Northeastern University
College of Professional Studies

Module 3 - Assignment

Technique Practice

ALY6040 Data Mining Applications

Instructor: Prof. Justin Grosz

Date: Jan 29, 2023

By,

Jeseeka Shah
Shah.jes@northeastern.edu

Introduction

The report aims to cover analysis on bike crash data set from the city of Austin. The main objective of the analysis is to find out the main reasons for bike crash as bikers in the city of Austin feel that it is not safe for the bikers. The specialty of the data is that it concentrates on city aspects that contribute to bike crashes rather than the bikers. This puts focus on city features like school zone, intersection, date and time of crash, roadway part, surface condition, and traffic control type.
The report uses predictive modelling methods like logistic regression, decision trees, and random trees to determine the main contributing factors to the crash.

Problem Background

The city of Austin has taken up a data initiative to use available bike crashes data to determine the aspects of the city that contribute to the crashes, so that actionable programs can be planned to rectify the crash contributors. The dataset has crash data from 2010 to 2017, which provides comprehensive data to analyze the accidents. There are a total of 2463 records of bike crashes with 16 attributes. Attribute variables include details on categorical variables like crash severity, intersection related, roadway part, surface condition, traffic control types, and if the person wore helmet. Boolean values include if the crash costed more than $1000 damage to one person's property, active school zone flag, intersection flag, and construction zone flag. Numerical attributes include average daily traffic, crash time/year/day of the week, and speed limit.

Data Clean-up

The dataset has 11 object values, 4 integer values, and one Boolean variable. There are no null values in the dataset. However, there are some missing values in "Average Daily Traffic Amount". 85.7% of data of the data is reported as "No data". Hence this column can be deleted as majority of the data is missing. Another issue of misinformation is in the column "Speed Limit". has value of ""-1" for "13.9%" of data and value of "0" for "11.3%" of data. As the percentage of data with these values are low, we can replace by the mean of the Speed Limit as it has a normal distribution when 0 and -1 values are taken out. Mean of speed limit without the "0" and "-1" values is 35.52. In accordance with other values, speed limit is rounded down to 35 and added to all the records with "0" or "-1" values. The column "Surface Condition" has some "Unknown" values, which indicate data missing. It constitutes 0.5% of the data and as it is low, the unknown values are imputed with the mode of the column "Dry". Columns 'Traffic Control Type', and 'Day of Week' are removed as they have high number of categories and will hinder the performance of the models. "Severity" column is generated based on the existing column "Crash Severity". If the "Crash Severity" value is "Possibly Injury", "Incapacitating Injury", or "Killed", then "Severity" is assigned a value of 1, otherwise 0.

Data Analysis and Interpretation

Upon doing exploratory data analysis on the cleaned dataset, insights that help understand the data more clearly were found. Insights on maximum number of accidents can be seen from grouped bar chart of Figure 1. 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 (Figure 2 – Histogram of traffic control type). The distribution of severity of the accident is approximately independent of active school zone flag, however, it showed that intersection matters and most of the severe accidents occurred in places which are not intersections. It can also be seen that most of the severe accidents occurred in places which were not construction zone flagged. A general trend of increase in the number of severe accidents can be seen through the years 2010 to 2017. Interestingly, Friday seems to be the day of the week which had maximum number of severe accidents. Service/Frontage Road and Main/Proper Lane roadway parts form more than 50% of the severe accidents over the 8 years. More severe accidents have occurred either in high-speed limit roads like 50mph/60mph or in very low speeds like 10mph. Surface conditions like wet/dry/other (explained in narrative) also contribute to the severity of the bike crash.

From box plots for numerical values of the dataset in Figure 3 of Appendix indicates the presence of ***outliers***. Crash Total Injury Count has extreme outlier values like 15. On closer look, all the records showing this extreme number of injuries happened in wet conditions, in main road areas, which resulted in incapacitating injuries. As these values seem legit, the outliers of the column Crash Total Injury are kept as such. Speed Limit also has varied outlier values, however, speed limit spread of data is higher than normal as bike crashes have occurred in varied speed limit areas. Hence decided to keep the values.

