Predicting car accident severity

1. Introduction

1.1 Background
Road accidents result in both injuries, material damage and in the worst-case loss of life. These accidents could in many cases be prevented if more caution was exercised. By understanding what the factors are that contribute to road accidents and their severity many accidents may be avoided through preventive measures.

1.2 Problem
Despite efforts to make roads and cars safer the number of accidents is still too high. Legislators, transportation agencies and car manufacturers all try to make traffic safer for all. However, in order to effectively prevent road accidents there is a need to understand the factors contributing to accidents. By utilizing data science to better understand what factors and their importance in contributing to accidents we can help reduce the number of accidents in the future. There will be common factors that correlate and contribute to road accidents more than others. By identifying these a model for preventing future accidents can be developed.

1.3 Interest
Legislators, transportation agencies and car manufacturers will be interested in understanding what factors that contribute to accidents and determine their severity in order to make their preventive work more efficient. It could also in the interest of drivers to understand what factors increase the likelihood of accidents in order to improve their driving and take precautionary measures when faced with accident prone situations.

2. Data acquisition and cleaning

2.1 Data source
I used the shared dataset available Clicking here. The dataset includes all types of collisions, factors which can be included in the machine learning model and the accident severity. The dataset contains many observations and various attributes (38 columns and 194673 rows), which makes it good for solving our problem. The collisions are provided by SPD and recorded by Traffic Records. This includes all types of collisions and will display at the intersection or mid-block of a segment between 2004 to present.

2.2 Data cleaning
There were some missing data. I replaced NaN data values and cleared empty columns. The data was also cleaned by removing excess columns. Furthermore, I also ensured all data types of each data was correct.

2.3 Feature Selection
There were many features in the dataset that are not relevant to determining the accident accident severity. However, the following features were selected for the machine learning algorithm due to their likely effect on accident severity.

ADDRTYPE - Collision address type:
-Alley

-Block
-Intersection

ROADCOND – the road condition when the accident occurred.

LIGHTCOND – the light condition when the accident occurred.

WEATHER – the weather condition when the accident occurred.

INCDTTM – time, later converted into day of week

4. Predictive Modeling

I use both KNN, Decision Tree, and Logistic Regression to classify new conditions as either 1 or 2 in severity code. Then I evaluate the model based on the evaluation metrics Jaccard, and F1 Score to determine the accuracy.

4.1 KNN

I determine the best accuracy with K=. Then compare test and training and after that I evaluate KNN with Jaccard and F1 Score. Model accuracy:

```
Train set Accuracy:  0.6932758000719166
Test set Accuracy:  0.6866545666244307
Train set jaccard similarity score:  0.6932758000719166
Valid set jaccard similarity score:  0.6866545666244307
Train set f1 score:  0.6474627015646283
Valid set f1 score:  0.6408401619312328
```

4.2 Decision Tree

I determine the decision tree accuracy. Then compare test and training and after that I evaluate the decision tree with Jaccard and F1 Score. Model accuracy:

```
DecisionTrees's Accuracy:  0.6945195940778284
Train set Accuracy:  0.6932758000719166
Test set Accuracy:  0.6866545666244307
Train set jaccard similarity score:  0.6932758000719166
Valid set jaccard similarity score:  0.6866545666244307
Train set f1 score:  0.6474627015646283
Valid set f1 score:  0.6408401619312328
```

4.3 Support Vector Machine

I determine the decision tree accuracy. Then compare test and training and after that I evaluate the decision tree with Jaccard and F1 Score. Model accuracy:

```
Logistic Regression's Accuracy:  0.6099049119322397
Train set Accuracy:  0.6932758000719166
Test set Accuracy:  0.6866545666244307
Train set jaccard similarity score:  0.6932758000719166
Valid set jaccard similarity score:  0.6866545666244307
Train set f1 score:  0.6474627015646283
Valid set f1 score:  0.6408401619312328
```

4.4 Logistic regression

I determine the logistic regression accuracy. Then compare test and training and after that I evaluate the logistic regression with Jaccard and F1 Score. Model accuracy: 61%

```
Logistic Regression's Accuracy:  0.700991975160668
Train set Accuracy:  0.6932758000719166
Test set Accuracy:  0.6866545666244307
Train set jaccard similarity score:  0.6932758000719166
Valid set jaccard similarity score:  0.6866545666244307
Train set f1 score:  0.6474627015646283
Valid set f1 score:  0.6408401619312328
```

5. Conclusions

Based on the models it is clear that these factors influence the car accident severity and the model has clear predictive value. Logistic regression shows a model accuracy of around 70%. However, the model accuracy could be higher, and more research is therefore needed in order to determine the best factors to predict car accident severity.