Northeastern University
College *of* Professional Studies

DATA MINING APPLICATIONS

ALY 6040, CRN 80440

PROFESSOR JUSTIN GROSZ

MODULE 4 ASSIGNMENT

TECHNIQUE PRACTICE

SUBMITTED BY:

Richa Umesh Rambhia

Abstract

The project deals with determining and predicting the type of accident taking place in the city of Austin. The city of Austin has heard of many complaints from cyclists that the city is not doing enough to protect them from motor vehicles. Thus, in order to deal with this problem and confirm the complaints that have been obtained, the city has complied the data of cyclists which are related to the various incidents that have occurred which would help in reviewing the findings to check if the data could be able to confirm this.

The dataset contains information about the accidents that have occurred based on the various factors and parameters. It consists of the various field values such as average daily traffic amount, crash severity, which is the type of accident, crash time, crash injury count, intersection related, speed limit, surface conditions, person helmet, etc. which would help in predicting the type of accident based on these external factors. This data would thus help in understanding what are the various possible factors which are leading to the accidents based on the severity of the incident that occurred, and also would help in determining to confirm the complains obtain from the cyclists based on the protection from motor vehicles.

In order to predict the type of accident and confirm the complaints obtained from the cyclists, various machine learning models such as Logistic Regression model, Decision Tree Classifier, and Random Forest Classifier, are implemented which would help to predict the type of accident and also based on the various metrices and feature importance it would help to determine which factors need to be considered in order to avoid such accidents. Initial exploratory data analysis and data visualizations would help in understanding the data and gaining more insights about the various types of accidents occurred based on their respective features.

Table of Contents

Introduction

The data from the city of Austin has been complied for the various incidents that have occurred of cyclists in order to review the data to check if the data confirms the complaints of the cyclists regarding the protection from the motor vehicles. This data is used in order to analyze and predict the type of accident that occurs based on the various parameters and whether the complains received hold true with respect to the cyclists. The features and field values are an important aspect of the dataset as these would help in determining the parameters which affects the accident and that can be taken care of in order to avoid the accident.

The objective of the project is to focus on recommending the areas and factors that need attention to order to prevent dangerous incidents and build a model in order to predict the severity of the accident which the city can be able to use for further data collected. The various models built would be Logistic Regression, Decision Tree Classifier, and Random Forest Classifier which would help in understanding and comparing these models to determine the best fit model which can then be used by the city to predict the severity of the accidents.

The various data analysis process are implemented like the exploratory data analysis, data visualizations, data cleaning, and model building in order to understand, explore, and analyze the dataset obtained. The field values of the dataset are with respect to the accidents such as the crash severity, crash total injury count, crash time, person helmet, surface condition, traffic amount, etc. which would be some of the key factors in determining and analyzing the accidents occurred to predict the type of accident.

Exploratory Data Analysis

In order to understand and explore the data various methods are implemented such as the descriptive and statistical analysis of the dataset and data profiling report in order to gain more insights about the dataset and its field values. From the descriptive analysis it is observed that there are **2463 rows of data and 16 field values**, from which there are **11 categorical** variable type**, 4 numerical** data type, and **1 boolean data type**. The figure below displays the total number of rows and columns in the dataset along with the various field values present.

```
Total number of Rows and Columns: (2463, 16)

Column Names:
 Index(['$1000 Damage to Any One Person's Property', 'Active School Zone Flag',
       'At Intersection Flag', 'Average Daily Traffic Amount',
       'Construction Zone Flag', 'Crash Severity', 'Crash Time',
       'Crash Total Injury Count', 'Crash Year', 'Day of Week',
       'Intersection Related', 'Roadway Part', 'Speed Limit',
       'Surface Condition', 'Traffic Control Type', 'Person Helmet'],
      dtype='object')
```

Figure 1. Descriptive Analysis of the dataset

The below figure gives the statistical analysis of the data where it is observed that the average of speed limit is **26.4** and total crash count is **2463**. This helps in understanding the overall speed limit and the crash count in the dataset.

	Crash Time	Crash Total Injury Count	Crash Year	Speed Limit
count	2463.0	2463.0	2463.0	2463.0
mean	1404.7	1.1	2013.4	26.4
std	559.0	1.2	2.2	17.0
min	1.0	0.0	2010.0	-1.0
25%	1008.0	1.0	2011.0	0.0
50%	1532.0	1.0	2013.0	30.0
75%	1822.5	1.0	2015.0	35.0
max	2358.0	15.0	2017.0	65.0

Figure 2. Statistical Analysis of the dataset

The data profiling report generated further helped in understanding the data values, missing and duplicate data, and the correlation between each variables in the dataset. From the dataset statistics it can be observed that there are 2463 observations, and 0 missing cells, along with no duplications.

Dataset statistics

Number of variables	16
Number of observations	2463
Missing cells	0
Missing cells (%)	0.0%
Total size in memory	1.8 MiB
Average record size in memory	758.7 B

Variable types

Categorical	11
Boolean	1
Numeric	4

Figure 3. Dataset Statistics and Variable types

Data Cleaning

In order to check for the missing values or null values in the dataset, the total of missing values in each of the columns are checked, and it is observed that the dataset is clean and has **no missing or null values**. Although, there are some field values which needed **renaming** and the datatype for the field values are checked for conversion of type of data for any **wrong datatype**. Apart from this, the average daily traffic amount has numerical and categorical data, and thus the categorical data, **'No Data'**, is replaced by **0** indicating that there is **no data available** for that particular row value, hence making it easier for datatype conversion and also for training of the model.

