Northeastern University

ALY6020: Predictive Analytics

Assignment 5: Crash Detection

Abhigna Ramamurthy, 002982276

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Feb 11, 2023

Introduction

This report helps shed light on the main issues faced during the bike crashes in the city of Austin. The city of Austin has been facing a growing concern over the safety of cyclists on its roads. With an increasing number of people taking up cycling as a means of transportation and recreation, the need to ensure their safety has become more pressing. Several complaints have been received from cyclists about the lack of measures taken by the city to protect them from motor vehicles. To address this issue, the city has collected data on various incidents involving cyclists. The aim of this analysis is to review the findings from this data to determine if the complaints made by cyclists are indeed accurate, and to understand the extent of the problem. With this information, the city can take appropriate measures to ensure the safety of its cyclists and make Austin a safer place for all road users.

Data cleaning

For this analysis, the data was presented in an excel format which was imported into Jupyter Notebook python where the analysis was conducted. Main libraries used include pandas, NumPy, and datetime for data cleaning and manipulation, seaborn, and plotly, for visualization, and statsmodels, and sklearn for data modelling. The dataset has crash data from 2010 to 2017, which provides comprehensive data to analyze the accidents. The initial data had 2463 records and 20 attributes with 4 integer columns, 1 boolean value column, and rest are categorical or object data type columns.

The first step in cleaning this data was to check for missing values and no missing data was found. The next step was to conduct feature engineering by inspecting each attribute of the dataset. 85% of “Average Daily Traffic Amount” data is "No Data" hence this column can be eliminated. Speed Limit has invalid values of -1 and 0 which needs to be rectified by replacing those values with the mean of the column without 0, and -1 values. The mean value of speed limit without 0 and -1 values is 35.52, rounding down to 35, 35v was filled instead of 0, and -1 values. “$1000 Damage to Any One Person\'s Property”, “Active School Zone Flag”, and “Construction Zone Flag”, were converted to Boolean values of 1 if “Yes” and 0 if “No”. More than 80% of the Highway System data has value of "No Data". Hence dropping this column as it would not provide any analytical advantage.

For the attribute “Intersection Related” there was clear distinction in dataset for intersection crashes, and others. Hence the attribute was converted to Boolean values of 1 if the record value is “Intersection” and “Intersection Related”, and others were given a value of 0. Deleted “At Intersection Flag” as Intersection Related column provides more data as compared. Majority of the crashes also happened in daylight hence the attribute “Light Condition” was converted to Boolean value of 1 for Daylight and 0 for all others. Similarly, “Road Class” data was converted to 1 for City Street and 0 for all other values. 87% of the roadway part for the bike crashes happened in main/proper lane. Hence eliminated this attribute. More than 90% of the bike crashes happened in dry surface condition. As this could pose an imbalance in the dataset, I decided to delete this categorical attribute.

Due to highly divergent data of traffic control type, I decided to eliminate this attribute. On the other hand, Weather Condition attribute shows unified data of more than 80% attributed to clear, hence got rid of this attribute. About 55% of the times the person helmet was not worn. Hence keeping the attribute but converting to Boolean value to show 0 for not worn and 1 for other situations. Day of week was deleted as it did not hold analytical value as day of the crash did not help as a fact for the crash.

Finally, defining crash severity or the dependent variable for the model analysis in the next sections. 10% of data is incapacitating injury which is any injury, other than a fatal injury, which prevents the injured person from walking, driving or normally continuing the activities the person could perform before the injury occurred. Around 6% of data is killed. Both of which pose high severity, hence grouped together as 2 indicating high severity. Possible injury accounts for 22% of the data, which poses medium severity. This is indicated as 1 severity. 6% of data is non incapacitating and 7% of data is not injured, both of which posse less severity, hence grouped as 0.

Exploratory Data Analysis

Exploratory data analysis is a visual way to uncovering hidden patterns that can help the government officials to make better decisions knowing the pattern of the bike crashes. Count plot of all the categorical data in the dataset in Figure 1, Appendix shows interesting insights about the data and convers hidden patterns of data. Maximum number of accidents occurred in intersections on main/proper roads with dry surface conditions, more often the biker was not wearing a helmet, and at the signal light. Although the distribution of accident severity is essentially independent of the presence of a school zone flag, it did demonstrate the importance of intersections and the fact that most serious accidents happened away from intersections. Additionally, it is evident that many fatal incidents took place in areas that were not clearly marked as construction zones.

