**Abstract**

Road crashes each year result in thousands of lives lost and injured victims, and billions of dollars in property damage in the USA. Accurate data are required to support the development, implementation, and assessment of highway safety programs aimed at reducing this toll. The analysis of highway-crash data has long been used as a basis for determining vehicle and highway designs. It also has great influences on directing and implementing a wide variety of regulatory policies aimed at improving safety. With the improvement in statistical methodologies, researchers are able to extract more information from crash databases to guide a wide array of safety design and policy improvements. In this project, 2018 Crash Report Sampling System (CRSS) data were used which contains all types of police-reported crashes, ranging from property-damage-only crashes to those that result in fatalities. The ACCIDENT data file containing crash data and PERSON data file including motorist and non-motorist data in CRSS were mainly analyzed to find the correlation between injury information and several representative variables such as age, sex, alcohol involvement, manner of collision, ejection status, day of the week, light condition and so on. To start with, basis statistic information of crashes was obtained by ACCIDENT Data File. Then we built models to predict the maximum injury severity in crashes on the basis of ACCIDENT Data File, which can be a criterion of how dangerous a crash is. Finally, models predicting the injury severity of a driver or passenger during the crash are created based on the information of PERSON Data File.

**Introduction and Background**

More than 1.2 million people die every year in highway-related crashes and almost 50 million people are injured all over the world1. It is reported that the highway-related crash is the 5th leading causes of the death worldwide2. Apart from that, highway-related crashes result in billions of dollars in property damage1. To provide vital information on motor vehicle traffic crashes, the National Highway Traffic Safety Administration (NHTSA) annually publishes nationally representative estimates of police-reported motor vehicle traffic crashes and their characteristics. From 1988 to 2015, NHTSA created national estimates using data from the National Automotive Sampling System General Estimates System (NASS GES), which sampled police crash reports from police jurisdictions across the United States. In 2016 NHTSA replaced NASS GES with the Crash Report Sampling System (CRSS), which is a sample of police-reported crashes involving all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that resulting in fatalities. CRSS is used to estimate the overall crash details, identify highway safety problem areas, measure trends, drive consumer information initiatives, and form the basis for cost and benefit analyses of highway safety initiatives and regulations3. The CRSS obtains its data from a nationally representative probability sample selected from the more than six million police-reported crashes which occur annually. Although various sources suggest that there are many more crashes that are not reported to the police, the majority of these unreported crashes involve only minor property damage and no significant personal injury or fatalities. By restricting attention to police-reported crashes, the CRSS concentrates on those crashes of greatest concern to the highway safety community and the general public. The datasets we are going to use are from CRSS 2018 data repository.

Several popular injury severity model structures have been studied in the extant literature. Sequential binary logit models4, ordered-response probit models5 and multinomial or nested logit models6 are typical examples. The sequential binary logit and ordered-response probit models represent the ordinality in the discrete categories of the injury severity. Sequential binary logit models assume that the factors determining the level of the severity change according the level of the severity itself, while ordered-response probit models assume that the same factors are correlated with all levels of injury severity. In multinomial and nested logit models, ordinality is not theoretically implemented, thus information relating to ordering of severities is not inherently captured in those structures. Compared with the previous two models, multinomial and nested logit models are structurally flexible in the sense that independent variables are not forced to be the same across all severities.

In this project, crash statistic information was analyzed, such as which region has the highest number of crashes, which hour of the day has the highest number of injuries and so on. Then we built models to predict the maximum injury severity in a crash and injury severity of a driver or a passenger during the crash by logistic regression, neural network and random forest classifier method.

**Predictive Problem and Approach**

The maximum injury severity is of great importance to transportation analysis. We select some representative variables that have strong correlation with injury severity, and then logistic regression, neural network and random forest classifier are applied to generate models to predict maximum injury severity of a certain crash and injury severity of a driver or a passenger during the crash. To improve the precision of our models, pipeline is used to determine the best parameters for the random forest classifier method and ROC curve to evaluate our model.

The most difficult part is that maximum injury severity has five categories, including no apparent injury, possible injury, suspected minor injury, suspected serious injury and fatal. It means that models predicting multiclass variables should be built which is different from the binary variables models we learned in class.