Data Visualization

Data Visualization is performed on the bike dataset in order to gain more insights about the dataset and its various parameters. The below visual represents the crash severity data which shows the total count and frequency of the various types of accidents that have occurred which can be simplified further to severe and non-severe accident. This helps in understanding the major type of accident that occur which can then be categorized as severe or non-severe and as observed non-incapacitating injury has the total overall count of **1474** and killed has the least count which

indicates a severe accident. This indicates that the accidents occurred are highly non-incapacitating injuries and there are approximately **16 crash severity** incidents i.e., 0.6% that are categorized as severe or killed.

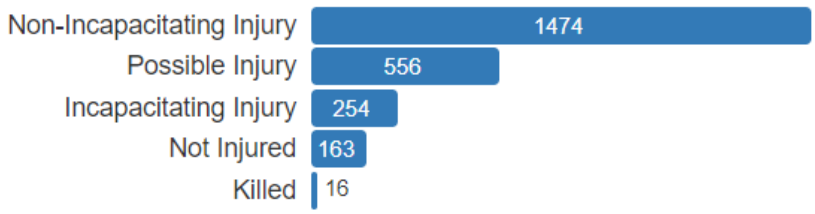


Figure 4. Total count of Crash Severity

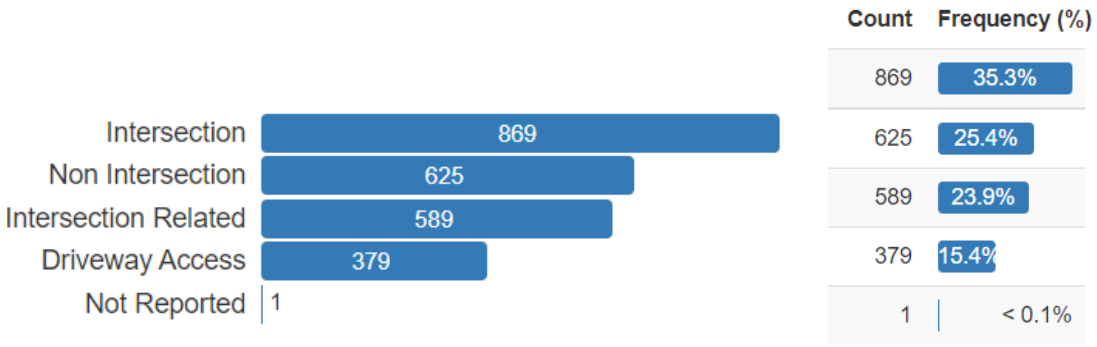


Figure 5. Intersection Related frequency count

The above visual represents the intersection related frequency count and total percent of accidents occurred due to the type of intersection. It is observed that the majority of the accidents have occurred for the type of **intersection roads**, which is **35.3%** of the times for a **total count of 869** whereas **15.4% for driveway access** at a total count if **589** and less than **0.1%** of accidents have not reported for any type of accidents that have occurred.

The next visual represents the speed limit histogram which helps understanding the various speed limits that have been determined and monitored during the accidents that have taken place. As observed, the **maximum speed limit frequency** lies in the **range of 40** whereas the **average** speed limit is approximately **26.41**.

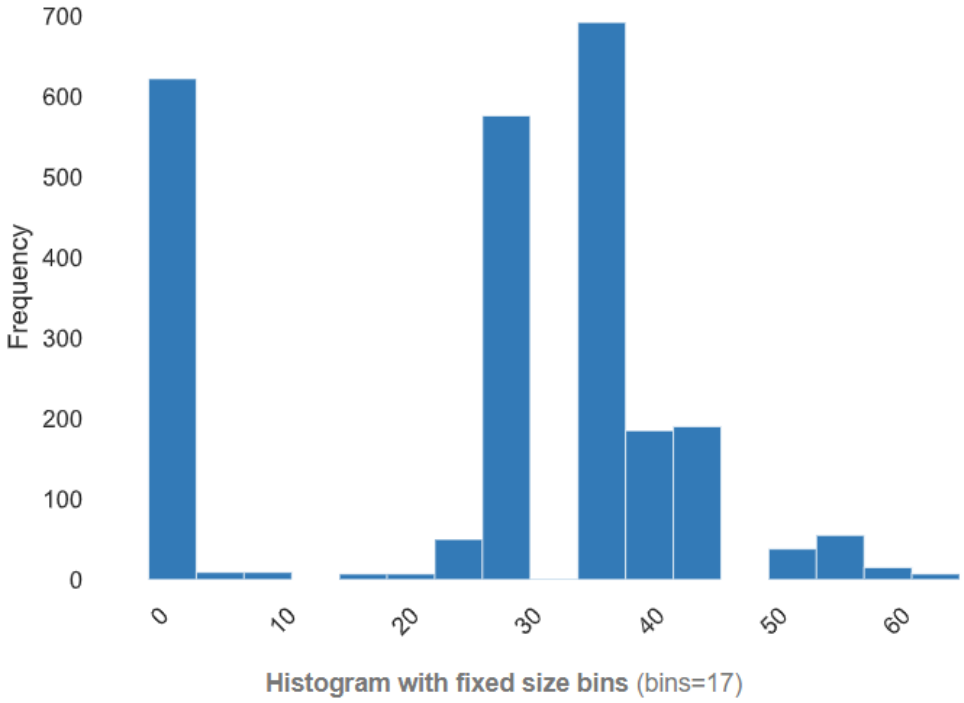


Figure 6. Speed Limit Histogram

The person helmet frequency plot below helps understand the count of accidents that have occurred when the helmet was worn by the person or not worn by the person and thus gives information to ensure that the riders are following the rules and wearing a helmet. As observed, **1356** of the total count have **not worn the helmet** causing the incidents to happen whereas on the other hand, **149** have **worn the helmet still faced damaged** and injury during the accident.