Correlation and Multicollinearity

Correlation analysis is a well-known method to quantify association between two continuous values. It produces the "Pearson Product Moment correlation coefficient" which indicates the direction and strength of linear association.
Figure 4, the heatmap on the left, we can see that severity has positive correlation with all other numerical values of the dataset. Using One-hot encoding method, dummy variable of categorical variables are created to analyze the effects on severity. This helps convert a categorical variable to numerical variable which is easy for the algorithm to include. Figure 5 from the appendix shows correlation plot after dummy variables are created. Some of the values contributing to severity positively are crash total injury count, $1000 damage, surface condition wet, non-intersection, and service/frontage road. Values negatively contributing towards severity are surface condition dry, driveway access, and main/proper road.

High correlations between the variables of the model can affect the performance. Hence handling of multi-collinear variables help better the performance of the models. Variance Inflation Factor (VIF) help get a measurable value for multicollinearity. The check resulted in many infinite values. Hence, I removed several dummy variables to rectify the issue. The variables are: '$1000 Damage to Any One Person\'s Property_No', 'At Intersection Flag_True', 'Active School Zone Flag_Yes', 'Intersection Related_Not Reported', 'Intersection Related_Driveway Access', 'Roadway Part_Main/Proper Lane', 'Surface Condition_Dry', 'Construction Zone Flag_Yes'. After removing the above variables, there were no infinite VIF values, however there were variables with very high values such as 'Crash Year', 'Active School Zone Flag_No', and 'Construction Zone Flag_No'. Hence the above three variables were also removed from the dataset. Every other variable other than Speed Limit have 1-6 VIF scores. Speed Limit can't be removed or transformed as it is a required numerical variable for analysis. Further through the analysis, reduction of principal

components is during through the models shows below. The final multicollinearity table is shown in Figure 5.

Data analysis is the action of examining, cleaning, manipulating, and modeling data with the objective of uncovering usable information, developing conclusions, and assisting decision making. Before we begin cleaning and preparing the data, we must first do a correlation on the dataset. In the appendix, we can see that the road accident occurred between 2010 and 2017 and that we will try to determine the severity of the collision. The goal of this research is to create a prediction model that can forecast whether or not borrowers would fail on their loans. Logistic Regression Let's use Logistics Regression to see how different factors influence dependent variables. I got the result using the Logistic model below in fig.5. For '$1000 Damage to Property, Speed Limit, Construction Zone Flag,' the P-value of each variable was calculated. We can see that total injury count has the highest count' in the diagram at appendix. First, I'll use Logistic Regression to analyze this data. For constructing dummies, I will transform the category variables into numerical variables in binary form before developing our model. The dataset will also be divided into two parts: a test set and a training set. The test set was used to store 30% of the data, while the train set was used to store the remaining 70%. After that, I looked at the model predictions categorization report. This model has an accuracy of 0.66, or nearly 70%, which is not terrible for forecasting the severity of a crash.

Logistic Regression: - Logistic regression model is the first model implemented on the clean dataset as severity is represented as a Boolean variable. 70-30% division of data into training and testing respectively are done. The resulting accuracy is 66.03% with 54.15% precision and 5.14% recall. The mean squared error is 33.96. The AIC value of the regression is 2189.0514. Figure 7 shows the confusion matrix of the model as a comparison between an ideal model and the model I created. True positive is 1.76%, true negative is 64.28%, false positive is 1.49%, and false negative is 32.48%. The false positives, i.e., the model predicted not severe, but the crash was severe should be reduced for the model to perform well. The model identifies different surface condition to highly contribute towards the severity of the crash. Ice, sand, mud, dirt, standing water negatively affect the severity, i.e., with these conditions, the severity of bike crash is likely to be of low impact. However, surface conditions mentioned as other (explained in the narrative) and wet contribute positively towards the severity.

