# Predicting Car Accident Severity

Korede Adegboye

September, 24 2020

## 1 Introduction

### 1.1 Background

Large amounts of accidents are reported each year. While those accidents are taking place, traffic jams occur. People tend to get frustrated and are presumably late for work. In addition, their gas usage increases. These occurences leave drivers wondering, if there is something in place that will enable drivers to drive more carefully or even change their travel. The importance of safe driving is recognized by the government, individuals and insurance companies. News channels keep drivers updated about traffic in major roads and the current weather conditions. Therefore, it is advantageous for all drivers to accurately predict whether an accident will occur during their duration on the road.

### 1.2 Problem

In this problem we will try to detect collisions that occur around the city. We are particulaly interested in the possibility of individuals getting into a car accident and how severe it would be. Furthermore, the aim is to enable drivers to drive safely given certain variables or suggest a different route of travel. Data that might contribute to predicting car accident severity, includes weather, road conditions, age of driver and speeding.

### 1.3 Interest

This report targets those driving around the cities of developing countries. It is also of interest to vehicle insurance companies, and fraud detection in regards to false claims. In addition, companies in relation economic costs such as medical, insurance administration, and legal and court. Lastly, car manufactures as they can develop many safety meausres in their car.

# 2 Data

## 2.1 Data sources

All collision data is provided by the Seattle Polic Department and recorded by Traffic Records. The time for the data is from 2004 to present, and is updated weekly. The data was download from the Courseras Applied Data Science Capstone, but can also be scraped from data-seattlecitygis.opendata.arcgis.com for the most recent verision. In total there are, 194,673 rows and 38 features in the raw data set.

## 2.2 Data cleaning

There were a couple problems with the dataset that would make our analysis and results difficult. First, the missing values with dealt with. There was less than 5% rows with incomplete entries. Therefore, those rows were dropped.

Second, irrelevant data existed across features. Those being 'Unknown' and 'Other'. Here, dropping the rows with irrelevant entries results in less than 5% rows being removed. This measure is cumulative with dropping rows with missing data.

Third, the data SEVERITYCODE data was unbalanced. There was 136,485 entries belonging to class 1 and 58,188 belonging to class to. Respectively, class 1 and class 2 correspond to property damage and injury. To balance the data a downsampling approach was used. Where 58,188 were randomly sampled from class 1.

## 2.3 Feature selection

An informal approach to feature selection was used. Given our problem, we are trying to minimize the amount of features needed to predict car accident severity. Intuitively, the features selected were weather, light conditions and road conditions. These features are easily detectable and verifiable by drivers. Thus, our data set is now left with 3 features used to predict the target (severity). Furthermore, the features were converted from categorical to numerical values. Whether the data was normalized or not did not make a difference. Due to the fact that the ranges of numerical values are close to eachother. Thus, for easy interpretability the features are not standardized.

# 3 Methodolgy

## 3.1 Exploratory Data analysis

We noticed that majority of the accidents occur during clear weather, daylight and dry road conditions. This goes for both property damage and injury. Thus, this indicates moderate amounts of entropy between the classifications of severity(1 and 2).