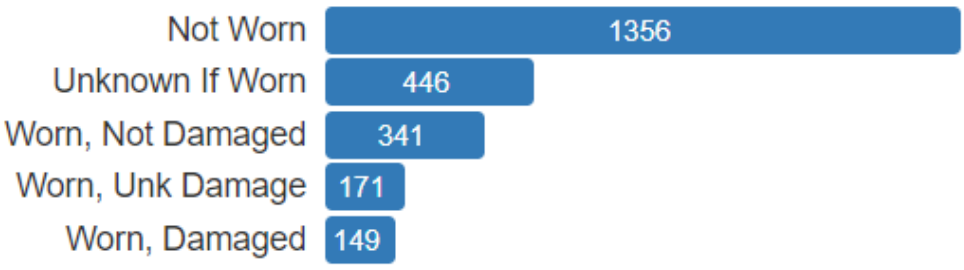


Figure 7. Person Helmet frequency count

Pre Modeling Steps

In order to understand the features for the implementation of model to predict the crash severity various pre modeling steps are implemented in order to understand the important features to be considered for modeling and to check if there is any collinearity that exists between the variables as that would highly affect the model built. The first step performed is automatic feature selection and extraction which helps in determining the important features that need to be considered for the training of the model. The next steps are label encoding for categorical variables, and correlation plot to understand the correlation between each of the variables of the dataset.

Feature Selection & Extraction

The automatic feature selection and extraction helped in understanding what the important features are with respect to the model building in order to predict the crash severity. Thus, with respect to each of the variables in comparison with each other, the top 10 features were selected as shown in the below figure.

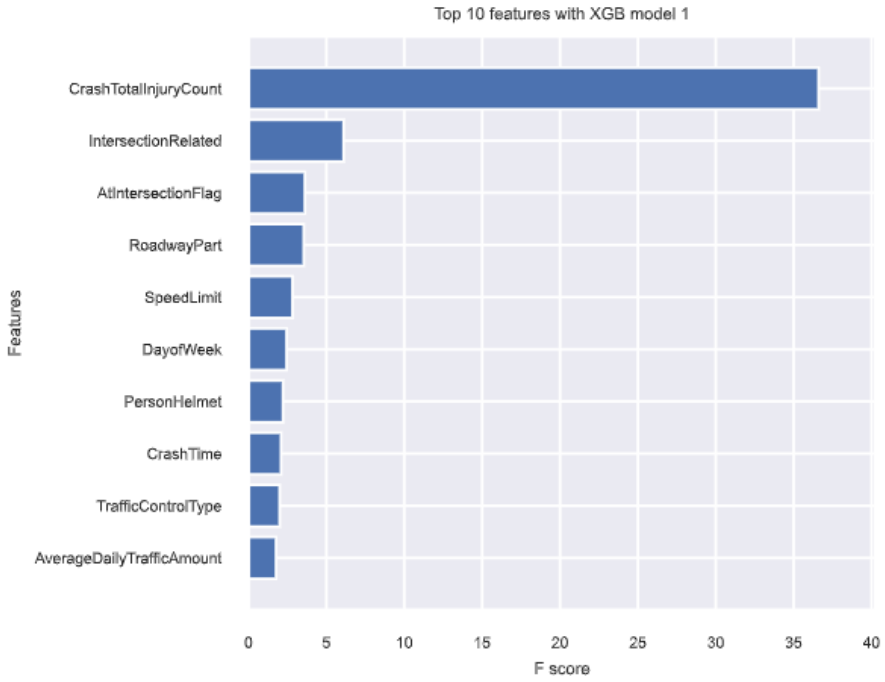


Figure 8. Features with XGB model

['CrashTotalInjuryCount', 'IntersectionRelated', 'AtIntersectionFlag', 'RoadwayPart', 'SpeedLimit', 'DayofWeek', 'PersonHelmet', 'CrashTime', 'TrafficControlType', 'AverageDailyTrafficAmount', 'SurfaceCondition', 'ActiveSchoolZoneFlag']

Thus, these are the features that are extracted by the algorithm which can be considered as features for training of the model based on the correlation between each variable.

