PREDICTING CAR ACCIDNET SEVERITY

10/19/2020

Dana Andrea

Table of Contents

1. Introduction

1.1 Background

First responders (firefighters, policeman, ambulance drivers, paramedics, EMTs) are all professions that are the first to arrive to the scene of a traffic accident. However, first responders and their stations are not always in the optimal location to reach the scene of the next accident. The longer it takes first responders to get to an accident, the more risk there is to those involved in the accident and to the first responder themselves.

For example,

- Firetruck accidents rank as the second leading cause of on-the-job deaths for firefighters.
- Approximately 500 firefighters are involved in fatal firetruck crashes every year; on average, 1 in 100 of those occupants die as a result of the crash.
- Between 2006 and 2016, more than one police officer per week was killed on average from a collision or from being struck directly by another vehicle.
- More than 10,000 ambulance-related collisions occur annually; from 1993 to 2010, approximately 97 EMS technicians were killed in collisions.

Source: https://insurance.glatfelters.com/first-responder-safety-roadside-safety

1.2 Project Goal

The goal of this project is to help minimize the risk posed to first responders on their way to an accident. By using machine learning categorization and prediction models, we can discover the most common accident locations and conditions. This information can be used to optimally staff and position first responders so they can minimize travel risk on the job.

1.3 Project Goal & Target Audience

The target audience for this project are the government entities and policymakers who are interested in optimizing their emergency services regarding car accidents. This can also be used within the emergency response service entities to optimize their own response time.

2. Data acquisition and cleaning

2.1 Data Sources

The data used in this project contains 194,673 records of motor vehicle accidents and corresponding weather conditions from Seattle, WA. The date range for these accidents is from 2004-2020. This dataset was provided by Coursera.

The data set includes key fields like accident severity (personal or property damage), coordinates of the accident, weather on the day of the accident, time of the accident, collision type, junction type, and number of persons involved. This information can be extracted from the raw data and used to find the most common accident conditions that cause severe

personal injury and provide a recommendation for where first responders should be located/focus patrols to reduce their travel time to an accident.

2.2 Data Cleansing

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

This car accident data a few basic data inconsistencies that first need to be remedied. First, the excess identifier columns will need to be removed as these refer to internal governmental agencies and have no bearing on the prediction of accident severity. Second, all Yes/No columns need to be converted to Boolean values. Third, all NaN data needs to be converted to 'Unknown' category.

Boolean Attributes:

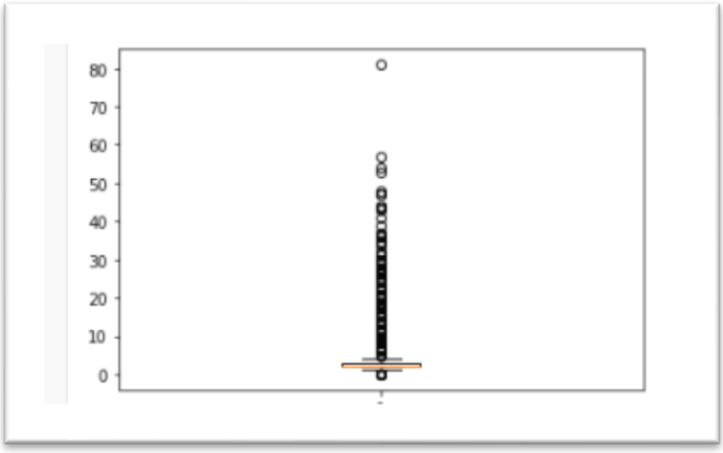
- INATTENTIONIND
- SPEEDING
- HITPARKEDCAR
- PEDROWNOTGRNT
- UNDERINFL
- SEVERITYDESC

Attributes where blanks were converted to 'Unknown':