Figure 1: Count plot of the severity reported in each weather label

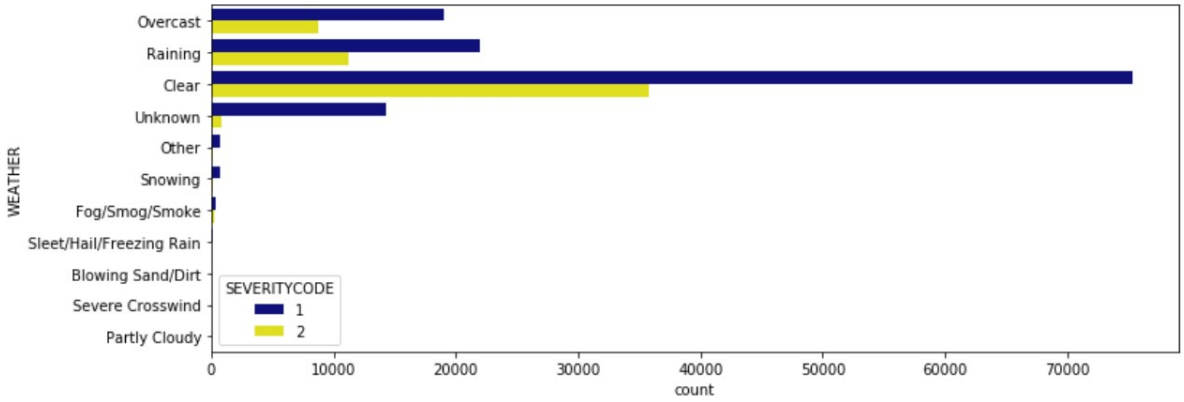



Figure 2: Count plot of the severity reported in each light condition label

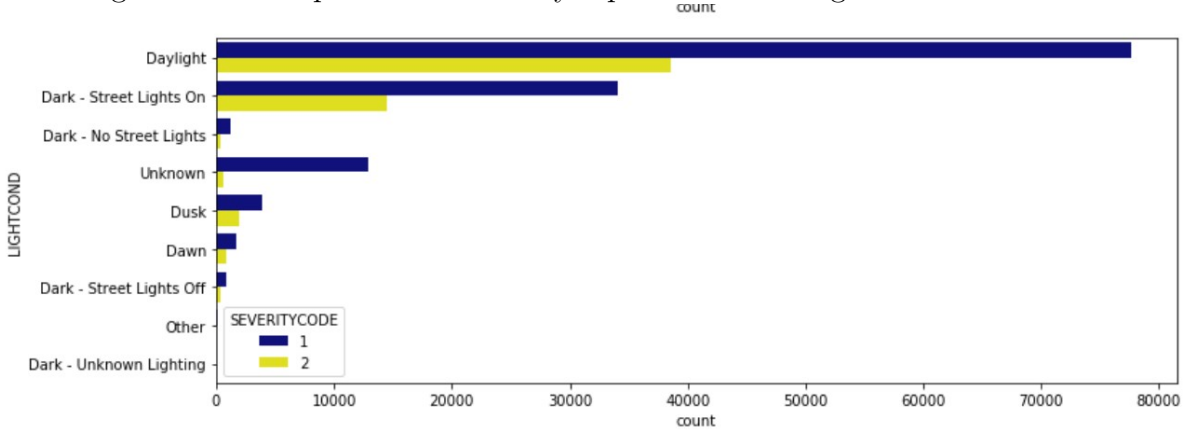
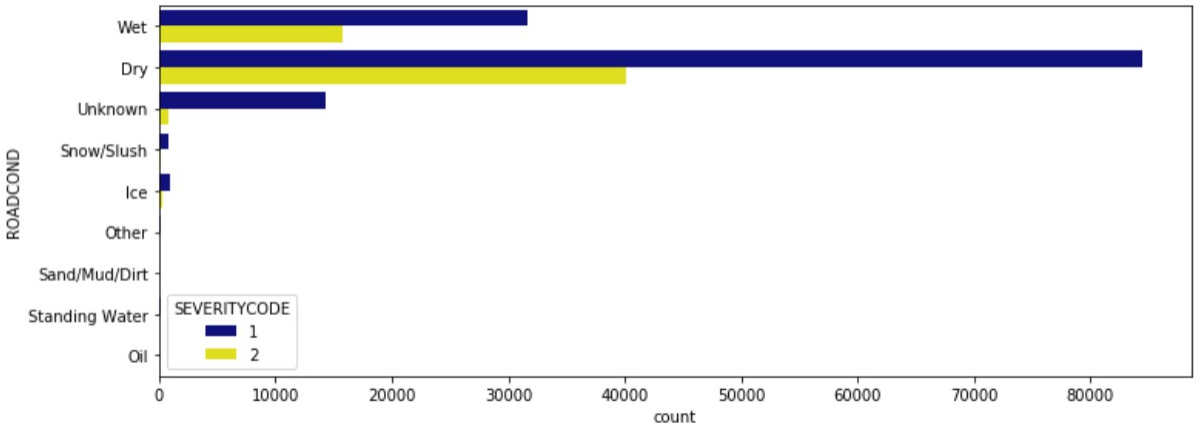