Correlation Plot

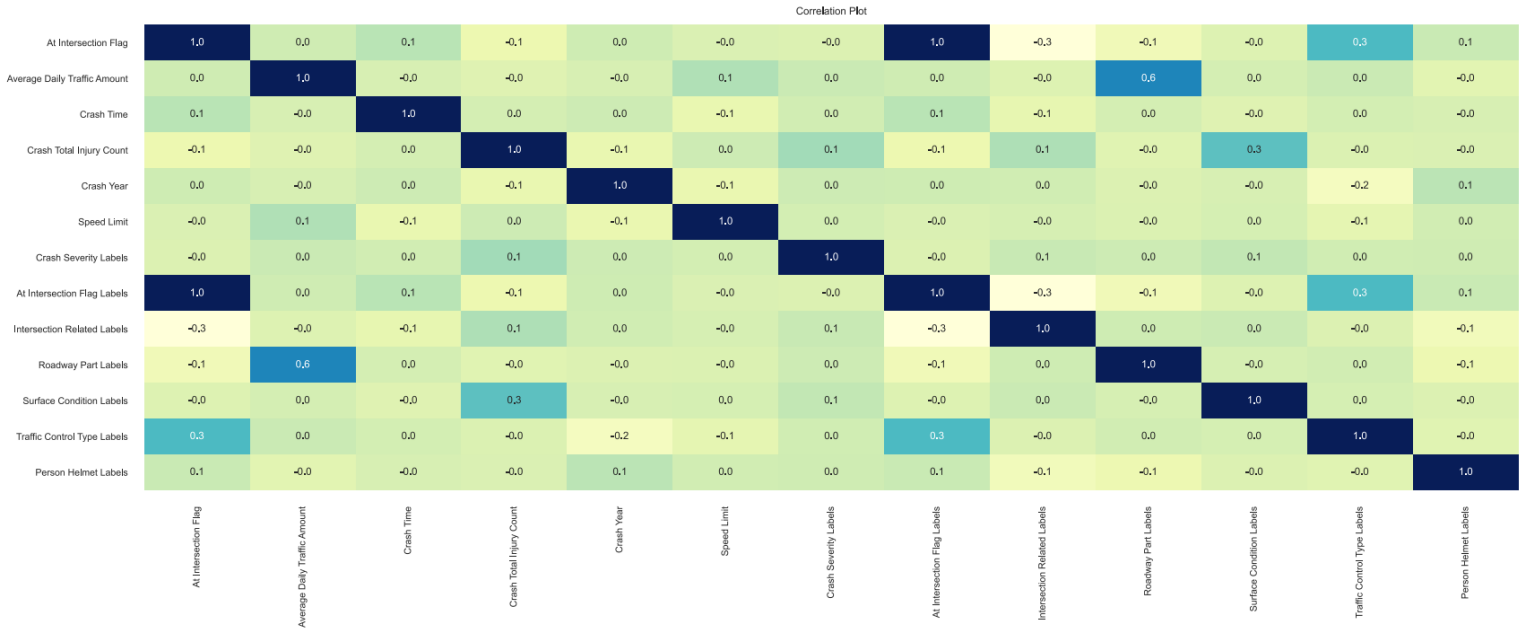


Figure 9. Correlation Plot

The correlation plot helps understand the collinearity between each of the parameters of the dataset in order to analyze the variables for multicollinearity which needs to be removed to avoid overfitting of the machine learning model. The variables with high collinearity value needs to be removed from the features selected before they are given to the model. Thus, as observed in the correlation plot above there is a **high collinearity** between the variables **Average Daily Traffic Amount and Roadway Part with a correlation value of 0.6**, and hence either of the two variables need to be removed before implementing the model, which would help reduce errors in the model. Also, there is a **high collinearity** between **Traffic Control Type & At Intersection Flag, and**

Crash Total Injury Count & Surface Condition with a correlation value of 0.3, and thus either of these variables need to be removed from the features selected to avoid multicollinearity.

Grouping Categories

Since the crash severity field which is the predictor variable has multiple categories, it can be grouped into two categories which becomes easy for classification and understanding the data with respect to the two categories. Hence, the values for the crash severity column are categorized as two values namely, Severity and Non-Severe Incident where **'Incapacitating Injury', 'Possible Injury', 'Killed'** are categorized into Severity class and **'Non-Incapacitating Injury', 'Not Injured'** are categorized into Non-Severe Incident class. Thus, these two categories would be now considered for the prediction of the model.

Label Encoding

Since there are categorical values present in the dataset which need to be extracted as the features for the model building, label encoding is performed on the categorical data in order to able to select the features for the training of the model. Thus, along with the **predictor variable** i.e., **Crash Severity**, the **independent variables** such as **Surface Conditions, Traffic Control Type, Roadway Part**, etc. are label encoded into a new column as numerical data.

At Intersection Flag	Average Daily Traffic Amount	Construction Zone Flag	Crash Severity	Crash Time	Crash Total Injury Count	Crash Year	Day of Week	...	Traffic Control Type	Person Helmet	Crash Severity Labels	Damage to any persons property Labels	At Intersection Flag Labels	Intersection Related Labels	Roadway Part Labels	Surface Condition Labels	Traffic Control Type Labels	Person Helmet Labels
False	15262	True	Severity	239	1	2010	Friday	...	Marked Lanes	Worn, Damaged	1	0	0	3	1	0	6	2
False	0	True	Non Severe Incident	310	2	2010	Friday	...	Center Stripe/Divider	Not Worn	0	0	0	3	1	0	1	0
False	0	True	Non Severe Incident	310	2	2010	Friday	...	Center Stripe/Divider	Not Worn	0	0	0	3	1	0	1	0
False	0	True	Non Severe Incident	310	2	2010	Friday	...	Center Stripe/Divider	Not Worn	0	0	0	3	1	0	1	0

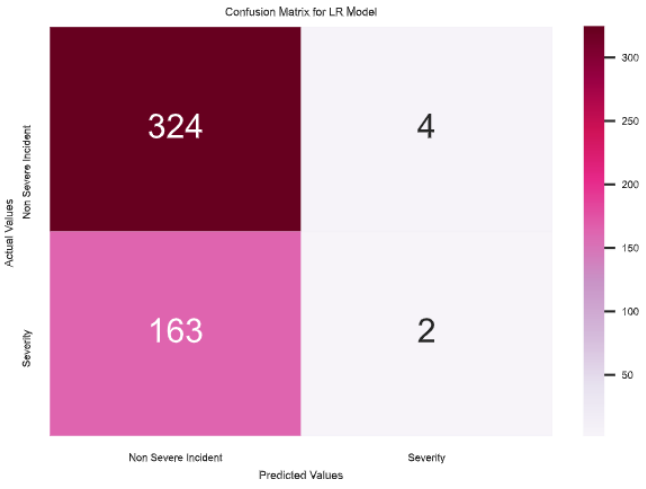
Figure 10. Label Encoding

Model Building

After the pre modeling steps are performed and implemented, the independent variables and predictor variables now can be given to the various machine learning models for the prediction based on the problem statement. The three machine learning models that are implemented for the prediction of the type of accident are **Logistic Regression Model, Decision Tree Classifier, and Random Forest Classifier.** Based on the feature selection, and correlation plot the various features that are extracted for the implementation of the model building phase are, *'Average Daily Traffic Amount', 'Crash Total Injury Count', 'Intersection Related Labels', 'Roadway Part Labels', 'Speed Limit', 'Surface Condition Labels', 'Traffic Control Type Labels', and 'Person Helmet Labels'*. Based on the correlation matrix and the accuracy of the prediction of each model, the features extracted differ.