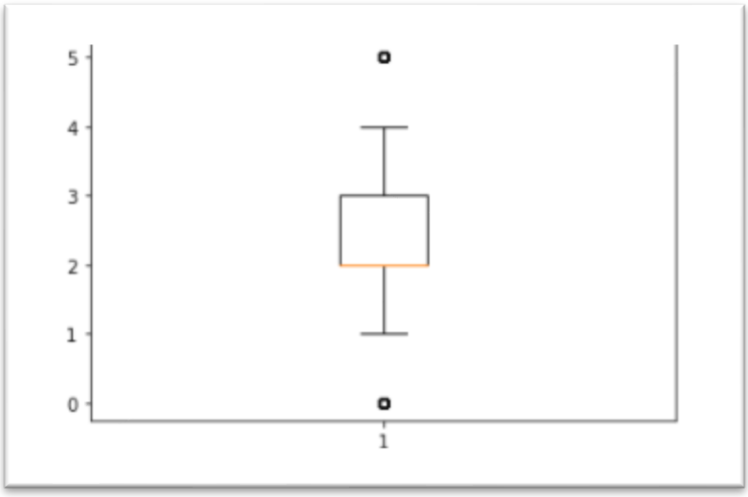
- COLLISIONTYPE
- WEATHER
- ROADCOND
- LIGHTCOND
- ST_COLDESC
- JUNCTIONTYPE
- ADDRTYPE

Now that basic data cleansing is completed, some more difficult data issues need to be addressed before analysis can begin.

First, outliers from the PERSONCOUNT attribute need to be removed. This attribute plays an important role in predicting accident severity. However, the mean number of persons involved in an accident is 2.4. The 95^{th} percentile of persons involved in an accident is 5. However, there are accidents involving up to 81 people in this data set. These need to be removed as they are extreme cases and will not aid in prediction.



PERSONCOUNT with outliers.



PERSONCOUNT without outliers.

Next, the following columns need to be hot encoded so they can be used in analysis: ADDRTYPE, SEVERITYDESC, COLLISIONTYPE, WEATHER, ROADCOND, LIGHTCOND. The resulting data set contains 70 columns.

The last thing to do before selecting features for the model is to balance the dataset. Data is unbalanced when one class has many more records than another. In the case of this car accident data, they are more than twice as many accidents with property damage (Severity 1) as compared to accidents with personal injury (Severity 2). Since the data set has so many valid records, it makes sense to under sample the property damage class (Severity 1).

```
In [91]: from sklearn.utils import resample

         # Separate majority and minority classes
         df_majority = hc_collisiondata[hc_collisiondata.SEVERITYCODE==1]
         df_minority = hc_collisiondata[hc_collisiondata.SEVERITYCODE==2]

         # Downsample majority class
         df_majority_downsampled = resample(df_majority,
                                            replace=False,    # sample without replacement
                                            n_samples=55406,     # to match minority class
                                            random_state=123) # reproducible results

         # Combine minority class with downsampled majority class
         df_downsampled = pd.concat([df_majority_downsampled, df_minority])

         # Display new class counts
         df_downsampled.SEVERITYCODE.value_counts()

Out[91]: 2.0    55406
         1.0    55406
```

As you can see in the output, bother Severity 1 and Severity 2 accidents both have 55,406 records. This will be sufficient to build our model.

Now that we have cleansed the data, we can perform some basic data analysis and select the features to be used in the model.

3. Methodology

3.1 Data Analysis

Weather Analysis

It makes sense to analyze the impact of weather on accident severity because intuitively, it is expected that as weather conditions worsen, accident severity increases. Let's take a look at the proportion of severity 1 and severity 2 accidents based on road conditions, weather conditions, and light conditions.

ROADCOND	Dry	Ice	Oil	Other	Sand/Mud/Dirt	Snow/Slush	Standing Water	Unknown	Wet
SEVERITY_1	47.5%	59.4%	48.7%	46.1%	46.3%	68.9%	50.0%	80.2%	46.3%
SEVERITY_2	52.5%	40.6%	51.3%	53.9%	53.7%	31.1%	50.0%	19.8%	53.7%

WEATHER	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	Unknown
SEVERITY_1	51.9%	47.4%	46.0%	71.2%	48.1%	25.0%	45.7%	53.8%	54.4%	66.0%	79.6%
SEVERITY_2	48.1%	52.6%	54.0%	28.8%	51.9%	75.0%	54.3%	46.2%	45.6%	34.0%	20.4%

