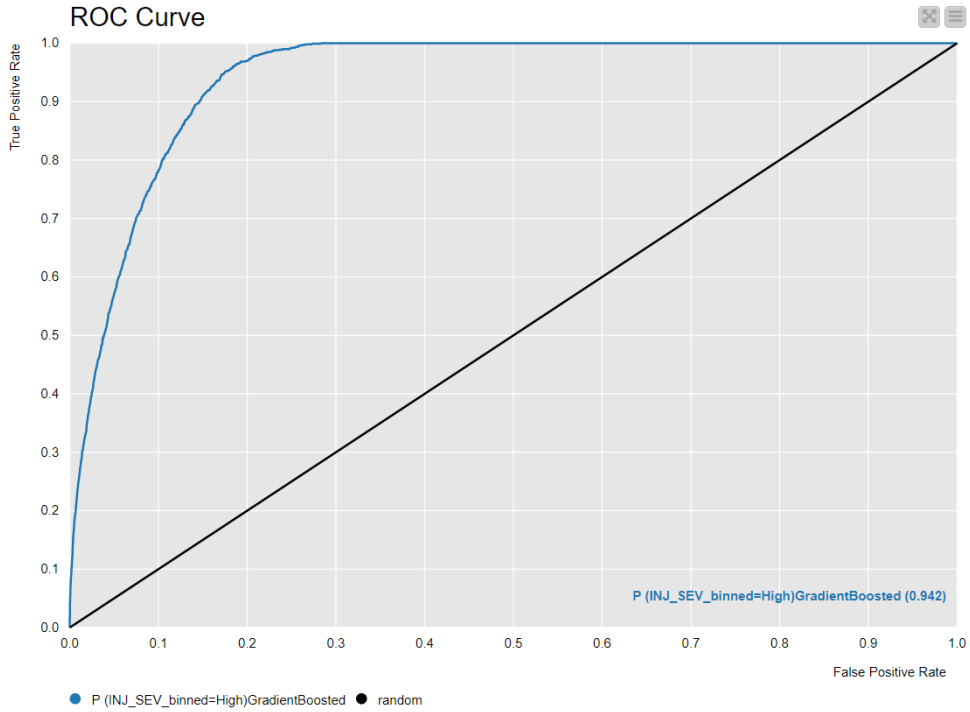


Figure J: *The results of the Gradient Boosted Trees ROC value. The value was 94.2%.*

The next model built was a **naive bayes**. The naive bayes' algorithm operates by predicting the probability of observations belonging to particular classes. The class that has the largest probability is deemed to be most likely. The results of the ROC curve for the naive bayes is 88.2%, which is performing better than the decision tree, but worse than the previous models. Figure K displays the results of this.

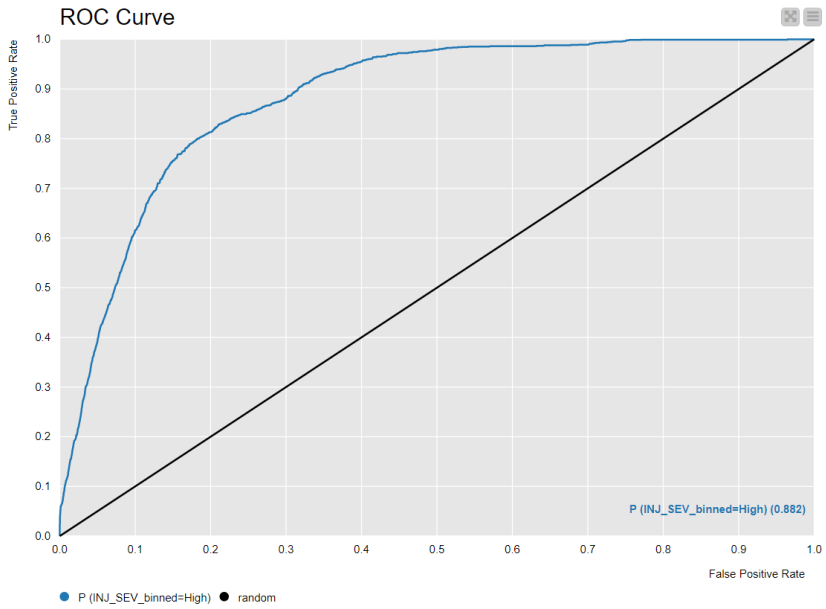


***Figure K**: The results of the Naive Bayes ROC value. The value was 88.2%.*

The final model created was a **k-nearest neighbor**. The k-nearest neighbor functions by classifying each observation based on similarity. The value of K is used to be the reference value that is looked at for the number of nearest neighbors to that observation. The results of the k-nearest neighbor values are a poor 73.4% for the accuracy rate and 77.2% for the sensitivity rate. This is not nearly good enough to qualify and is the worst producing model we have built.