Logistic Regression Model

For the implementation of the Logistic Regression Model, the features extracted for training of the model are 'Average Daily Traffic Amount', 'Speed Limit', 'Surface Condition Labels', 'Traffic Control Type Labels', and 'Person Helmet Labels'. The data is split into 80% training data and 20% as test dataset. The accuracy obtained for both the training and testing set of the LR model is **66%** for which the confusion matrix, model summary, and classification report is as follows.




```
Accuracy of Logistic Regression on training set: 0.66
Accuracy of Logistic Regression on test set:     0.66
0.6612576064908722
```

Figure 11. Confusion Matrix & Accuracy

As observed, **324 are correctly classified** as non-severe incident whereas **163 are incorrectly** classified to non-severe incident. Here, it is important to minimize the wrong prediction of severity into non-severe incident.

Logit Regression Results

Dep. Variable:	Crash Severity Labels	No. Observations:	2463
Model:	Logit	Df Residuals:	2458
Method:	MLE	Df Model:	4
Date:	Wed, 15 Jun 2022	Pseudo R-squ.:	-0.009814
Time:	15:36:23	Log-Likelihood:	-1586.6
converged:	True	LL-Null:	-1571.2
Covariance Type:	nonrobust	LLR p-value:	1.000

	coef	std err	z	P>|z|	[0.025	0.975]
Average Daily Traffic Amount	2.687e-06	1.29e-06	2.085	0.037	1.61e-07	5.21e-06
Speed Limit	-0.0086	0.002	-4.120	0.000	-0.013	-0.005
Surface Condition Labels	0.0817	0.028	2.963	0.003	0.028	0.136
Traffic Control Type Labels	-0.0493	0.007	-6.815	0.000	-0.063	-0.035
Person Helmet Labels	-0.0108	0.031	-0.353	0.724	-0.071	0.049

Figure 12. Model Summary

From the model summary it can be observed that the p-value is less than the significance value of 0.05 for variables **Speed Limit, Traffic Control Type, Surface Condition, and Average Daily Traffic Amount**, which implies the variables to be significantly important that need to be considered for the prediction of the crash severity. Based on the p-value and significance of the variables, the positive coefficient values considered are average daily traffic amount and surface condition and for the negative coefficient we consider the speed limit and traffic control type. Hence, it is observed that the severity of the accident would increase in the increase of traffic amount and thus that needs to be taken into consideration. Apart from this, the surface condition and speed limit would also affect the severity of the accident, and thus for the decrease in speed limit, there is less chance for the accident to occur or have a non-severe incident. The surface condition needs to be checked as for an increase in the surface condition, the severity of accident might increase.

Classification report LogisticRegression():

	precision	recall	f1-score	support
0	0.67	0.99	0.80	328
1	0.33	0.01	0.02	165
accuracy			0.66	493
macro avg	0.50	0.50	0.41	493
weighted avg	0.55	0.66	0.54	493

Figure 13. Classification Report

Decision Tree Classifier

The decision tree classifier is implemented for a maximum depth of 4 branches where the above mentioned features are extracted for the prediction of crash severity but based on the correlation matrix and accuracy of the model, some of the features are either removed or added which help in improving the model performance. Thus, the features considered for the decision tree model are 'At Intersection Flag Labels', 'Average Daily Traffic Amount', 'Intersection Related Labels', 'Speed Limit', 'Surface Condition Labels', 'Traffic Control Type Labels', and 'Person Helmet Labels'. The accuracy obtained for this model is **68% for training set and 67% for testing set.** The decision tree, confusion matrix, classification report, and feature importance are as shown in the below figures.

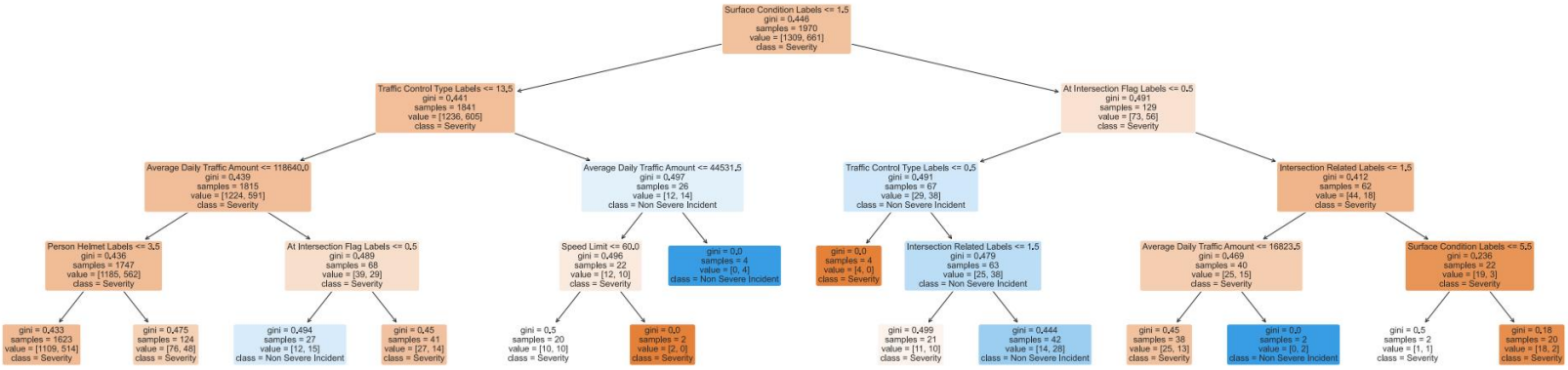
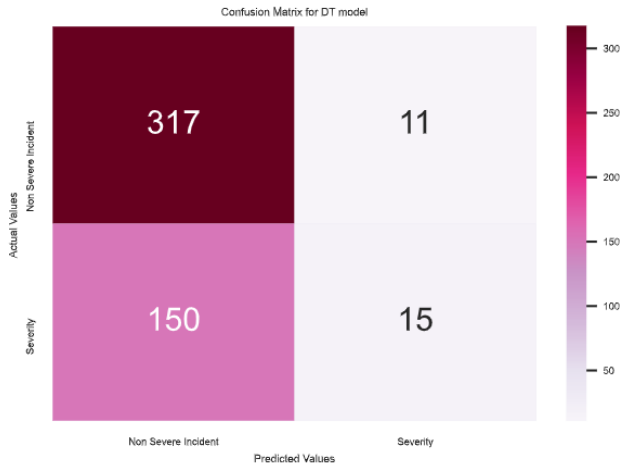