**Methodology**

**Data Collection**

In the CRSS 2018 data, a total of 48,443 representative crashes are selected from over six million police-reported crashes that occurred in 2018, involving 86,105 vehicles and 120,230 people. The following two table show the representative variables considered in the maximum injury severity models from ACCIDENT file and injury severity of a driver or a passenger during the crash from PERSON file respectively.

Table 1 representative variables of ACCIDENT file

|  |  |  |
| --- | --- | --- |
| 1. Maximum severity injury | 6. Alcohol involvement | 11. Number of Persons in Motor Vehicles In-Transport |
| 1. Hour | 7. Manner of collision | 12. Number of Persons Not in Motor Vehicles |
| 1. Day of the week | 8. Location of crash (whether in junctions or interchange areas) | 13. Number of vehicles in the crash |
| 1. Light condition | 9. Location of crashes (type of junctions or interchange areas) |  |
| 1. Weather | 10. First harmful event |

Table 2 representative variables of PERSON file

|  |  |  |
| --- | --- | --- |
| 1. Maximum severity injury | 6. Alcohol involvement | 11. Sex |
| 1. Hour | 7. Manner of collision | 12. Age |
| 1. Day of the week | 8. Location of crash (whether in junctions or interchange areas) | 13. Ejection status |
| 1. Light condition | 9. Location of crashes (type of junctions or interchange areas) | 14. Seating position |
| 1. Weather | 10. First harmful event |  |

Weight function is a mathematical device used when performing a sum, integral, or percentage to give some elements more influence on the result than other elements in the same set. In the pre-processing of obtaining basis statistic information of crashes, weight is added when analyzing distributions among crashes and maximum injury severity affected by hour, day, region, weather, manner of collision and alcohol involvement to ensure unbiased and robust estimate. To build models, we transform the hour of the day into two categories, where daytime is from 6am to 8 pm and nighttime is from 8pm to 6am, and day of the week is also divided into the weekday and weekend. The rest of categorical variables are all transformed into dummy binary variables. For the first harmful event related to the maximum injury severity, there are four primary categories and many secondary categories, so we choose fifteen categorical first harmful events whose frequencies are larger than 0.007, including motor vehicle in-transport, parked motor vehicle, live animal and so on. There are totally 63 and 88 variables included in those two models respectively. We apply logistic regression, random forest classifier and neural network methods to build models.

**Data Analysis Theory**

**Model predicting maximum severity injury**

In this model, data are from ACCDIENT data file which contains information of each crash. we use Logistic regression with solver = 'lbfgs' and solver = 'newton-cg', neural networks, random forest classifier, pipeline to determine the best parameters for the random forest and ROC curve to evaluate the model.

Logistic regression is a classification algorithm which assigns observations to a discrete set of classes of an output variable. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to a discrete variable7. First, we try logistic regression with solver = ‘lbfgs’ and set multi\_class = ‘multinomial’ so we can predict multiclass variable.

Table 3 Logistic Regression (solver = 'lbfgs') Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 71.7 | | | | |
|  | Precision | recall | F1-score | support |
| 0 (no apparent injury) | 0.72 | 1.00 | 0.84 | 41016 |
| 1 (possible injury) | 0.00 | 0.00 | 0.00 | 9422 |
| 2 (suspected minor injury) | 0.00 | 0.00 | 0.00 | 5061 |
| 3 (suspected serious injury) | 0.00 | 0.00 | 0.00 | 1417 |
| 4 (fatal) | 0.00 | 0.00 | 0.00 | 260 |
| accuracy |  |  | 0.72 | 57176 |
| Macro avg | 0.14 | 0.20 | 0.17 | 57176 |
| Weighted avg | 0.51 | 0.72 | 0.60 | 57176 |

The precision is the ratio true positives / (true positives+ false positives), which means the ability of the classifier not to label as positive a sample that is negative. The recall is the ratio true positives/ (true positives+ false negatives), which means the ability of the classifier to find all the positive samples. The F1 -score is a weighted harmonic mean of the precision and recall, where an F1-score reaches its best value at 1 and worst score at 0. The precision of the model is 71.7, but the results show the precision and F1-score are ill-defined and being set to 0.0 in labels with no predicted samples. Then we try logistic regression with solver = 'newton-cg'.