Feature selection is used to reduce the number of predictors in the model. The next two models implemented are forward logistic regression feature selection models. Forward logistic model starts with one predictor and adds the predictor sequentially until arriving at the best-case model. Sequential Feature Selection (SFS) from mlxend is used for the implementation. Using SFS with score as negative mean squared error, I identified the optimal number of features to be selected for the best model performance. Figure 8 shows a line plot visualization of performance versus the number of features for the dataset. As the performance peaks at 6 features, I selected the top six features to perform the logistic regression based on SFS results.

The top six features are 'Crash Time', 'Crash Total Injury Count', 'Speed Limit', '$1000 Damage to Any One Person's Property_Yes', 'Roadway Part_Entrance/On Ramp', and 'Surface Condition_Wet'. The logistic regression model produces accuracy of 66.58% with 63.64% precision and 5.53% recall. The mean squared error value is 33.42. AIC value is 2186.1764 which is slightly lower than the logistic regression model. The confusion matrix of the model is shown in Figure 7. The true positives are 1.89%, true negatives are 64.68%, false positives are 1.08% and false negatives are 32.34%. False positives have dropped from the first logistic regression model. Hence can conclude that forward selection model has

produced a better model than logistic regression. The top two features of this model are 'Roadway Part_Entrance/On Ramp', and 'Surface Condition_Wet'.

Decision Tree:- The next model implemented is a decision tree model with a tree depth of 4. Decision tree selects the best feature using attribute section measures (ASM) and splits the records. Here a maximum of 4 depth model is considered for analysis. The model produces an accuracy of 66.71% with 100% precision and 2.77% recall. The mean squared error value is 33.29. The confusion matrix and decision tree model with branches are shown in Figure 11. The model has predicted true positives of 0.95%, true negatives are 65.76%, false positives are 0% and false negatives are 33.29%. The false positives are 0% which is great. The accuracy has also slightly increased from the forward selection logistic regression model with reduced false positives. The MSE value has also reduced. The top features by importance are "Crash Total Injury Count", "Crash Time", "Roadway Part_Service/Frontage Road", "Intersection Related_Non-Intersection", "$1000 Damage to Any One Person's Property_Yes" and "Speed Limit". Decision trees work well with small datasets and the bike crash dataset considered in this report qualifies for one. Although logistic regression classifier is good for predicting Boolean values, decision tree model can handle skewed data better. With the nature of the dataset being highly categorical attributes, decision tree model helps in proper implementation.

Random Forest is the next model that is implemented with minimum of 5000 estimators and a tree depth of 4. RF is made up of multiple decision trees. Even though there is more flexibility in terms of number of iterations of the model, RF can result in overfitting which can create biased models. The RF model created for the data has 66.84% accuracy with 100% precision and 2.39% recall. The mean squared error value is 33.15. The accuracy of this model is the highest compared to the previous models. MSE value is also the lowest amount the four models. The true positives for the model are 0.81%, true negatives are 66.04%, false positives are 0% and false negatives are 33.15%. The top five features are "Crash Total Injury Count", "Crash Time", "Speed Limit", "$1000 Damage_Yes", "Surface Condition_Wet".

Model Comparison

	Model	Accuracy %	Precision %	Recall %	MSE	Time (ms)	AIC	False Positives %
0	Logistic Regression	66.03	54.17	5.14	33.96	103.0	2189.0514	1.49
1	Logistic Regression with Forward Selection	66.58	63.64	5.53	33.42	42.5	2186.1764	1.08
2	Decision Tree	66.71	100.00	2.77	33.29	14.4	NA	0.00
3	Random forest	66.84	100.00	2.39	33.15	7270.0	NA	0.00

From the above model metrics table, we can compare the accuracies, precision, recall, mean squared error, time in milliseconds, AIC and percentage of false positives when test data is fed into the model. These form the main metrics on which a suitable model can be chosen. I believe Random Forest is a clear winner here. It has the highest accuracy and precision. Precision is a good measure for this data as cost of false positives, i.e., the model predicting not severe and, it is severe accident, are high. Recall is a good measure of false negatives, i.e., the model predicting that the accident is severe but, it is not. The humanitarian cost is less; however, this can help improve the model. Mean squared error represents average squared difference between the estimated and actual value. MSE is the lowest in case of RF. Although RF takes a long time (because of 5000 estimators and tree depth of 4), RF produces the best results for the given data. Hence features given importance in the RF model helps in analyzing the factors contributing to severity of the bike accidents.