Figure 14. Decision Tree

From the confusion matrix below, it is observed that there are **317** classes correctly classified into non-severe incident whereas **15** of them correctly classified as severe incident. Hence, it is important to **minimize the False Negative** value here as the accidents are classified as non-severe accidents despite being classified into severe accidents. But the performance of decision tree model when compared to logistic regression model is better when it comes to the minimization of false negative values as LR model could only classify **2 of the classes into severe category** whereas DT model was able to classify **15 into the correctly classified** class. Hence, Decision Tree Classifier would be preferred for the prediction of crash severity over Logistic Regression model.




```
Accuracy of Decision Tree on training set: 0.68
Accuracy of Decision Tree on test set:     0.67
0.6734279918864098
```

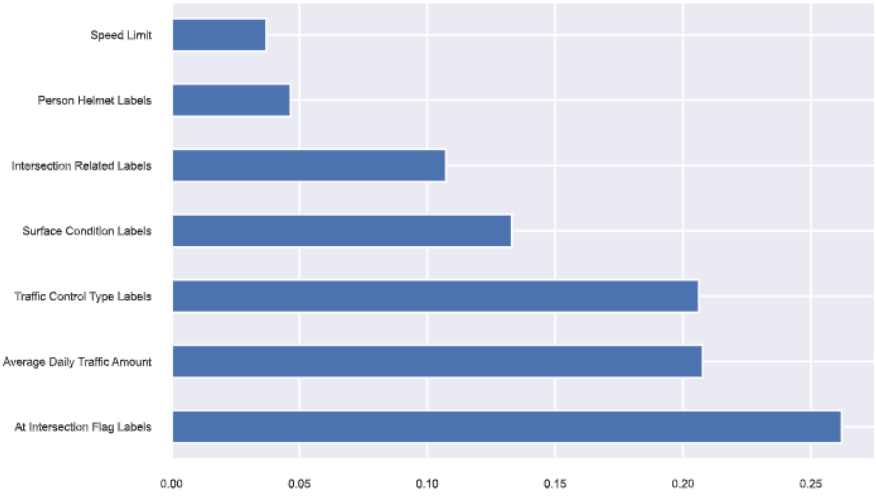
Figure 15. Confusion Matrix & Accuracy

```
Classification report DecisionTreeClassifier(max_depth=4, random_state=42):
              precision    recall  f1-score   support

           0       0.68      0.97      0.80       328
           1       0.58      0.09      0.16       165

    accuracy                           0.67       493
   macro avg       0.63      0.53      0.48       493
weighted avg       0.64      0.67      0.58       493
```

Figure 16. Classification Report



	Feature	Importance
0	At Intersection Flag Labels	0.262
1	Average Daily Traffic Amount	0.208
5	Traffic Control Type Labels	0.206
4	Surface Condition Labels	0.133
2	Intersection Related Labels	0.107
6	Person Helmet Labels	0.047
3	Speed Limit	0.037

Figure 17. Feature Importance

Based on the feature importance plotted for the Decision Tree model, it can be observed that the important features that are considered during the prediction of crash severity by the model are **At Intersection Flag, Average Daily Traffic Amount, and Traffic Control Type**. These features thus affect the prediction of the crash severity which needs to be considered in order to avoid or minimize the severity of the accidents. The road intersections thus affects the accidents to occur and thus it is recommended to keep a check on these intersections where the severity of the accident is more. Also, the traffic amount is one of the factor affecting the accidents and thus some solution needs to be implemented with respect to the traffic rules in order to avoid such accidents.

Random Forest Classifier

For the Random Forest Classifier, the model was implemented with a minimum of **5000 tress** and a maximum depth of **4 branches**. Similar to Decision Tree Classifier, the features selected and extracted for the Random Forest model are, 'Damage to any persons property Labels', 'At Intersection Flag Labels', 'Intersection Related Labels', 'Speed Limit', 'Surface Condition Labels', and 'Person Helmet Labels'. The accuracy of the Random Forest Classifier model obtained on both training and testing set is **67%**. The confusion matrix, accuracy, classification report, and feature importance for RF model are as follows.