Almost 70% of the crashes occurred during the daylight and in city streets which points to the fact that more traffic in those regions can cause the issue. 25% of the bike crashes happened at signal light. A close second did not happen at any traffic control. Interesting about 6% of the accidents happen in bike lane! however severity of these accidents is low. Severity is high in traffic control types of "Signal Light with Red Light Running Camera", "Officer" and "Flagman". From 2010 to 2017, a general trend of an increase in the number of serious accidents was evident, however 2011 saw the highest number of accidents. Majority of the bike crashes recorded are non-injury, or non-incapacitating, however, crash enough to be recorded, number of incapacitating or killed bike crashes have been more than 27 each year which is concerning.

The correlation plot in Figure 2, Appendix shows some interesting correlations. Highest positive correlation between severity is with crash total injury count. Higher the count of injuries, higher the severity of the bike crash. Another positive indicator of severity is to know whether there was $1000 damage to nay one person’s property. Along with speed limit restrictions, crash time and person helmet positively correlate with severity of the accident. Which means, higher speed limit zones, later in the day, and not wearing a helmet can be lead to sever damage during the bike crashes.

Analysis

Data Pre-processing

Upon data cleaning and exploratory data analysis, I did several steps to get the data ready for modelling. The check for multicollinearity is where correlation between the predictor variables is quantified using variance inflation factor and helps get rid of unnecessary attributes to improve the model performance. At this stage I found crash year to have high VIF value and hence deleted the value form the independent attributes. All other VIF values were in range. The dataset was further divided into training and testing data as an 70-30 split for model training and testing respectively. The training data was further scaled using standard scaler to ensure all the independent attributes are in the same range. This is especially needed for neural networks.

Data Modeling

Logistic Regression: The aim of this analysis is to understand the properties contributing to bike crashes. The logistic regression model output displays significant variables which can be determined by the P(>|t|), p-value, and depicts the positive and negative impact over dependent variable price through the coefficients. The Significance of the variable is determined if the P-Value is less than the significance level (0.05 A significance level of 0.05 indicates a 5% risk) then it terms that the model fits the data well. However, that features with more than 0.05 and less than 0.1 are also considered significant here as most of the variables have higher p-values. The results of the regression model are shown in Figure 3 of Appendix. The 4 most significant attributes obtained in this analysis are Crash Total Injury Count (1.4), Road class (-0.16), $1000 Damage to Any One Person's Property (0.198), and Speed limit (0.11)

The confusion matrix shown below in Figure 4, Appendix represents the summary of performance for the model. As seen in the Confusion Matrix result, our Model has an accuracy of 66.8%. The model also yielded 54.8% precision which means the model makes a truly positive prediction 54.8% of the times, which is very low, and recall of the model is 66.8% which means if the response is actual positive then the model can predict it 66.8% of the times. The F1-score is 0.545 which indicates low model performance. The mean squared error value is 0.632 which tells how close a regression line is to a set of points. This can be used in comparison of the models.

Random Forest: As this is a multiclass classification model, ensemble models tend to work better. The Random Forest algorithm trains multiple decision trees using different samples of the training data, and different features are used as the root node in each tree. The final prediction is based on the collective predictions of all the decision trees in the forest. The number of estimators is set to 5000 for this analysis which represents the minimum number of trees. The feature importance that the model output is shown in Figure 5, Appendix. We can see that crash time, speed limit, crash total injury count, person helmet, intersection related, and road class are the top 6 significant variables. We can see that slight changes in significant variables as compared to logistic regression.

Evaluating the performance of the model using confusion matrix shown in Figure 1 below, accuracy of the mode is 59.9% which is lower than that of logistic regression. The model also