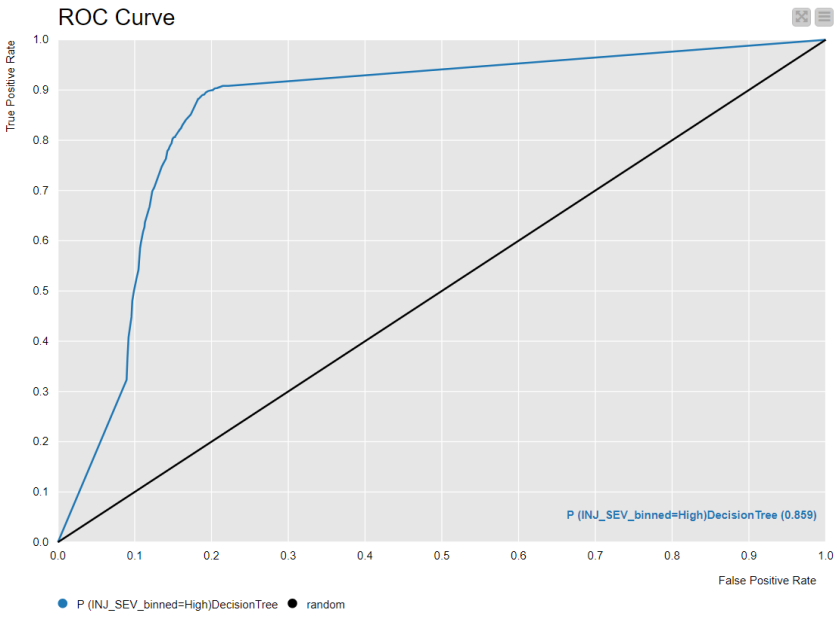


***Figure J**: The results of the Naive Bayes ROC value. The value was 88.2%.*

Lastly, a diagram of the KNIME workflow is shown to see the different nodes, connections, and models used to have a better understanding of the CRISP-DM process. This can be seen in Figure K.

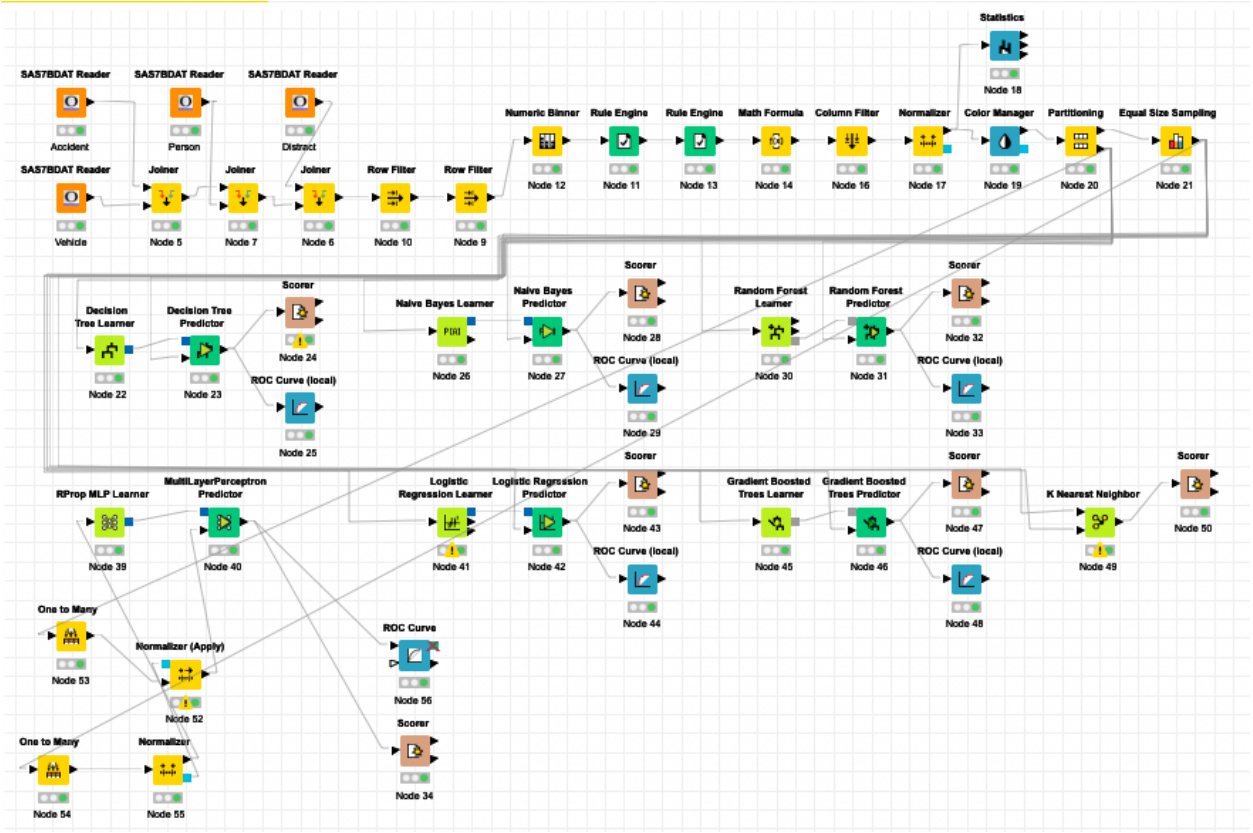


Figure K: *Overview of the KNIME workflow. The diagram displays the nodes, connections and models used.*

Evaluation

Once the models were built, we created a model comparison chart. This chart displays the metrics from the scorer and ROC curve nodes. That would be accuracy, specification, sensitivity, and the ROC Curve value. These metrics are compared to each other in order to determine the best type of model for our problem. Determining which metric is best for this model is heavily dependent on what exactly we are trying to solve. For instance, when building models to predict for diseases or other health related issues it is crucial to tune the model in a way where false negatives are mitigated. As a false negative is a horrible outcome for this test.

