**ALY 6040: Data Mining Applications**

Technique Practice 4

College of professional studies, Northeastern

Rohit Meena

Prof. Justin Grosz

May 8, 2022

Abstract:

In this report, I've compiled information about cyclist-related incidents in Austin. The city of Austin has received numerous complaints from cyclists alleging that the city is not doing enough to protect them from motor vehicles. As a result, I'm going to look over these findings to see if the data confirms these incidents through prediction.

Report

**PART 1**

I can build a logistic regression model to predict severity in the first part. When the dependent variable is dichotomous, logistic regression is the best regression analysis to use (binary). Logistic regression, like all regression analyses, is a predictive analysis. To describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables, logistic regression is used. The target or 'y' column in the data set is "crash severity," which has five different unique values that I converted into binary form using the "map" and "dictionary" functions, as shown in the code in the appendix. Then, with the exception of three columns ("Active School Zone Flag," "Average Daily Traffic Amount," and "Construction Zone Flag"), I chose all of the columns as features. I used the "get dummie" function to convert some of the feature columns' into categorical values to numerical values. When all of the data was collected, I trained the model and obtained an accuracy of 89 percent, as shown in the appendix. You can also see which features are most important in the appendix. "$1000 Damage to Any One Person's Property”, “Not Worn”, “Center Stripe/Divider" are the top three features According to model. Appendix I also shows the confusion matrix, where false positive is 5 and false negative is 185 which is different with other models

**PART 2**

In the second part, I'll explain how a decision tree model differs from a logistic regression model in terms of severity prediction. A supervised machine learning algorithm is a decision tree. It is used in both classification and regression algorithms. A node-based tree resembles the decision tree. The branches are influenced by a number of factors. It divides data into these types of branches until a certain value is reached. A decision tree is made up of root nodes, child nodes, and leaf nodes. I used the previous data and did not make any changes to it. I trained the model with max depth 4 and got an accuracy of 87%, as shown in the appendix. In the appendix, you can also see which features are most important. The top three features are "$1000 Damage to Any One Person's Property", "Crash Time" and "Crash Total Injuery Count”. When I compare the decision tree and logistic regression results, I notice that the decision tree model's top feature is completely different from the logistic regression model, as shown in the Appendix. It's also different when I talk about a prediction's accuracy score. The previous modal accuracy score was 89 percent, so this one is around 87 percent. We can also see the confusion matrix with false positive is 48 and false negative is 175 which is different with other models and decision tree in the appendix, where we can see the root node and branches.

**PART 3**

In the thired part, I'll explain how a random forest model differs from a logistic regression model in terms of severity prediction. Random Forest is a widely used supervised machine learning algorithm for classification and regression problems. It constructs decision trees from various samples and uses the majority vote for classification and the average for regression. One of the most important characteristics of the Random Forest Algorithm is that it can handle data sets that contain both continuous and categorical variables, as in regression and classification. For classification problems, it produces better results. I used the previous data and did not make any changes to it. I trained the model with max depth 4 and minimum threes 5000 and got an accuracy of 89%, as shown in the appendix. In the appendix, you can also see which features are most important. The top three features are "$1000 Damage to Any One Person's Property”, “Crash Total Injury Count” and “Crash Time”. When I compare the random forest model with decision tree and logistic regression, Then I notice that the random forest model top feature is completely different from the logistic regression model and decision tree model, as shown in the Appendix. It's also different when I talk about a prediction's accuracy score. The random forest and logistic regression model have approximately same accuracy score was 89 percent, but decision tree have different from others around 87 percent. We can also see the confusion matrix of random forest model with false positive is 0 and false negative is 187 which is different with other models.

**Conclusion**

The random forest model, in my opinion, has the best accuracy score of 89 percent and model fit for this data set, as well as having low false positive and false negative values when compared to other models. It has a top feature that can be adjusted and used to predict severity.

Appendix

Chart

Description automatically generated with medium confidence

Figure 1 Top 10 Feature of Logistic Regression

A picture containing chart

Description automatically generated

Figure 2 Top 10 Feature of Decision Tree

Chart

Description automatically generated

Figure 3 Top 10 feature of Random Forest

Chart, bar chart

Description automatically generated Chart, bar chart, waterfall chart

Description automatically generated Chart, bar chart, waterfall chart

Description automatically generated

*Fig 4 confusion Matrix LR Fig 5 Confusion Matrix DT Fig 6 Confusion Matrix RF*

Diagram

Description automatically generated

*Fig 7 Decision Tree*