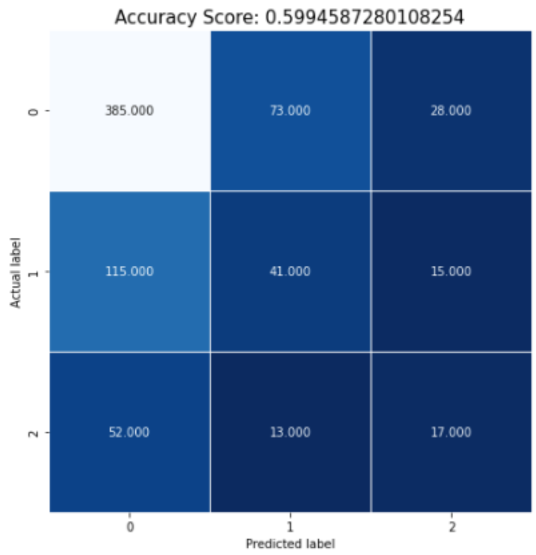


Figure 1: Confusion Matrix for RF

yielded 56.5% precision which means the model makes a truly positive prediction 56.5% of the times and recall of the model is 59.9% which means if the response is actual positive then the model can predict it 59.9% of the times. The precision has improved from logistic regression however recall shows lower performance in random forest. The F1-score is 0.578, slightly better than logistic regression, which indicates good model performance. Overall, compared to logistic regression, better performance is seen in random forest even though accuracy has gone down slightly.

Neural Networks: Attributed as a complex model, neural network method is based on artificial neural networks. In this report they are used to classify severity of the bike crash. One disadvantage of this method is that the significant features of the model can't be easily extracted. However, we can use other methods to get an understanding of the relative importance of each feature for the classification task. One common method is to use permutation feature importance, which is implemented in scikit-learn and used in this example to understand the features affecting severity of bike crash. Hyperparameters used in this implementation are hidden layers as 10, maximum iterations as 10,000, solver as 'sgd', and learning rate unit as 0.01.

The model achieved accuracy of 65.8% as shown in confusion matrix Figure 6, Appendix. It is better than random forest accuracy, but slightly lower than that of logistic regression. The model also yielded 43.2% precision which means the model makes a correct prediction 43.5% of the times and recall of the model is 65.8% which means if the response is actual positive then the model can predict it 65.8% of the times. Precision is the lowest among the three models, but recall value is better than that of random forest. The F1-score is 0.52, which indicates very low model performance and the lowest among the three models. The mean squared error value is 0.675 which is lower than random forest but slightly higher than logistic regression. Finally, to understand the feature importance through the neural network permutation importance was calculated as shown in Figure 7, appendix. The top 3 significant features are intersection related, crash time, and speed limit. We can see similar features in random forest model.

Model comparison using benchmarking metrics

Model comparison is done using benchmarking metrics, which are quantitative measures to evaluate the performance of the models. The benchmarking metrics chosen for evaluation in this report are accuracy, precision, recall, F1 score, and speed of model execution. Accuracy is a simple and intuitive measure of performance, but it can be misleading in the presence of class imbalance, which is true in the case of bike crash data, where one class has many more samples than another. Precision is the fraction of true positive predictions among all positive predictions, recall is the fraction of true positive predictions among all actual positive samples, and F1 Score is the harmonic mean of precision and recall. These metrics are often used in imbalanced class problems to get a better picture of the performance for the minority class. In a multiclass classification, false positive rate is not suitable for model comparison, similarly, MSE may not be appropriate loss function for this classification.

Model	***Accuracy***	***Precision***	***Recall***	***F1 score***
Logistic Regression	66.8%	54.8%	66.8%	0.545
Random Forest	59.9%	56.5%	59.9%	0.578
Neural Networks	65.8%	43.2%	65.8%	0.522

Table 1: Benchmarking metrics used in the models

The above table shows different model metrics that are used to compare the models. Given that the main aim of the analysis is to understand what attributes affecting the severity of bike crashes, it is important to have a balance between all the different metrics to arrive at the optimal model predictors. As mentioned earlier, precision, recall, and F1 score is given higher importance than accuracy. Random forest has a good balance of precision and recall values and has the highest F1 score showing better performance than that of logistic regression or neural network.

Another aspect to investigate is the confusion matrix. This is a table that shows the number of true positive, true negative, false positive, and false negative predictions for each class. It provides a detailed breakdown of the classifier's performance for each class and can be useful for identifying class-specific errors. When we examine the confusion matrix of logistic regression and neural network in Figure 4 and 6 respectively, we can see that the models are good in predicting low severity or the non-incapacitating or non-injury categories. They are also the highest number of records we have in the dataset. Hence it can be concluded that logistic regression and the neural networks are affected by the imbalance in the dataset.

However, random forest predicts better when severity is 1 or 2 along with good prediction when severity is 0. Hence can say with evidence that random forest is the best performing model for the given scenario. Thus, the attributes of importance from the random forest model are crash time, speed limit, crash total injury count, person helmet, intersection related, and road class.