This scenario is similar to the problem we have been modeling. Because we separate the severity of the injuries into 2 seperate bins, the "high" severity value is something that we don't want to have false negatives when predicting. Because of this, we have decided the gradient boosted trees model is the best model for us to implement. We chose this because it has the highest sensitivity at 96.2% with only the random forest model coming close behind at 95%. Also, overall accuracy was right on par with the other models at 82% and this model had the largest area under the curve at 94.2%.

Because of the severe unbalance in the original data, it is difficult to come up with a model that is more accurate than guessing. However, this model does help in limiting the number of false negatives.

Model	Accuracy	Sensitivity	Specificity	ROC Value
Decision Tree	82.6%	85.4%	82.6%	85.9%
Random Forest	82.0%	95.8%	81.3%	94.1%
Logistic Regression	81.6%	88.4%	81.2%	90.5%
Gradient Boosted Trees	82.0%	96.2%	81.4%	94.2%
Naïve Bayes	87.5%	66.4%	88.6%	88.2%
K Nearest Neighbors	73.4%	77.2%	73.2%	85.9%
Multilayer Perceptron	79.8%	79.0%	95.9%	93.3%

Figure L: *Model comparison chart showing the performance metrics for each model. It highlights the best performing model for each metric.*

Deployment

When it comes to deployment of these models we think that additional work should be done before investing heavily in any of the following recommendations. There are two key reasons why we feel the need to state this; any changes capable of making a significant impact on the rate of severe injury and death in auto accidents is likely to be very expensive, and the stakes quite literally are life and death. We would suggest expanding the scope of the data that is being looked at to data collected in other countries as well, since driver behavior and manufacturing standards may vary widely based on location. In addition to this expansion to the work we think that further variable experimentation would be advisable before commiting sizable resources to confronting any of the key factors found in the model. With all that said, we

do still believe that the modeling done here can serve as a starting point to working on this public health crisis. The key way to deploy these models is to use them to pull out the key factors, such as seat belts etc, and use that to shape policy both within the manufacturing space as well as externally in public information campaigns as well as legal policies as well. One change that we would advise based on our findings is increasing the legal penalty for driving without a seatbelt/properly installed age appropriate restraint. Currently in most states you can only receive a nominal fine (in the realm of $25) for not wearing a seatbelt, and on top of that, many states only allow for a fine to be issued for not wearing a seatbelt as a secondary law, meaning that you have to be pulled over for a different violation entirely before being able to be punished for not wearing a seatbelt. We would suggest making it a national requirement that not wearing a seat belt be considered a primary law and carry a minimum fine of $150. We also suggest adding an excelating scale that after the third violation within a given time frame (say 2 years) the offense is upgraded to a reckless driving violation with the additional fines and potential jail time still included. Dramatically increasing the legal and financial punishment for not wearing a seatbelt while driving could go a long way in reducing the amount of serious injuries and deaths that occur in auto accidents on a yearly basis.

Conclusion

We began with data spreading across four different sources with over 200 variables in total. We were able to not only reduce this to 28 highly important variables, but we were able to create meaningful predictive models that could tell us if a crash was likely to have minimal or severe consequences for the passengers compared to our equal sampling. We are then able to look into the models and determine what factors they found to be most significant when predictive a severe outcome. By pulling out these key factors we are able to make suggestions that could potentially help reduce this public health crisis. We are also confident that even though all of the data that we are working with comes from accidents occuring in the United States, our findings can be applied more broadly and could potentially help guide policies abroad as well. However, we don’t think that this should be the end of this data exploration. Any meaningful change will require significant financial inputs or legal overhaul, so we heavily encourage greater depth of study, as well as running data from other countries through the models, before fully moving on to the implementation stage of the process.

When sifting and filtering through the data and pulling variables in and out of the modeling, it is easy to lose perspective inside the intricacy of the numbers. Every record and every individual data point also represents worry, stress, and potentially heartbreak for those close to the passengers. We aren't building models for the sake of building models, and while it is critical to be able to distance yourself from the work while you are doing it, so that we may mitigate the personal bias that we are building into the models, it is also a must to not lose the human side of the equation as well. We are building these models to help improve the world we live in and decrease this public health crisis that we are faced with. We are excited and encouraged by the potential for change that big data and its analysis carries, but we recognize that there is still a long way to go in building public trust in this type of modeling to allow for better policy creation and implementation.