LIGHTCOND	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	Other	Unknown
SEVERITY_1	60.2%	55.1%	50.4%	50.0%	45.8%	46.3%	45.7%	60.3%	80.0%
SEVERITY_2	39.8%	44.9%	49.6%	50.0%	54.2%	53.7%	54.3%	39.7%	20.0%

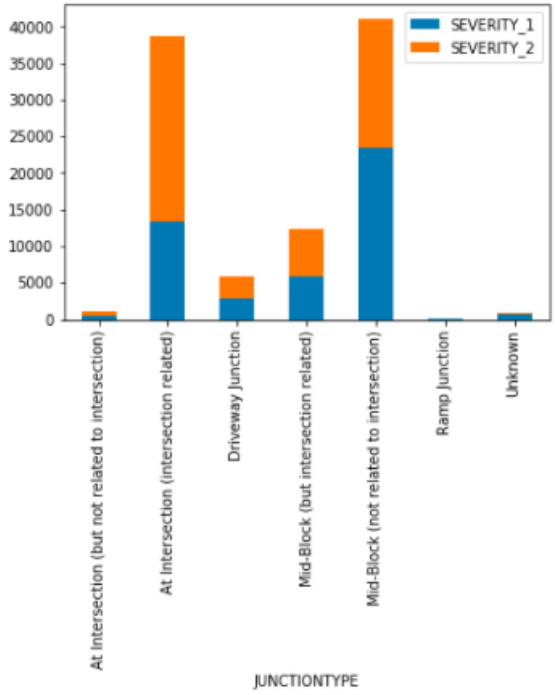
After considering the proportion of severity 1 and 2 accidents in different weather conditions, it appears that no weather condition heavily increases the likelihood of severity 2 accidents. The proportions for nearly all conditions are close to 50/50. This indicates that weather, road and light conditions do not actually influence accident severity. This is important to note as it is counterintuitive.

However, there are two notable exceptions to this observation.

The first is snow conditions show 60% severity 1 accidents and 30% severity 2 accidents. This indicates the opposite of what one may naturally assume. Snow conditions do not cause more severity 2 accidents. We might explain this by saying that in the snow, people are more conscious of how they drive or are less likely to drive at all. Perhaps the most severe accidents occur when people are least prepared for it - in clear and dry conditions.

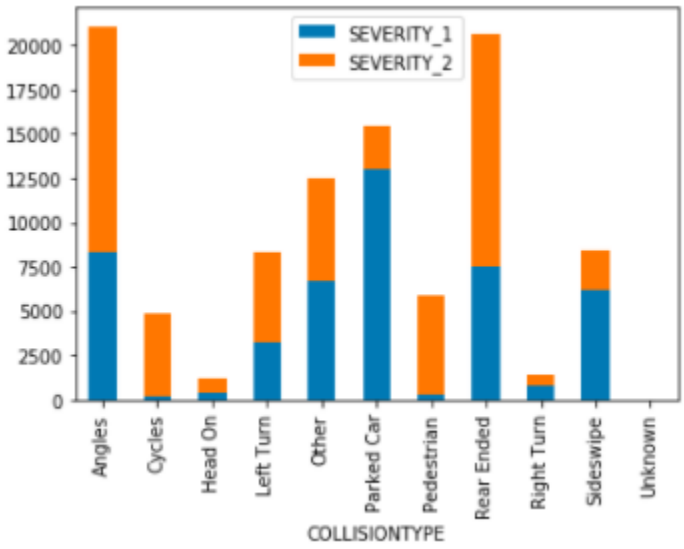
Secondly, it appears that "unknown" weather, road, and light conditions are 80% severity 1 accidents. Perhaps this is because the weather was so mild the officer forgot to capture it in the report. Therefore, an accident with weather unknown is more likely to be a severity 1 accident. However, since there is no way to confirm this, it is best to exclude "unknown" and "other" categories from consideration.

Junction Type Analysis



Most accidents occur at a midblock or at an intersection. However, the proportion of severity 2 accidents is slightly higher than 50%. Therefore, these junction types seem like a good predictor of accident severity.

