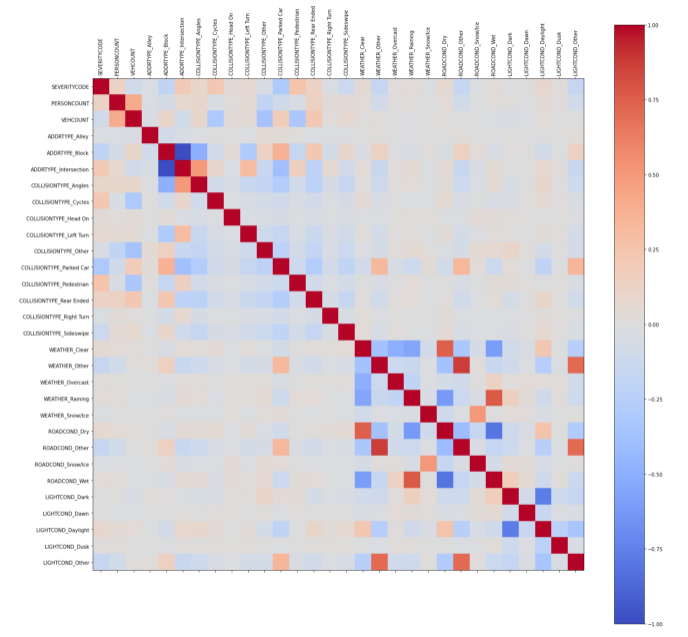
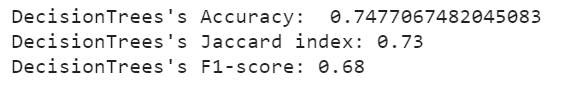
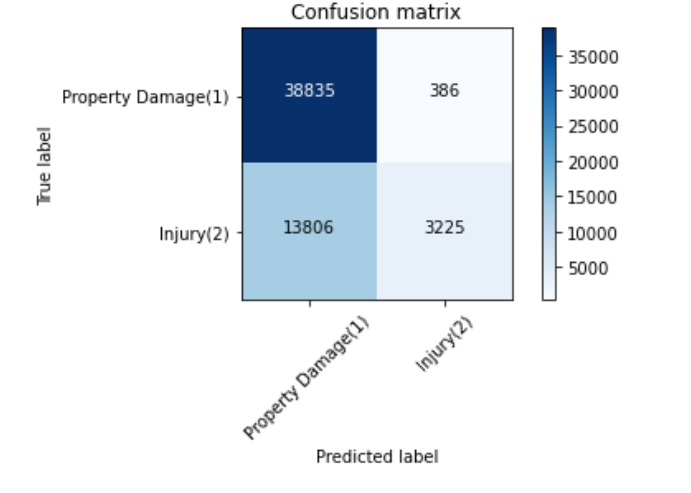
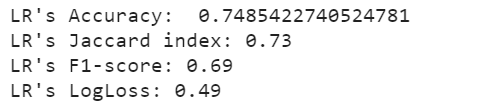
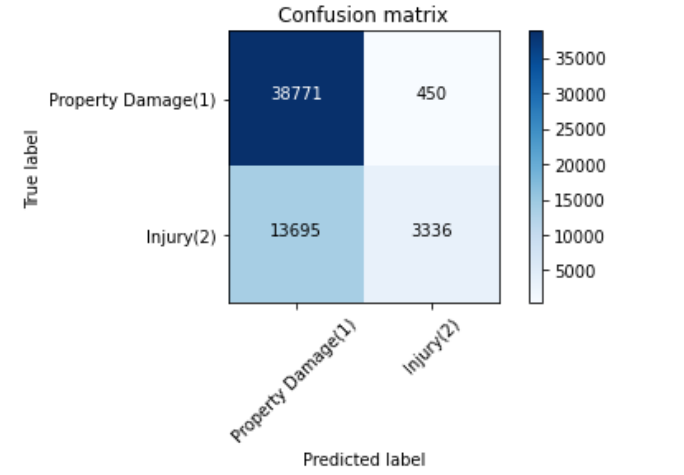
1. **Introduction / Business problem**
2. Vehicle accidents differ in severity. First responders to vehicle accidents include both emergency first aiders, fire fighters and the police department. In cases of accidents involving injuries, the duration and composition of such first responders can determine the survivability of the people involved in the accident. However, the departments providing first response to accidents usually have limited resources and manpower. As such, the committed resources and personnel being deployed to respond to an accident should comprise the optimal composition so that it does not unnecessarily use up such resources, which could be better utilised by being on standby for other accidents. Data can help to predict the severity of an accident being reported such that the departments will know how many people and what type of personnel may need to be involved in responding to the accident.