Table 4 Logistic Regression (solver = 'newton-cg') Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 72.5 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.74 | 0.99 | 0.85 | 41016 |
| 1 (possible injury) | 0.40 | 0.05 | 0.08 | 9422 |
| 2 (suspected minor injury) | 0.37 | 0.05 | 0.09 | 5061 |
| 3 (suspected serious injury) | 0.00 | 0.00 | 0.00 | 1417 |
| 4 (fatal) | 0.19 | 0.01 | 0.02 | 260 |
| accuracy |  |  | 0.73 | 57176 |
| Macro avg | 0.34 | 0.22 | 0.21 | 57176 |
| Weighted avg | 0.63 | 0.73 | 0.63 | 57176 |

Though the algorithm doesn't converge at 10000 iteration, we now have predicted values of possible injury, suspected minor injury, suspected serious injury and fatal. The precision is still very low at 72.5. So, we try Neural networks.

Neural networks are created by adding the layers of several perceptrons together, known as a multi-layer perceptron model. There are three layers of a neural network - the input, hidden, and output layers. The input layer directly receives the data, whereas the output layer creates the required output (training and testing data). The layers in between are known as hidden layerswhere the intermediate computation takes place8. Neural networks are useful tool for both classification and regression problems.

Table 4 Neural networks Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 71.9 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.72 | 1.00 | 0.84 | 41016 |
| 1 (possible injury) | 0.91 | 0.01 | 0.01 | 9422 |
| 2 (suspected minor injury) | 1.00 | 0.01 | 0.01 | 5061 |
| 3 (suspected serious injury) | 0.62 | 0.00 | 0.01 | 1417 |
| 4 (fatal) | 0.00 | 0.00 | 0.00 | 260 |
| accuracy |  |  | 0.72 | 57176 |
| Macro avg | 0.65 | 0.20 | 0.17 | 57176 |
| Weighted avg | 0.77 | 0.72 | 0.60 | 57176 |

The precision is still low at 71.9. From recall in the table, we can see that when maximum injury severity is possible injury, suspected minor injury, suspected serious injury and fatal

, the recall is almost 0. Thus, maximum injury severity cannot be correctly predicted by neural networks.

Random forest classifier is an excellent classifier for predicting multi-class dependent variable. It is a meta estimator that fits a number of decision tree classifiers on a variety of subsets of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The size of subsets is always the same as the original input sample size. Then we try random forest classifiers with default settings.

Table 5 Random forest classifier (default setting) Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 96.5 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.98 | 1.00 | 0.99 | 41016 |
| 1 (possible injury) | 0.95 | 0.93 | 0.94 | 9422 |
| 2 (suspected minor injury) | 0.93 | 0.90 | 0.92 | 5061 |
| 3 (suspected serious injury) | 0.88 | 0.62 | 0.73 | 1417 |
| 4 (fatal) | 0.88 | 0.61 | 0.72 | 260 |
| accuracy |  |  | 0.97 | 57176 |
| Macro avg | 0.92 | 0.81 | 0.86 | 57176 |
| Weighted avg | 0.96 | 0.97 | 0.96 | 57176 |

The table shows that random forest classifier does much better than previous models. Precision jumps from 72 to 96.5. We can see that for recall, all no apparent injury and more than 90% possible injury and suspected minor injury cases are correctly predicted, and more than 60% of suspected serious injury and fatal cases are correctly predicted.

Now we change some settings to see if the precision can be improved, setting n\_estimators = 20, max\_features = 10.

Table 5 Random forest classifier (n\_estimators = 20, max\_features = 10), Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 97.3 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.98 | 1.00 | 0.99 | 41016 |
| 1 (possible injury) | 0.96 | 0.94 | 0.95 | 9422 |
| 2 (suspected minor injury) | 0.94 | 0.91 | 0.92 | 5061 |
| 3 (suspected serious injury) | 0.88 | 0.65 | 0.75 | 1417 |
| 4 (fatal) | 0.87 | 0.63 | 0.73 | 260 |
| accuracy |  |  | 0.97 | 57176 |
| Macro avg | 0.92 | 0.82 | 0.87 | 57176 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 57176 |

We can see that the recall of possible injury, suspected minor injury, suspected serious injury and fatal increase slightly by 0.01. Now we use pipeline to help us determine the parameters of random forest classifier and use cross-validation to avoid over-fitting. We set feature\_selection\_k = [1,6,11,16,21,26,31,36,41,46,51,56,61], n\_estimators = [80, 90, 100] and min\_samples\_split = [2,4]. With the combination of these three parameters, there are totally 78 candidates. It takes almost one hour to run. Rank\_text\_score shows that when k=46, n\_estimators=90, min\_samples\_split=2, we have the highest test\_score.