Conclusion

From the above analysis we can conclude that higher performance is obtained from the random forest model. The 6 independent variables are chosen from the analysis as significant and contribute to severity in bike crashes and help the city of Austin to confirm the claim that bike crashes happen due to lack of support for bikers from motor vehicles, and appropriately amend laws to rectify that. The 6 features are crash time, speed limit, crash total injury count, person helmet, intersection related, and road class. These are the significant features shown by the chosen algorithm random forest.

Crashes that occur during peak traffic hours or during times of heavy vehicle and pedestrian traffic are more likely to result in severe injuries because there are more potential sources of conflict and obstacles on the road. High speed limits can contribute to the severity of bike crashes because they increase the force of impact in the event of a collision. Additionally, a cyclist's ability to react and avoid a crash is limited at high speeds. Crash involving higher number of injuries lead to fatalities which can be handled by both the city as well as the cyclists. Given that there are numerous possible points of conflict between bicycles, pedestrians, and automobiles at intersections, they are frequently high-risk locations for bicycle accidents. A crash at an intersection can be more severe if it involves a motor vehicle making a left-hand turn, as the cyclist may not be visible to the turning driver. The severity of a collision may also depend on the type of road it occurs on. For instance, severe injuries or fatalities are more likely to occur in collisions that happen on highways with high-speed restrictions or large traffic loads. Serious injuries can also be more common in collisions that happen on poorly maintained roads with poor visibility or with inadequate bicycle infrastructure.

The top 6 features as explained above show that the biker safety is not just in the hands of the city of Austin, but also the biker. Choosing low traffic times and routes with lower speed limits, proper bike lanes by the bikers and wearing helmet can help reduce the bike crashes. On the other hand, the city of Austin, can evaluate high traffic times and provide guidelines on the routes that can be followed to the bikers. Proper maintenance of bike lanes and penalties to any motor vehicles in bike lanes can incentivize bike lane culture.

References

Can the mean squared error be used for classification? (n.d.). Cross Validated. https://stats.stackexchange.com/questions/46413/can-the-mean-squared-error-be-used-for-classification

Confusion Matrix for Multi-Class Classification - Analytics Vidhya. (2021). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/

Multiclass Classification using Random Forest on Scikit-Learn Library | Codementor. (2018). Codementor | Get live 1:1 coding help, hire a developer, & more.

https://www.codementor.io/@agarrahul01/multiclass-classification-using-random-forest-on-scikit-learn-library-hkk4lwawu

Multinomial Logistic Regression — DataSklr. (2023). DataSklr. https://www.datasklr.com/logistic-regression/multinomial-logistic-regression

Multinomial Logistic Regression With Python - MachineLearningMastery.com. (2021). MachineLearningMastery.com. https://machinelearningmastery.com/multinomial-logistic-regression-with-python/

NN - Multi-layer Perceptron Classifier (MLPClassifier) - Michael Fuchs Python. (2021). Michael Fuchs Python. https://michael-fuchs-python.netlify.app/2021/02/03/nn-multi-layer-perceptron-classifier-mlpclassifier/#mlpclassifier

sklearn.ensemble.RandomForestClassifier. (2023). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

sklearn.neural_network.MLPClassifier. (2023). scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

T, B. (2021, June 9). Comprehensive Guide on Multiclass Classification Metrics. Medium. https://towardsdatascience.com/comprehensive-guide-on-multiclass-classification-metrics-af94cfb83fbd

Teng, A. (2019, August 31). Dealing with Multiclass Data. Medium. https://towardsdatascience.com/dealing-with-multiclass-data-78a1a27c5dcc