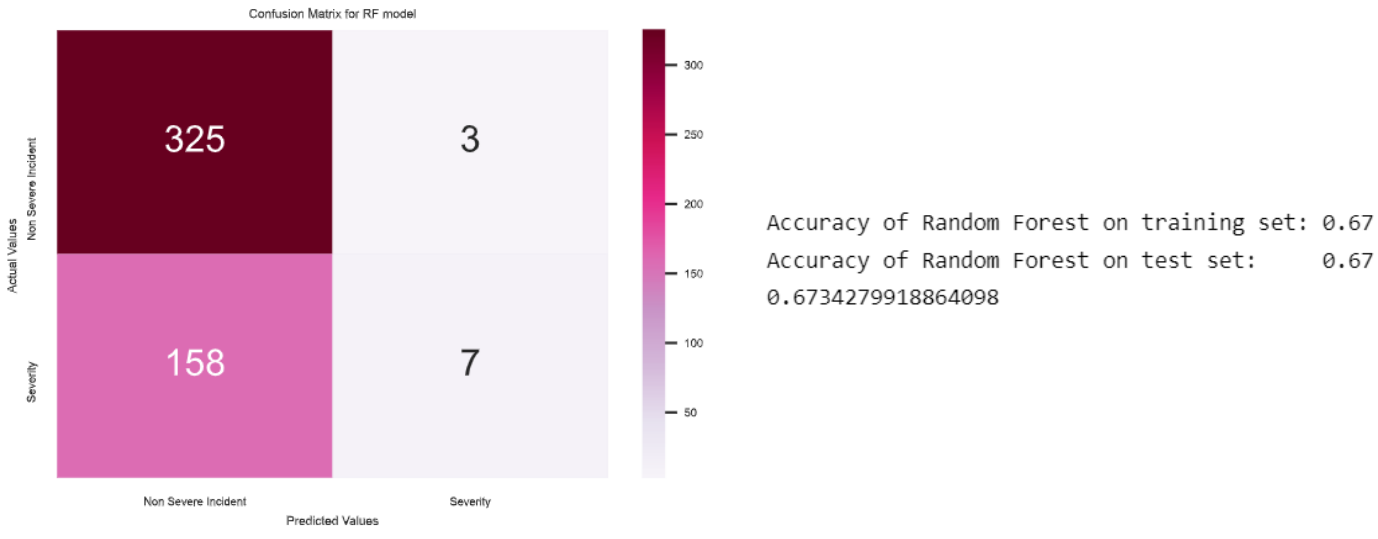


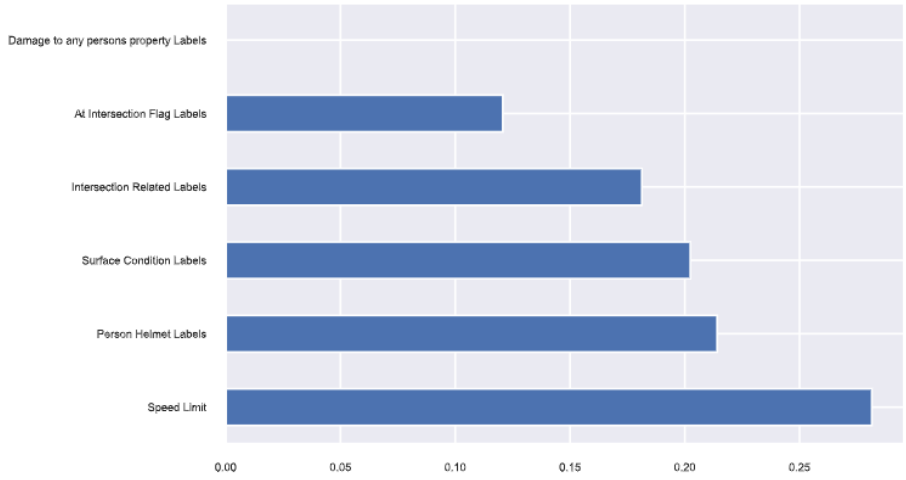
Figure 18. Confusion Matrix & Accuracy

Based on the confusion matrix generated, it is observed that **325 classes** have been correctly classified as **non-severe incident class** which is the most accurate classification obtained as compared to the Logistic Regression model and Decision Tree Classifier model. The severity class have been correctly classified only **7 out of 165 times**, which is not accurate if compared to DT model, as it was able to classify **15 out of 165 classes** correctly, which is half of what RF model classified. Hence, it is again needed to minimize the false negative values here as the incidents are predicted as non-severe despite them being in the severe category, which could prove dangerous and lead to even more accidents to occur for any wrong predicted value.

Classification report RandomForestClassifier(max_depth=4, n_estimators=5000, random_state=42):

	precision	recall	f1-score	support
0	0.67	0.99	0.80	328
1	0.70	0.04	0.08	165
accuracy			0.67	493
macro avg	0.69	0.52	0.44	493
weighted avg	0.68	0.67	0.56	493

Figure 19. Classification Report



	Feature	Importance
3	Speed Limit	0.281
5	Person Helmet Labels	0.214
4	Surface Condition Labels	0.202
2	Intersection Related Labels	0.181
1	At Intersection Flag Labels	0.121
0	Damage to any persons property Labels	0.000

Figure 20. Feature Importance

Here, the feature importance is considered for **Speed Limit, Person Helmet, and Surface Conditions** variables which are key aspects that need to be considered when monitoring the severity of the incidents. These parameters are related to the person riding where he/she needs to control the speed limit and wear helmet in order to avoid the accidents.