Figure 3: Count plot of the severity reported in each road condition label

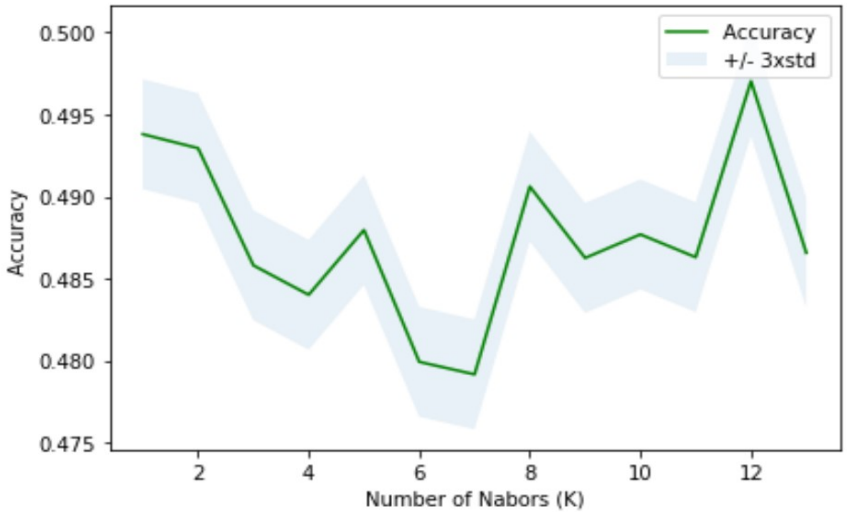



## 3.2 Predictive Modelling

The features were strings and thus, categorical variables. In addition, the we are trying to classify labeled data. Therefore, this is a supervised learning problem and a classification approach is needed.

K-Nearest Neighbors, classifies classifies the severity of the accidents based on similar cases. To find the optimal k, I produced a plot, showcasing accuracy against the amount of neighbors. The best was attained with 12 nearest neighbors.

Figure 4: Accuracy vs Number of Neighbors, K



Decision Trees, will map out all possible decision paths to predict the severity of an accident. Similar to K-Nearest Neighbors a plot was produced. Here, instead of K, we are searching for the optimal D. Meaning max depth of the tree. The plot shows that a depth of 6 yields the best result. Note that a max depth of 4 results in a ccuracy very close to that of 6. Therefore, max depth of 4 was chosen for the Decision Tree model. Taking a closer look at the decision tree, the entropy remains high throughout. This indicates significant concern in the model, and may not be the best classification approach.