3. **Description of data**
4. The dataset consists of vehicle collisions reported and compiled by the SPD and the Traffic Department of Seattle from 2004 to present.
5. The data set consist of 38 columns and 194673 rows of data. The full list of attributes and what they represent can be found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).
6. The data contains the severity of the accident (i.e. SEVERITYDESC), which is the target variable ad has been coded in the data set (i.e. SEVERITYCODE).
7. The dataset contains other attributes that are commonly reported or collected in the course of an accident report, such as the type of junction that the accident took place (i.e. ADDRTYPE), the type of collision (i.e. COLLISIONTYPE), the number of people (i.e. PERSONCOUNT), pedestrians (i.e. PEDCOUNT), bicycles (i.e. PEDCYLCOUNT) and vehicles (i.e. VEHCOUNT) involved, the weather (i.e. WEATHER), road (i.e. ROADCOND) and lighting conditions (i.e. LIGHTCOND), as well as whether the accident involves a parked vehicle (i.e. HITPARKEDCAR). Such attributes can be used to model and predict the severity of an accident being reported.
8. **Methodology**
9. In an initial accident report, the commonly reported observations are likely to be the junction where the accident took place, the type of collision, how many vehicles appear to be involved, the weather, road and lighting conditions. The number of people who are involved are also likely to be data that is reported.
10. As such, we extract the columns "SEVERITYCODE", "ADDRTYPE", "COLLISIONTYPE", "PERSONCOUNT", "VEHCOUNT", "WEATHER", "ROADCOND", "LIGHTCOND" into a dataframe.
11. We then identify missing data from this dataframe. Since the number of missing data is about 13k rows (i.e. about 6.7% of total number of rows), we proceed to drop the rows with missing data.
12. We then perform some basic description of the dataframe, including a value count for each column. We simplify the values for columns "WEATHER", "ROADCOND", "LIGHTCOND" into values that will usually be raised in an initial report.
13. We then proceed to obtain dummy variables for columns "ADDRTYPE", "COLLISIONTYPE", "WEATHER", "ROADCOND", "LIGHTCOND".
14. We proceed to compute a correlation matrix on the dataframe, and found that ROADCOND\_Other s highly correlated to WEATHER\_Other. Hence we dropped ROADCOND\_Other from our dataframe.
15. The features and target variables are then coded into a X and y dataframe, with a MinMax scaler being used to preprocess the values in X, given the different values for PERSONCOUNT and VEHCOUNT. MinMax scaled would not change the values for the binary options for the dummy variables.
16. We the split the dataset into training and testing sets, and build a machine learning prediction model using the Decision Tree and Logistic Regression algorithm to the training data. We test the models on the test data set to predict whether an accident will involve property damage only or also personnel injury. We then compute accuracy, Jaccard index, F1 and LogLoss scores for our models.
17. Finally, we draw a confusion matrix to visualise the true and predicted outcomes using our model.
18. **Results**
19. The following figures shows the scores and confusion matrix for our Decision Tree model.



1. Accuracy score and Jaccard index appear to be reasonably high, but F1 score does not appear as high. Looking at the confusion matrix, the model appears to be able to predict well when it is a property only damage. However, the model appears to be weak in predicting when personnel injury is involved, with a majority of predictions being property only damage. This could have led to the lower F1 score.
2. The following figures shows the scores and confusion matrix for our Logistic Regression model.



1. Again the scores and confusion matrix are quite similar to the Decision Tree model, where the Logistic Regression model is able to predict well when there is a property damage, but is weak at predicting true personnel injury.
2. **Discussion**
3. We set out to build a model to be able to predict the severity of an accident being reported such that the government departments will know how many people and what type of personnel may need to be involved in responding to the accident.
4. Our models were able to predict well those accidents which only involve property damage, but was weak in predicting those accidents involving injury to people. As it stands, the models may not be as useful in ruling out those accidents where there may not be any injury to people such that medical staff would not have to move out to attend to such accidents and can be on standby for other accidents which involve injury.
5. We observe that in the dataset, there were many more instances of property damage type of accident as compared to injuries, which may have skewed the dataset. Possible improvements may be to balance out the number of accidents involving property damage only and those involving injury, for example by randomly replicating the data which shows that the accident involves personnel injury. This may increase the precision of the model in predicting when an injury occurs from an accident.
6. The training and test set split is also a static one. Possible improvement will be to utilise cross validation so that the model has more data to train on.