Features suggested by Random Forest model that contribute to high severity are count of injuries (higher the number of njuries, more severe is the accident), crash time (maximum number of crashes occur between 6am to 9pm), speed limit (30mph to 35mph have the highest number of accidents), if there is more than $1000 damage to one property, and finally a wet surface condition contribute to severe bike accidents. However, only crash time, speed limit and surface condition would provide business outlooks for the city of Austin council with actionable insights.

Conclusion and Recommendations

From the above report, it can be concluded that the city of Austin has some aspects that can be improved to better ensure the safety of bikers. Based on the model metrics comparison in the above section, Random Forest classification model best predicts the severity of the bike crashes. Maximum number of accidents happen between the time 6am to 9pm. Strategies to prioritize the safety of bikers during this time need to be implemented at an administrative level such as re-allocating road space for safe bike infrastructure, clearly marking biker lanes which are visible during day and night times, signposts indicating biker activity, and setting up a hotline for biker emergencies. Speed Limit regulations such as not allowing bikers in high-speed limit roads, putting up slow down signs where the speed limit is on the lower spectrum can help reduce the bike crashes. Another major take away from the analysis is that wet surface conditions lead to severe bike accidents. Providing a comprehensive guide to riding during wet surface conditions to the bikers or easily available on city council website helps avoid accidents. Shelters along with roadside for bikers to take rest in the areas where biker activities are high can help prevent crashes.

Reference:-

Kat, S. (2019, August 8). Logistic Regression vs. Decision Tree - DZone Big Data. dzone.com. https://dzone.com/articles/logistic-regression-vs-decision-tree

Koehrsen, W. (2018, August 30). An Implementation and Explanation of the Random Forest in Python. Medium. https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76

Shung, K. P. (2018, March 15). Accuracy, Precision, Recall or F1? Medium. https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Tan, J. (2020, October 13). Feature Selection for Machine Learning in Python—Wrapper Methods. Medium. https://towardsdatascience.com/feature-selection-for-machine-learning-in-python-wrapper-methods-2b5e27d2db31

Wu, S. (2020, May 19). Multi-Collinearity in Regression. Medium. https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea#:~:text=We%20should%20check%20the%20issue,high%20correlation%20with%20the%20rest.