Appendix

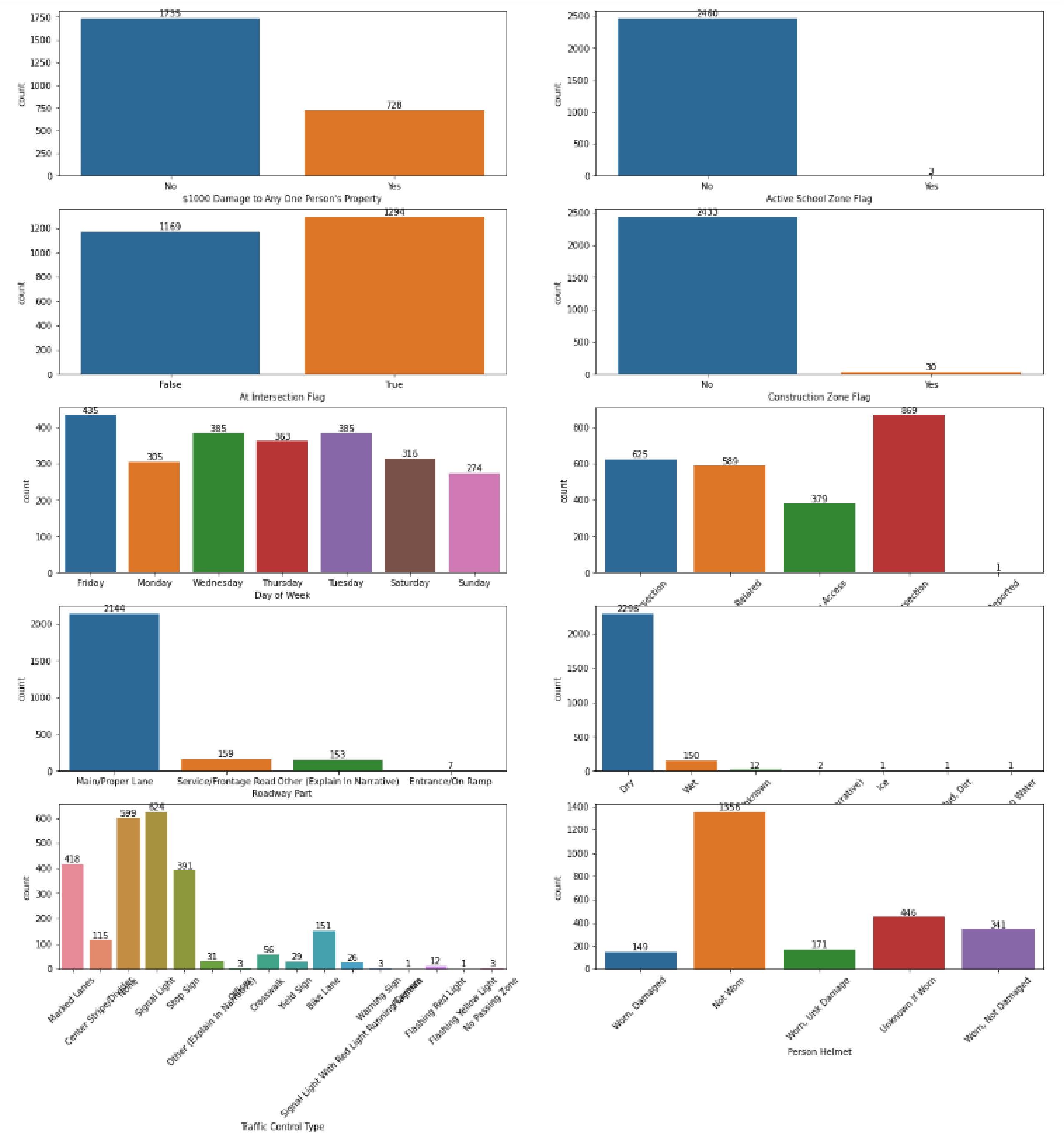