Figure 5: Accuracy vs Max Depth, d

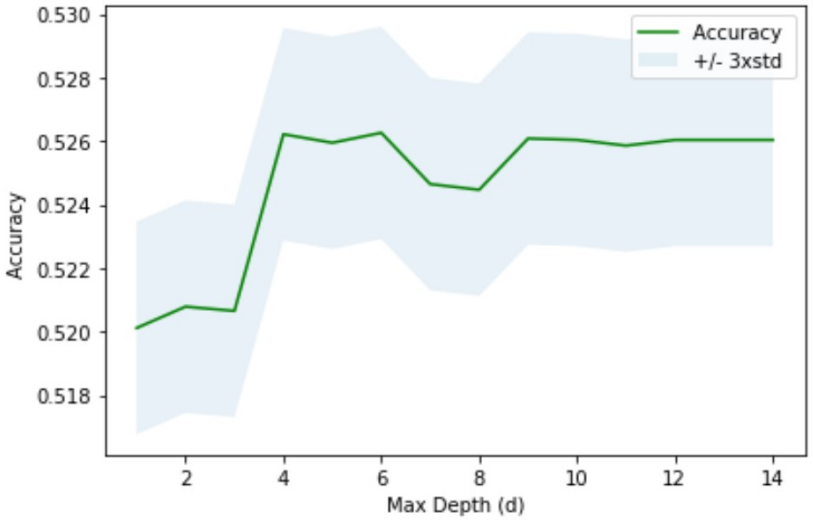
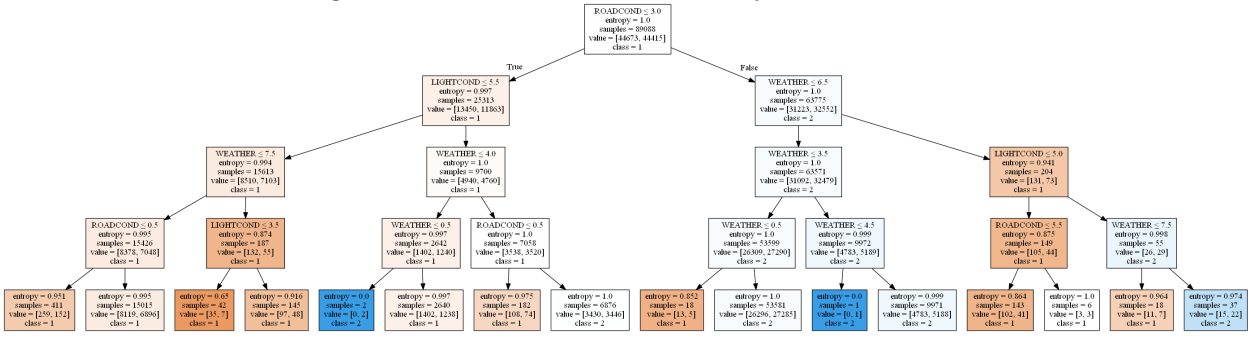



Figure 6: Decision Tree of severity classification



Support Vector Machines, classifies cases by trying to find a seperator. Preferably, non-linear kernels would be best for our model. Those being, radial basis function, polynomial and sigmoid. Note, support vector machines are computationally expensive, as they are highly dependent on the size of the data set. Therefore, this is some cause for concern when new data is frequently being added. Furthermore, a linear kernel was chosen for the support vector machine classifier

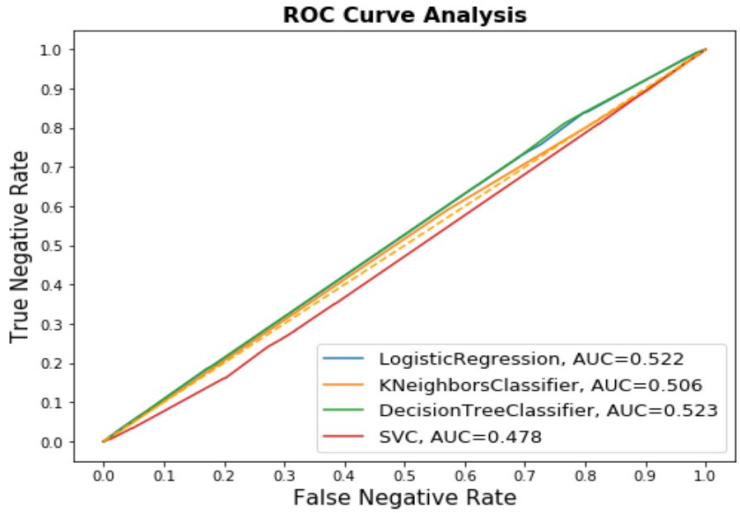
Logistic Regression, predicts binary labels. Here, we only have 2 classes for severity in our data. Note, it is possible to have multiclass labels since the data documentation indicates that the SEVERITYCODE can be 1,2,2b,3 or 4. Presently, logistic regression is a good apprach to try out. The logistic regression model, has different solvers. All of which, resulted in the same outcome. Logistic regression is computationaly faster, and works well with binary data. Note, if more Severity codes appear in our dataset it is best to go with SVM.