Appendix

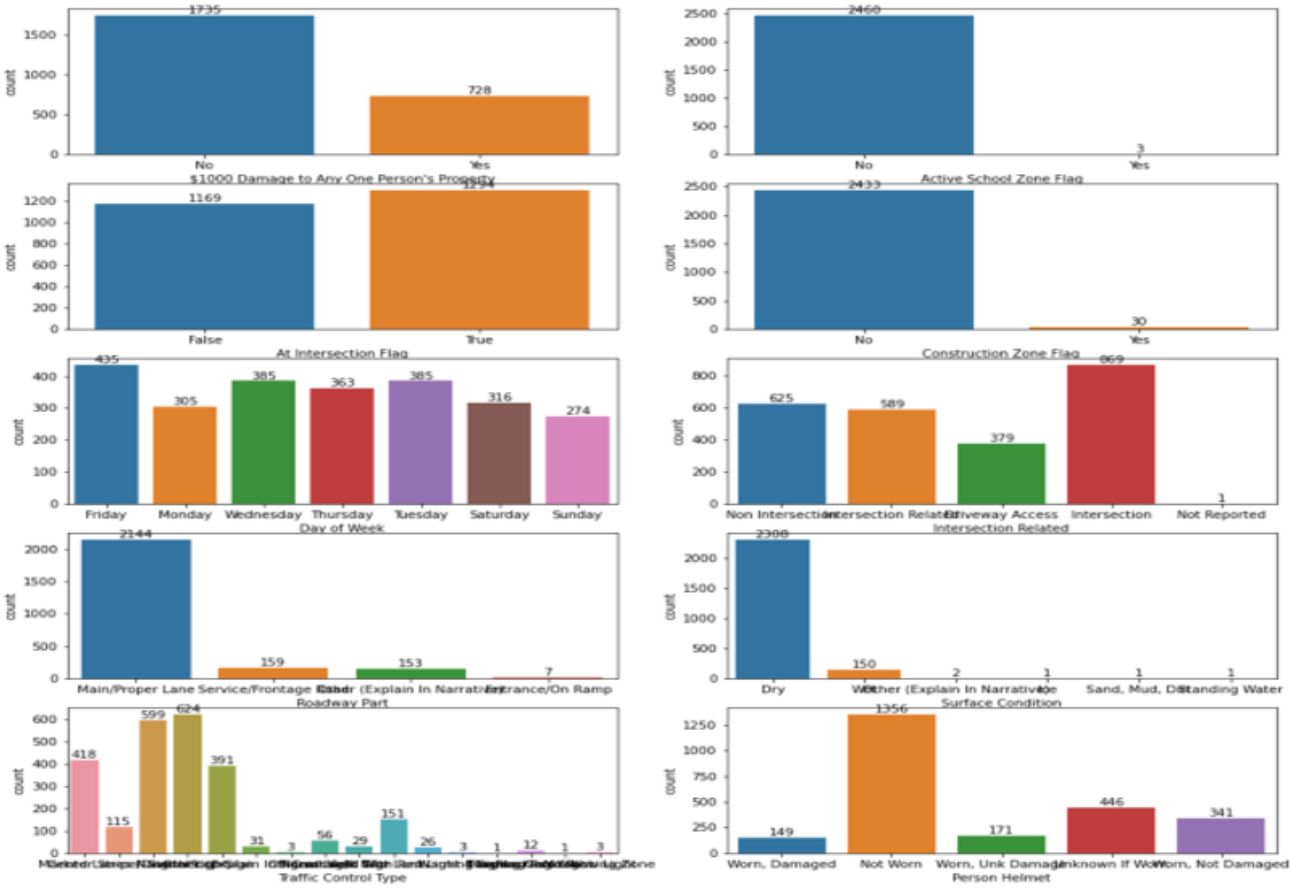

Figure 1: Understanding the dataset

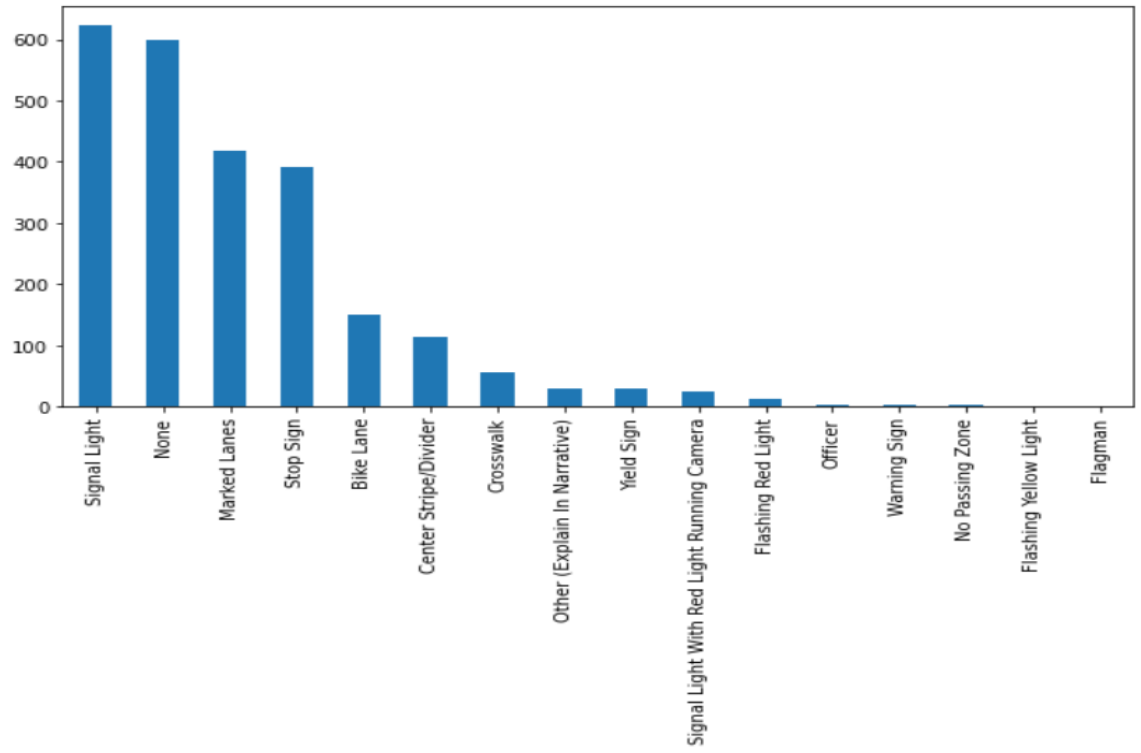

Figure 2:- Histogram of Traffic control

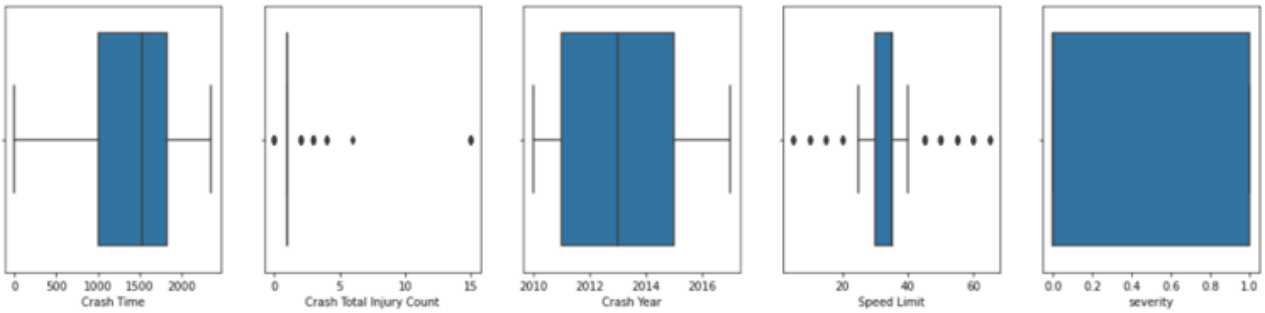

Figure 3:- Outlier shown by boxplot

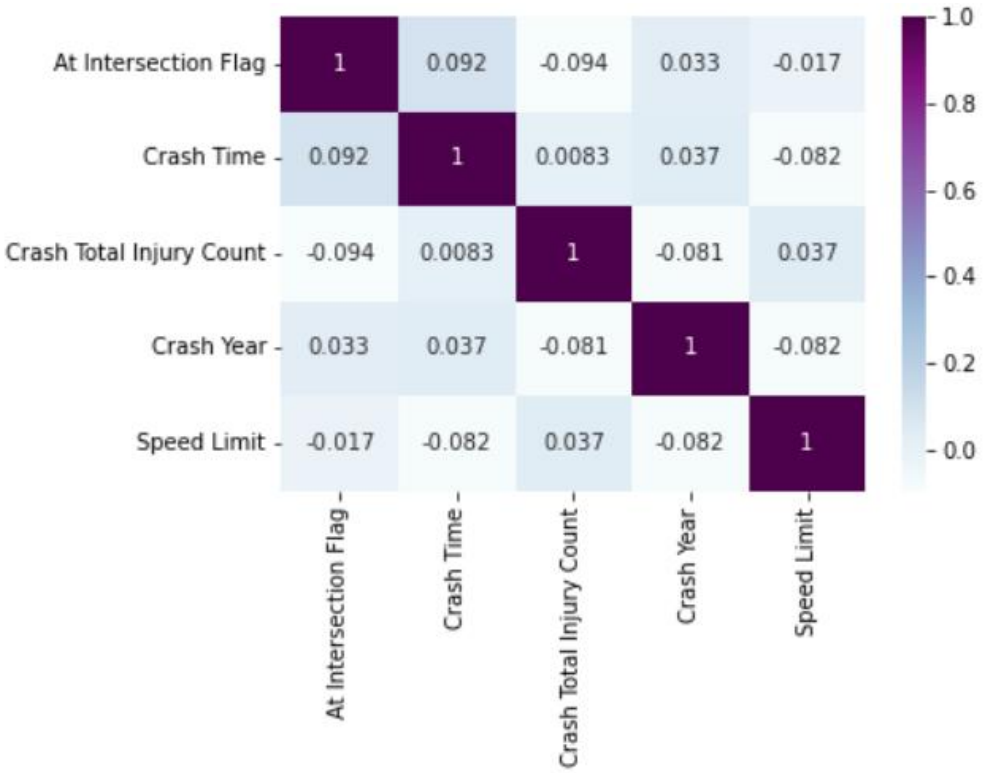

Figure 4:- heat map correlation

	variables	VIF
0	Crash Time	7.325432
1	Crash Total Injury Count	1.865175
2	Crash Year	38.436360
3	Speed Limit	30.026236
4	severity	1.539147

Figure 5:- Multicollinearity

Results: Logit

Model:	Logit	Pseudo R-squared:	0.016
Dependent Variable:	severity	AIC:	2189.0514
Date:	2023-01-29 22:24	BIC:	2276.2898
No. Observations:	1724	Log-Likelihood:	-1078.5
Df Model:	15	LL-Null:	-1096.2