Table 5 Random forest classifier (feature\_selection\_k = 46, n\_estimators = 100, and min\_samples\_split = 2), Precision, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 97.3 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.98 | 1.00 | 0.99 | 41016 |
| 1 (possible injury) | 0.96 | 0.95 | 0.95 | 9422 |
| 2 (suspected minor injury) | 0.94 | 0.92 | 0.93 | 5061 |
| 3 (suspected serious injury) | 0.89 | 0.67 | 0.76 | 1417 |
| 4 (fatal) | 0.85 | 0.63 | 0.73 | 260 |
| accuracy |  |  | 0.97 | 57176 |
| Macro avg | 0.93 | 0.83 | 0.87 | 57176 |
| Weighted avg | 0.97 | 0.97 | 0.97 | 57176 |

The precision increases again with the average score of 3 cross-validation reaching 97.3, so we can say we have a pretty good model to predict the maximum injury severity.

Finally, ROC curve is applied to evaluate the model. ROC curve is a performance measurement for classification problem, created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It tells how much model is capable of distinguishing between classes.  The area under the ROC curve is called AUC.

A close up of a map

Description automatically generated

Figure 1 ROC curve of Random forest classifier (feature\_selection\_k = 46, n\_estimators = 100, and min\_samples\_split = 2)

From ROC curves we see that AUC of class 0,1,2,3 and 4 are 0.98, 0.97, 0.96, 0.83, 0.82 respectively, and the AUC of the macro-average ROC curve is 0.91, which means our model is very successful.

**Model predicting injury severity of a driver or passenger**

In this model, data are from the PERSON data file which contains information of driver and passages during the crash. We use logistic regression with solver = 'newton-cg', neural networks, random forest classifier to build models and ROC curve to evaluate the model.

Table 4 Logistic Regression (solver = 'newton-cg') Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 84.1 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.84 | 1.00 | 0.91 | 116454 |
| 1 (possible injury) | 0.20 | 0.00 | 0.00 | 14284 |
| 2 (suspected minor injury) | 0.39 | 0.04 | 0.07 | 6322 |
| 3 (suspected serious injury) | 0.43 | 0.05 | 0.08 | 1467 |
| 4 (fatal) | 0.42 | 0.04 | 0.08 | 223 |
| accuracy |  |  | 0.84 | 138750 |
| Macro avg | 0.46 | 0.23 | 0.23 | 138750 |
| Weighted avg | 0.75 | 0.84 | 0.77 | 138750 |

From the table, it is clear that the total precision is 84.1, but the recall for possible injury, suspected minor injury, suspected serious injury and fatal are quite low at near 0.00. Then we try Neural networks.

Table 4 Neural networks Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 83.9 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.84 | 1.00 | 0.91 | 116454 |
| 1 (possible injury) | 0.71 | 0.00 | 0.00 | 14284 |
| 2 (suspected minor injury) | 1.00 | 0.00 | 0.00 | 6322 |
| 3 (suspected serious injury) | 0.00 | 0.00 | 0.00 | 1467 |
| 4 (fatal) | 0.00 | 0.00 | 0.00 | 223 |
| accuracy |  |  | 0.84 | 138750 |
| Macro avg | 0.51 | 0.23 | 0.18 | 138750 |
| Weighted avg | 0.82 | 0.84 | 0.77 | 138750 |

From the table, we can see that the total precision is 83.9, even lower than previous logistic regression model. The recall for possible injury, suspected minor injury, suspected serious injury and fatal is 0.00. Finally, we try random forest classifier.