Recommendation

The various machine learning models implemented in order to predict the crash severity are Logistic Regression model, Decision Tree Classifier, and Random Forest Classifier. Based on the features selected and extracted for the model building phase, the features differed in each scenario for various models implemented which helped in determining the important features of the model and accuracy of the model. Thus, the significant features that are considered based on the implementation of the three models are **Speed Limit, Person Helmet, Surface Condition, Road Intersection, and Average Daily Traffic Amount.**

In order to prevent dangerous incidents, the model was able to provide significant features that can be considered for the same. The recommendation based on the model results are such that the individuals should focus on the speed limit and wear a helmet which would reduce the count of accidents and also reduce the severity of the accidents to some extent. As observed, the speed limit if decreased and helmet worn could decrease the accident count which would help in prevention of dangerous accidents. Thus, here the features to be considered are **speed limit and person helmet.**

Also, the other important factors that can be considered with respect to the safety precautions on the road are related to **surface conditions, traffic amount, and road intersections**. These are some of the other parameters that could lead to an accident and thus to prevent it the road surface conditions needs to be checked and maintained depending on the weather conditions. The traffic amount is somewhat not completely in control, but measures need to be taken in order to have a strict traffic rule followed to prevent such severe accidents.

With respect to the model selection for the prediction of crash severity based on the three models implemented, Decision Tree and Random Forest Classifier performed best as compared to the Logistic Regression model as the accuracy of the **LR model is 66%** whereas **67% accuracy is obtained for Decision Tree and Random Forest model**. Not only the accuracy, but when considered the confusion matrix, **Logistic Regression** was able to predict **324 non-severe incidents correctly** and **2 severity classes correctly**, whereas **Decision Tree** predicted **15 severity classes correctly**, and **Random Forest** was able to predict **325 classes of non-severe incidents correctly**.

Thus, either Decision Tree Classifier or Random Forest Classifier can be considered for the prediction of the accident severity. Now, based on the classification of crash severity, it is important to **minimize the false negative values** as that could be dangerous in any situation, and thus **Decision Tree Classifier** which has 150 wrong classification of severity classes whereas Random Forest Classifier had 158 wrong classification, despite the correct classification for non-severe incident class is more in the case of RF model, can be considered as the model for prediction of crash severity as we need to minimize the false negatives more importantly.

Therefore, either Decision Tree or Random Forest model can be used as the accuracy obtained for both the models is same of 67%, but when the classification and minimization of false negatives is considered, DT model could be the best fit model for the same. The comparison of accuracy and confusion matrix for the machine learning models implemented are as follows which helps to decide the best fit model for prediction of crash severity.

Table 1. Accuracy Comparison of ML models

Machine Learning Model	Accuracy
Logistic Regression Model	66.13%
Decision Tree Classifier	67.34%
Random Forest Classifier	67.34%

Table 2. Confusion Matrix Comparison of ML models

Machine Learning Model	**Confusion Matrix**			
Logistic Regression Model	*Actual Values*		*Predicted Values*	
		Non-Severe Incidents	**324**	**4**
		Severity	**163**	**2**
			Non-Severe Incidents	Severity
Decision Tree Classifier	*Actual Values*		*Predicted Values*	
		Non-Severe Incidents	**317**	**11**
		Severity	**150**	**15**
			Non-Severe Incidents	Severity
Random Forest Classifier	*Actual Values*		*Predicted Values*	
		Non-Severe Incidents	**325**	**3**
		Severity	**158**	**7**
			Non-Severe Incidents	Severity

Conclusion

The city of Austin has heard many complaints from the cyclists regarding the protection from the motor vehicles. In order to confirm these complaints, the city complied data of cyclists related incidents that occurred in order to analyze the data and find if the data confirms the same. The objective here is to predict the type of accident occurred and recommend factors that are affecting the accidents to occur in order to prevent the dangerous accidents.

The crash severity column consists of several types of accidents which are then categorized into two categories namely, Severity and Non-Severe Incidents where Possibly Injury, Incapacitating Injury, and Killed are categorized into Severity class and Non-Incapacitating Injury, and Not Injured are categorized into Non-Severe Incident. Based on the feature selection and extraction and the correlation matrix, certain features were considered for the implementation of model whereas some of the features were dropped in order to avoid the issue of multicollinearity.

The significant and important features for each model were observed which could help in prevention of the dangerous accidents to occur. Some of the important features considered and

recommended are speed limit, and person helmet at an individual level, & surface conditions, traffic amount, and road intersection with respect to precautions to be taken on the road. These features considered would help in preventing any dangerous accidents or crash severity to happen.

The models implemented for prediction of the crash severity performed well with an overall accuracy of 66-67% where Decision Tree and Random Forest Classifier performed best with an accuracy of 67% and were able to correctly classify majority of the class. The main goal here is to minimize the false negative values as the model has incorrectly classified severe accidents to non-severe accidents which can be dangerous leading to even more incidents to occur. Since decision tree model could correctly classify 15 of the classes into severe class, minimizing the false negative values as compared to logistic regression and random forest model, the decision tree model can be used for the implementation and prediction of crash severity.

Thus, an individual needs to limit the speed and wear a helmet in order to prevent dangerous accidents and the road precautions such as surface condition, traffic amount, and intersection road needs to be maintained and checked as these are the factors affecting the crash severity. Lastly, Decision Tree Classifier can be used for the implementation and prediction of the severity of the accidents for the city of Austin.

References

Bedre, R. (2021, January 3). Logistic regression in Python (feature selection, model fitting, and prediction). Data Science Blog. https://www.reneshbedre.com/blog/logistic-regression.html

Bedre, R. (2021b, March 13). Multicollinearity and variance inflation factor (VIF) in the regression model (with Python code). Data Science Blog. https://www.reneshbedre.com/blog/variance-inflation-factor.html

sklearn.tree.DecisionTreeClassifier. (2022). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

sklearn.ensemble.RandomForestClassifier. (2022). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html