Df Residuals:	1708	LLR p-value:	0.0022051
Converged:	0.0000	Scale:	1.0000
No. Iterations:	35.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Crash Time	-0.0000	0.0001	-0.5534	0.5800	-0.0002	0.0001
Crash Total Injury Count	0.2837	0.0914	3.1039	0.0019	0.1046	0.4629
Speed Limit	-0.0242	0.0057	-4.2462	0.0000	-0.0354	-0.0130
$1000 Damage to Any One Person's Property_Yes	0.2265	0.1139	1.9881	0.0468	0.0032	0.4498
At Intersection Flag_False	-0.1972	0.1617	-1.2196	0.2226	-0.5141	0.1197
Intersection Related_Intersection	-0.1857	0.1948	-0.9530	0.3406	-0.5675	0.1962
Intersection Related_Intersection Related	-0.1113	0.1813	-0.6140	0.5392	-0.4666	0.2440
Intersection Related_Non Intersection	0.0203	0.1594	0.1275	0.8986	-0.2921	0.3328
Roadway Part_Entrance/On Ramp	0.4524	0.9253	0.4889	0.6249	-1.3611	2.2658
Roadway Part_Other (Explain In Narrative)	-0.0345	0.2169	-0.1590	0.8737	-0.4596	0.3906
Roadway Part_Service/Frontage Road	0.6577	0.2016	3.2616	0.0011	0.2625	1.0529
Surface Condition_Ice	-19.1546	18572.7720	-0.0010	0.9992	-36421.1188	36382.8096
Surface Condition_Other (Explain In Narrative)	20.2991	19315.8059	0.0011	0.9992	-37837.9847	37878.5830
Surface Condition_Sand, Mud, Dirt	-18.7847	18277.5735	-0.0010	0.9992	-35842.1704	35804.6010
Surface Condition_Standing Water	-18.1910	17658.0007	-0.0010	0.9992	-34627.2365	34590.8545
Surface Condition_Wet	0.3355	0.2153	1.5582	0.1192	-0.0865	0.7574