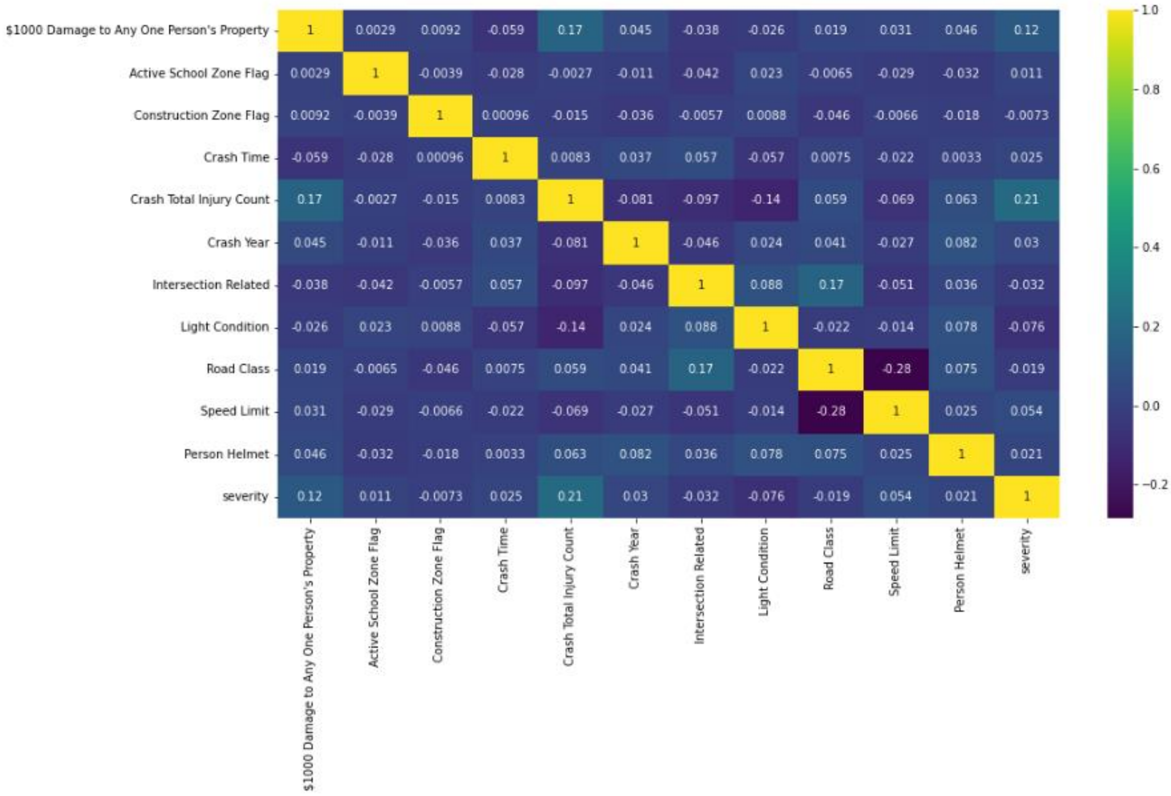
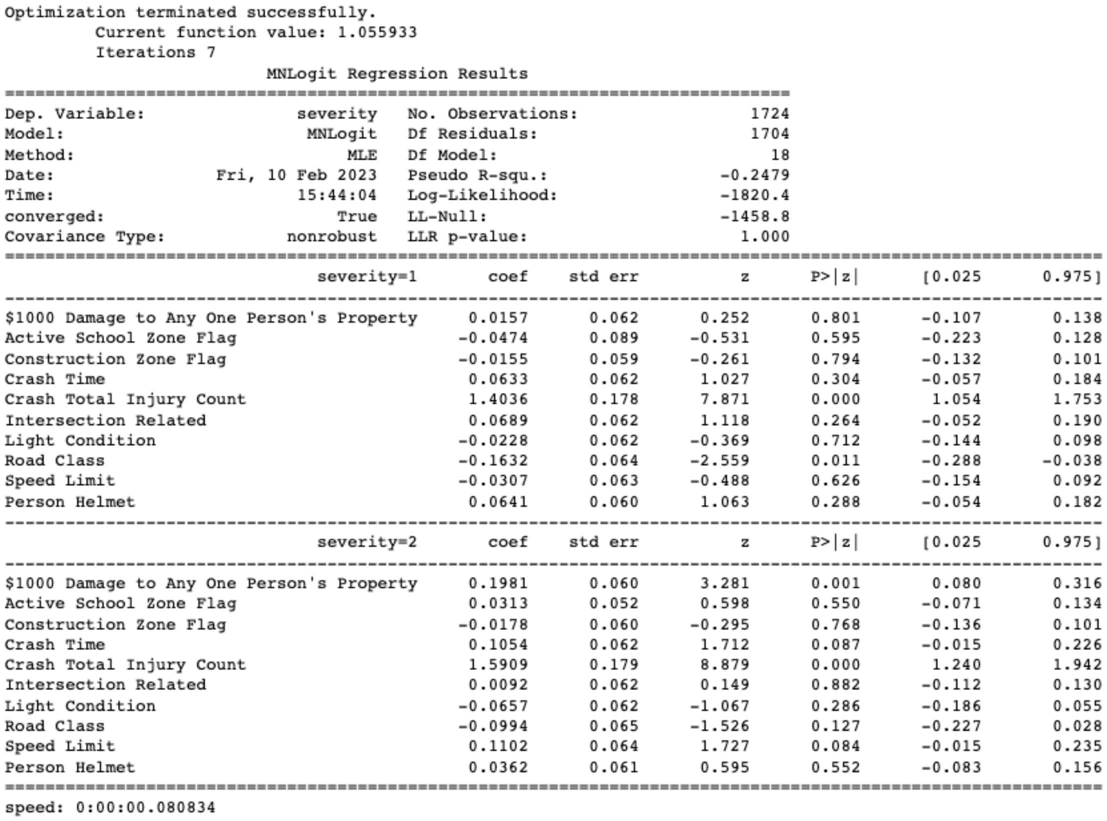