# 4 Results

Table 1: Performance of classification models. Best performance labeled in red

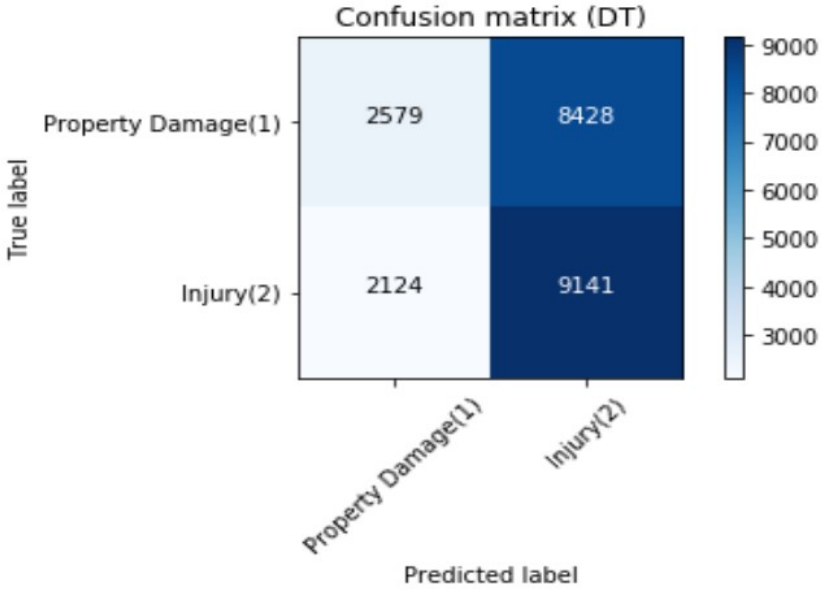
| | K-Nearest Neighbors | Decision Tree | Support Vector Machine | Logistic Regression |
|---|---|---|---|---|
| Jaccard | 0.5106 | 0.5262 | 0.5261 | 0.5199 |
| F1-score | 0.5075 | 0.4830 | 0.4856 | 0.4986 |
| Log-loss | 0.7400 | 0.6917 | 0.8091 | 0.6921 |
| No. of True Positives | 4733 | 2614 | 3372 | 3434 |
| No. of False Positives | 6274 | 8393 | 7635 | 7573 |
| No. of False Negatives | 4626 | 2158 | 3054 | 3119 |
| No. of True Negatives | 6639 | 9107 | 8211 | 8146 |

Figure 7: A section of ROC curves of different classification models




The ROC curve analysis shows us that the decision tree clasifier is the best model. In this problem we are more interested in lower false negatives than higher true negatives. The reason being that, it is important to be sure that an accident involving injuries occurs as predicted rather than predicting one with property damage. Furthermore, out of the models, decision tree also has the smallest log loss. Note if it were the case that lower false positives than higher true positves were of interested, then K-Nearest Neighbors would be the best model.

Figure 8: Confusion matric for Decision tree clasifier



# 5 Discussion

Note, most of the accidents appear to occur during clear weather, daylight and dry roads. This implies that drivers are less careful during these conditions, and seem to be more careful otherwise. Thus, there must be other factors besides the ones used that impact the possibility and the severity of an accident.

# 6 Conclusion

Our analysis has showed that past weather, light condition and road condition data are somewhat significant when predicting future car accidents. Thus, our model somehwhat determines if an accident will occur in Seattle. Our models have room for improvement. Other features will need to be considered. Those features may include, junctiontype, addtype and a description of the collision (whether it was a vehicle to vehicle or vehicle to bike accident). In addition, a map displaying the location of the accidents would help in identifying where most of the accidents occur. Further, enabling drivers to be more cautious when they are in the specific area. In regards to junctiontype and addtype, these features may be of interest to every driver, but specifically to driving instructures. These instructors can formulate better programs based on our insights.