Figure 6: Logistic regression

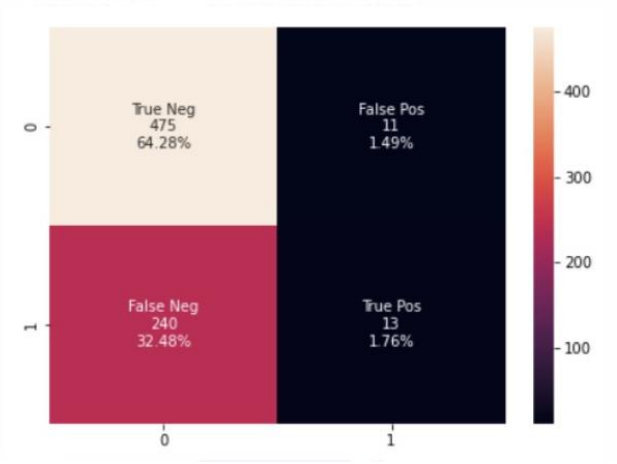

Figure 7: Correlation Logistic regression

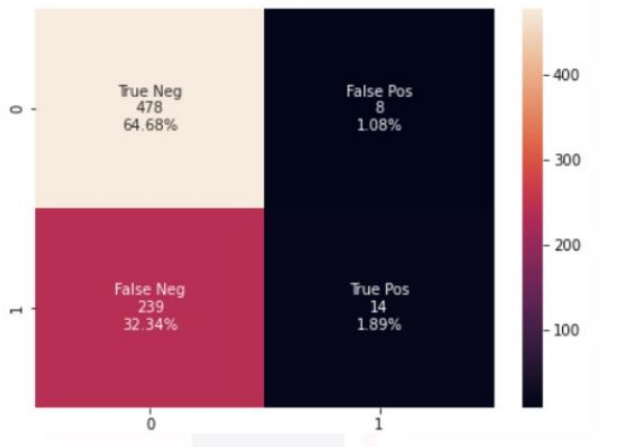

Figure 8: Correlation FSE

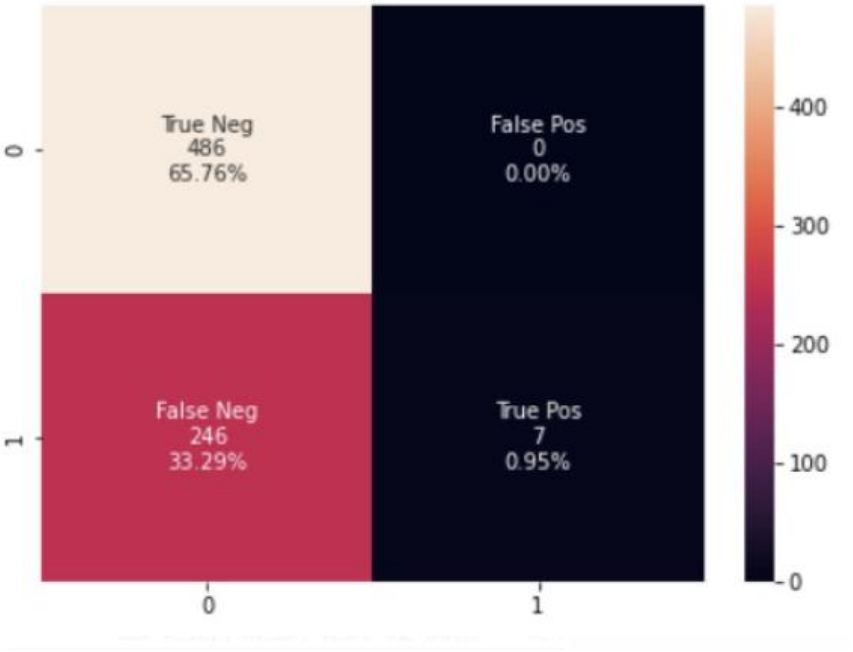

Figure 9:Decision Tree