Figure 1: Count plot of categorical data in the dataset.



	$1000 Damage to Any One Person's Property	Active School Zone Flag	Construction Zone Flag	Crash Time	Crash Total Injury Count	Crash Year	Intersection Related	Light Condition	Road Class	Speed Limit	Person Helmet	severity
$1000 Damage to Any One Person's Property	1	0.0029	0.0092	-0.059	0.17	0.045	-0.038	-0.026	0.019	0.031	0.046	0.12
Active School Zone Flag	0.0029	1	-0.0039	-0.028	-0.0027	-0.011	-0.042	0.023	-0.0065	-0.029	-0.032	0.011
Construction Zone Flag	0.0092	-0.0039	1	0.00096	-0.015	-0.036	-0.0057	0.0088	-0.046	-0.0066	-0.018	-0.0073
Crash Time	-0.059	-0.028	0.00096	1	0.0083	0.037	0.057	-0.057	0.0075	-0.022	0.0033	0.025
Crash Total Injury Count	0.17	-0.0027	-0.015	0.0083	1	-0.081	-0.097	-0.14	0.059	-0.069	0.063	0.21
Crash Year	0.045	-0.011	-0.036	0.037	-0.081	1	-0.046	0.024	0.041	-0.027	0.082	0.03
Intersection Related	-0.038	-0.042	-0.0057	0.057	-0.097	-0.046	1	0.088	0.17	-0.051	0.036	-0.032
Light Condition	-0.026	0.023	0.0088	-0.057	-0.14	0.024	0.088	1	-0.022	-0.014	0.078	-0.076
Road Class	0.019	-0.0065	-0.046	0.0075	0.059	0.041	0.17	-0.022	1	-0.28	0.075	-0.019
Speed Limit	0.031	-0.029	-0.0066	-0.022	-0.069	-0.027	-0.051	-0.014	-0.28	1	0.025	0.054
Person Helmet	0.046	-0.032	-0.018	0.0033	0.063	0.082	0.036	0.078	0.075	0.025	1	0.021
severity	0.12	0.011	-0.0073	0.025	0.21	0.03	-0.032	-0.076	-0.019	0.054	0.021	1

Figure 2: Correlation plot.



```
Optimization terminated successfully.
         Current function value: 1.055933
         Iterations 7
                          MNLogit Regression Results
==============================================================================
Dep. Variable:               severity   No. Observations:                 1724
Model:                        MNLogit   Df Residuals:                     1704
Method:                           MLE   Df Model:                           18
Date:                Fri, 10 Feb 2023   Pseudo R-squ.:                 -0.2479
Time:                        15:44:04   Log-Likelihood:                -1820.4
converged:                       True   LL-Null:                       -1458.8
Covariance Type:            nonrobust   LLR p-value:                     1.000
===========================================================================================================
                              severity=1       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
$1000 Damage to Any One Person's Property    0.0157      0.062      0.252      0.801      -0.107       0.138
Active School Zone Flag                     -0.0474      0.089     -0.531      0.595      -0.223       0.128
Construction Zone Flag                      -0.0155      0.059     -0.261      0.794      -0.132       0.101
Crash Time                                   0.0633      0.062      1.027      0.304      -0.057       0.184
Crash Total Injury Count                     1.4036      0.178      7.871      0.000       1.054       1.753
Intersection Related                         0.0689      0.062      1.118      0.264      -0.052       0.190
Light Condition                             -0.0228      0.062     -0.369      0.712      -0.144       0.098
Road Class                                  -0.1632      0.064     -2.559      0.011      -0.288      -0.038
Speed Limit                                 -0.0307      0.063     -0.488      0.626      -0.154       0.092
Person Helmet                                0.0641      0.060      1.063      0.288      -0.054       0.182
-----------------------------------------------------------------------------------------------------------
                              severity=2       coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------------------
$1000 Damage to Any One Person's Property    0.1981      0.060      3.281      0.001       0.080       0.316
Active School Zone Flag                      0.0313      0.052      0.598      0.550      -0.071       0.134
Construction Zone Flag                      -0.0178      0.060     -0.295      0.768      -0.136       0.101
Crash Time                                   0.1054      0.062      1.712      0.087      -0.015       0.226
Crash Total Injury Count                     1.5909      0.179      8.879      0.000       1.240       1.942
Intersection Related                         0.0092      0.062      0.149      0.882      -0.112       0.130
Light Condition                             -0.0657      0.062     -1.067      0.286      -0.186       0.055
Road Class                                  -0.0994      0.065     -1.526      0.127      -0.227       0.028
Speed Limit                                  0.1102      0.064      1.727      0.084      -0.015       0.235
Person Helmet                                0.0362      0.061      0.595      0.552      -0.083       0.156
===========================================================================================================
speed: 0:00:00.080834
```

Figure 3: Logistic regression results.

