**Predicting Crash Injury from Seattle Traffic Accident Data**

**1. Introduction**

**Background**

There have been many traffic accidents causing injuries or property damage in Seattle between 2004 and 2020. The data set for this capstone has the severity of accidents as either property damage or injury. I am going to create a predictive model to inform drivers of the important factors that contribute to injuries caused by traffic accidents. The objective of the model is to inform drivers so they can make safer driving decisions and reduce their chances of being involved in a traffic accident causing an injury.

**Problem**

When accidents occurred in Seattle between 2004 and 2010 there were many factors recorded at the time of the crash such as, road conditions, weather, light, speed, people involved and the date. All of these environmental factors can be used to determine to predict traffic accidents that are more likely to cause injury. This project aims to predict if a traffic accident will result in an injury or property damage.

**Interest**

**There are many different groups of people that would be interested in knowing what factors are more likely to cause injury. For example, drivers would like to know which situations are most likely to cause injury so they can drive more safely. The auto industry would be interested in this model so they can improve the safety of their cars. Insurance companies could use this data to identify areas of higher risk. Also local municipalities could use this data to make drivers more aware of higher risk areas by adding signage or lighting to reduce traffic accidents.**

**2. Data acquisition and cleaning**

**Data Source**

The traffic accident data set is a .csv file, the target variable to predict is an attribute with labeled data indicating the severity of the traffic crash as either property damage or injury. There is location indicating the latitude and longitude of the accidents as wells as collision type, the number of people involved, pedestrians, bicycles and vehicles involved as well as the date and time of the crash. The environmental factors could be important in predicting accident severity by including the road conditions for example was the road icy or wet or were the light conditions dark. There are environmental information available such as the weather, road and light conditions and if any of the drivers involved were under the influence.

**Data Cleaning**

There are two date attributes some have just the date while many others have the data and time. The time of the accidents could be an important factor so that attribute needs to be tested to confirm if including improves accident severity prediction. There are also entries that are blank in many attributes. The size of the original data set was 73 attributes with 194,673 rows. I decided to keep only rows of data that had complete data set so I removed rows where there were no important environmental attributes.

The date and time could be important factors in determining the severity of traffic accidents, so I split the date time column into day of the week, hour and month the accident occurred. I split this information out so machine learning models could use the newly created attribute.

By eliminating the empty cells of potentially important environmental data, the data set was reduced by 25%. The removal of empty cells did not affect the frequency rate that injury occurred compared to the original set. The frequency of injury was 30% of all traffic accidents in the original data set and after removing the empty cells, the frequency of injuries in the remaining accidents was still 30%. The final data set ended up with 155,403 rows compared to 194,673 rows in the original set.

Additional cleaning occurred where some important features had blank columns but the correct data could be inferred. For example the under the influence was assumed to be ‘N’ since officers would record is someone was under the influence with a ‘Y’ other wise a blank would indicate the individual was not under the influence. In a similar fashion, the speeding, distracted driving and pedestrian right of way features were also filled in with ‘N’ if the data was left blank.

**Feature Selection**

After cleaning the data there were 37 attributes in the data set. Some of the 37 attributes were not related to any environmental or driver related factors that could have influence on the severity of traffic accidents. I removed these features such as the OBJECTID or INCKEY which are purely unique identifiers.

There were also some columns that had duplicate or unnecessary information. For example there were 2 date columns, one had just the date, the other had the date and time. I removed the column that only had the date and kept the date time column but parsed out the month, day and year for machine learning purposes. Also there was a column that had the crash descriptions as a string but the same information was included as a numerical category with each number as a key code describing where the collision occurred in reference to the vehicle. Below is a table showing the feature selections.

Table 1. Features removed during data cleaning.

|  |  |
| --- | --- |
| Dropped features | Reason for dropping |
| ObjectID, inckey ReportNo, Status, Addrtype, Location, SeverityDesc, SDOTcolnum | Clerical information not related to environmental |

**3. Exploratory Data Analysis**

**Balancing the Data Set**

Since injury related traffic accidents occur in only 30% of the total traffic accidents, in an unbalanced data set machine learning models could always choose property damage only and achieve roughly 70% accuracy but would never choose an injury related crash. For this reason I created a balanced data set by down-sampling the majority class which was the property damage class. This was done by first splitting the data into injury or property related crashed. The property damage had 106,764 rows and the injury related set had 48,639 rows.

I then down sampled the majority by randomly choosing with replacement 48,639 rows from the property damage data set and then combined the 2 data sets. This resulted in a final data set with 97,278 rows and 22 columns. Half of this balance data set was property damage, and the other half was injury related crashes.

**Response Vector**

There are many categorical variables that needed to be converted from a string to a numeric category for the machine learning algorithms to process the information. There were X categorical features that needed to be converted. They were INATTENTININD, SPEEDING, WEATHER, JUNCTIONTYPE, ROADCOND, LIGHTCOND, PEDROWNOTGRNT. Some of these features had many categorys which were all converted to numeric categories by importing the preprocessing module from sklearn.

**4. Predictive Modelling**

Predicting if a crash will result in either an injury or only property damage is a classification problem and as a result, I chose decision trees, logistic regression and SVM algorithms to determine which was the better predictor. For each algorithm I used train test split to train the models on 70% of the data and test the models on 30% of the data. I used the same train test split and random state for each algorithm.

**Performance of Different Models**

The table below shows some the models and different evaluation metrics I chose. Many of the algorithms had similar accuracy and F1 scores but the decision tree was the highest in both categories so I chose this model as the most predictive.

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | F1 Score |
| Decision Tree | 64% | 66% |
| Logistic Regression | 63% | 60% |
| SVM ‘rbf’ | 63% | 65% |
| SVM ‘linear’ | 63% | 62% |
| SVM ‘poly’ | 63% | 62% |

**5. Conclusions**

In this study I analyzed traffic crash data in Seattle from 2004-2010. I balanced the data set so the algorithms could predict either property damage or injury crashes. I built 5 classification models and chose decision tree due to the higher rank on F1 score and accuracy compared to the other models.