Table 5 Random forest classifier (default setting) Precision, Recall, F1 -score and Support

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Precision: 97.6 | | | | |
|  | Precision | recall | F1 -score | support |
| 0 (no apparent injury) | 0.98 | 1.00 | 0.99 | 116454 |
| 1 (possible injury) | 0.96 | 0.89 | 0.92 | 14284 |
| 2 (suspected minor injury) | 0.94 | 0.89 | 0.91 | 6322 |
| 3 (suspected serious injury) | 0.87 | 0.65 | 0.74 | 1467 |
| 4 (fatal) | 0.81 | 0.61 | 0.69 | 223 |
| accuracy |  |  | 0.98 | 138750 |
| Macro avg | 0.91 | 0.81 | 0.85 | 138750 |
| Weighted avg | 0.98 | 0.98 | 0.97 | 138750 |

From the table, we can see that the total precision increases significantly at 97.6, much higher than logistic regression and natural network. The recall for possible injury, suspected minor injury, suspected serious injury and fatal is 0.89, 0.89, 0.65 and 0.61 respectively, meaning that the model predicts the level of injury severity of a driver or passenger well.

A close up of a map

Description automatically generated

Figure 1 ROC curve of Random forest classifier

From the ROC curve, we see that AUC of class 0,1,2,3 and 4 are 0.95, 0.94, 0.94, 0.82, 0.80 respectively, and the AUC of the macro-average ROC curve is 0.89. The AUCs are close to 1 meaning that our model has good precisions.

**Proposed Presentation and Utilization Plan**

Maximum severity injury has a strong correlation with weather and light condition. When the weather is dark or foggy, the crashes have more influence on injury severity. For drivers, they need to make the vehicle visible to others both ahead of you and behind you by using low-beam headlights. The department of transportation should be able to provide accurate warning messages to drivers in advance. In addition, crashes happen between 7pm and 5am are more likely to be involved with alcohol, and even more than 1/4 of the crashes happens at 2am are involved with alcohol, so drivers should not drink and drive. Front to front collision has higher injury severity in accidents, so drivers and passengers need to fasten their seat belt and keep distance between the car in front of them.

**Potential Problems and Fall-Back Strategies**

One potential problem is that the CRSS 2018 repository is lack of crashes with suspected serious injury and fatal. Compared with 41016 cases of no apparent injury for model predicting maximum severity injury, there are only 1417 and 260 cases of suspected serious injury and fatal. For model predicting the level of severity injury of driver or passenger, there are 1467 and 223 cases of suspected serious injury and fatal, while there are 116454 cases of no apparent injury. This is why the highest recall value of suspected serious injury and fatal is about 0.65 which far from what no apparent injury behaves at 1.00 of our random forest classifier models. The other problem is that there are some unknown and unreported data from the CRSS 2018 repository. We use imputed data which are createdfrom the original data elements having each unknown or not reported value substituted by the estimated value to ensure enough data for models.

**Conclusion**

In our models to predict the maximum injury severity in crashes, random forest classifier with pipeline determining parameters is the most successive algorithm with precision 97.3, but it takes more than one hour to run the codes. For the model to predict the level of injury severity of a driver or passenger, random forest classifier also behaves with precision 97.6. The reason why random forest classifier has the highest precision in both models is that random forest classifier can emphasize feature selection and utilize ensemble learning, meaning that random forest takes random samples, forms many decision trees, and then averages out the leaf nodes to get a clearer model. Although random forest classifier has such advantages, logistic regression is more timesaving.

**Reference**

1. World Health Organization. Dept. of Violence, Injury Prevention, World Health Organization. Violence, Injury Prevention, & World Health Organization. (2009). *Global status report on road safety: time for action*. World Health Organization.
2. World Health Organization. (2013). *Global status report on road safety 2013: supporting a decade of action: summary* (No. WHO. NMH. VIP 13.01). World Health Organization.
3. National Center for Analysis and Statistics. (2019, November). *Crash Report Sampling System CRSS analytical user's manual 2016-2018* (Report No. DOT HS 812 846). National Highway Traffic Safety Administration.
4. Saccomanno, F. F., Nassar, S. A., & Shortreed, J. H. (1996). Reliability of statistical road accident injury severity models. *Transportation Research Record*, *1542*(1), 14-23.
5. O'donnell, C. J., & Connor, D. H. (1996). Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. *Accident Analysis & Prevention*, *28*(6), 739-753.
6. Shankar, V., & Mannering, F. (1996). An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. *Journal of safety research*, *27*(3), 183-194.
7. Ayush Pant. (2012, January). Introduction to logistic regression. <https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148>
8. Deepika Singh. (2019, June). Machine learning with neural networks using scikit-learn. <https://www.pluralsight.com/guides/machine-learning-neural-networks-scikit-learn>