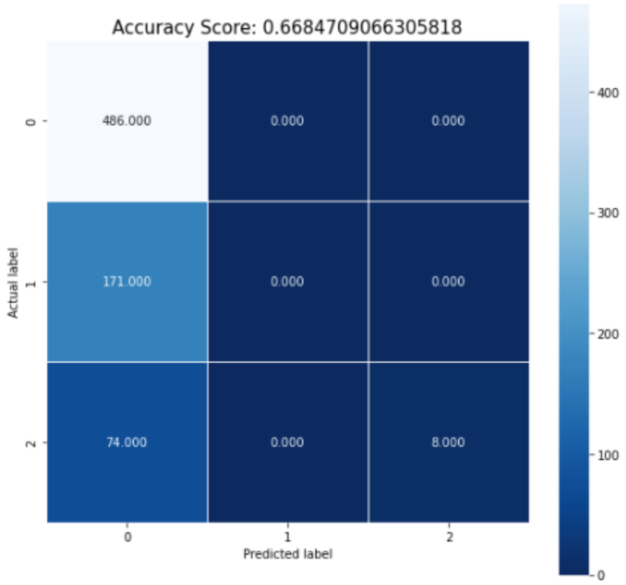


Figure 4: Confusion matrix for logistic regression.

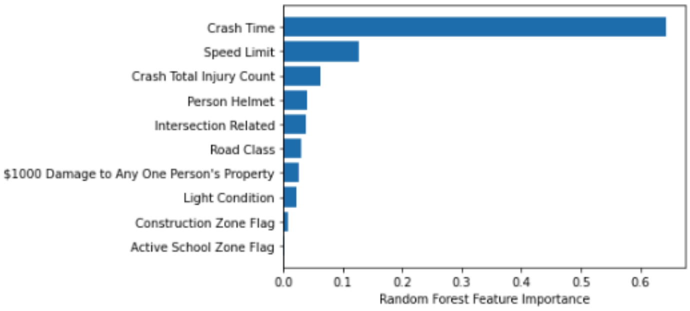


Figure 5: Feature importance from Random Forest

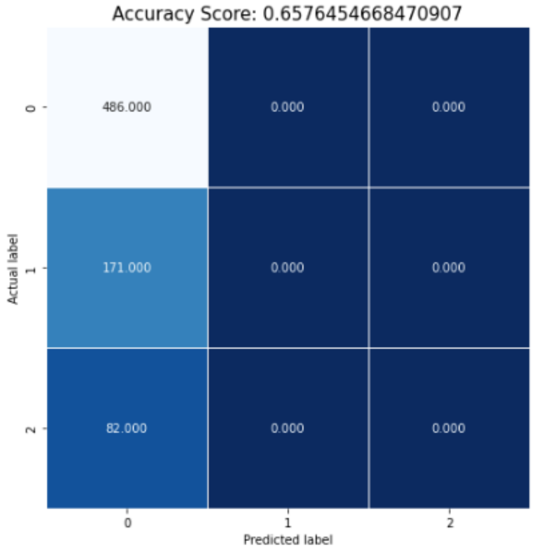


Figure 6: Confusion matrix for neural network

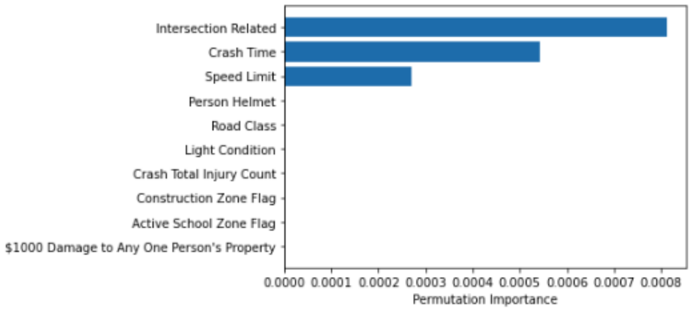


Figure 7: Permutation importance for neural network model