Collision Type Analysis



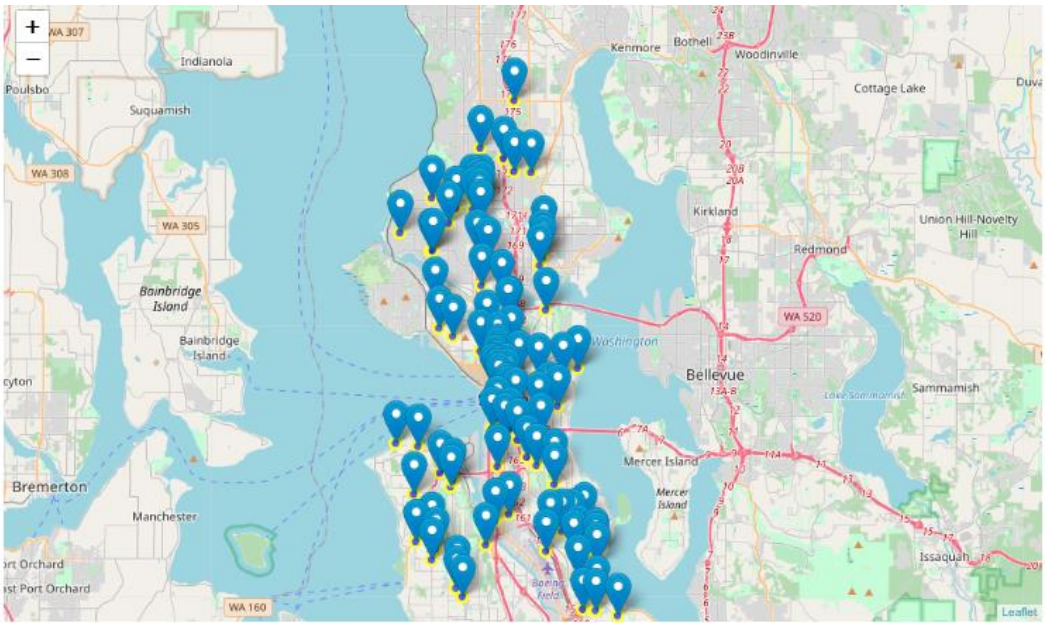
This graph gives us some good insights. It looks like nearly all accidents with a parked car, side swipes, and right turns are severity 1. This makes sense as they are most likely fender benders with cars moving at slow speed.

Read ended and angles show more than half of accidents are severity 2. So, these types of accidents are a good predictor of severity 2 accidents.

Lastly and most importantly, cycles and pedestrian have overwhelmingly severity 2 accidents. This makes sense because these most likely involve a car hitting a person or a bike. Therefore, these two categories are great predictors of severity 2 accidents.

Geospatial Analysis

Lastly, let's understand where on the map this data exists.



It looks like the accident data is centered around the Seattle area. While this information most likely will not change our modelling approach, this is good information to know because perhaps our model can be a good predictor of accident severity in other large US cities.

3.2 Machine Learning Model

Since we are trying to predict accident severity (class 1 or class 2), it makes since to use machine learning classification models so that, given the raw data of an accident, the model can classify the severity of that accident. We will use two machine learning models in this section. First will be K-nearest neighbor and second will be Logistic Regression.

Before we train the models, let's create a new dataframe which contains only those attributes which will be useful for the model and then define the X and Y. The X will be all attributes except for SEVERITYCODE, which is our predictor variable. Now, we set Y to the predictor variable.

```
In [23]: y = df_knn['SEVERITYCODE'].values
         y[0:5]

Out[23]: array([1, 1, 1, 1, 1], dtype=int64)
```

K-Nearest Neighbor

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. In this dataset we have sufficient domain knowledge and we don't expect predictions to be made frequently, but we do require high accuracy of predictions. These are all good reasons to use K-nearest neighbor, according to the IBM Knowledge Center article on the subject.

First, we split our data set into a training and test set. Next, we import KNeighborClassifier and train the model. Then we make predictions with the test set. Lastly, we import the metrics to get the accuracy of our model.

After optimizing the variable inputs and value of K, the accuracy of our model is 66%. This isn't terrible, but perhaps a different modeling technique would be better.

```
In [29]: from sklearn import metrics
         print("Train set Accuracy: ", metrics.accuracy_score(y_train, neigh.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat))

         Train set Accuracy:  0.6619509020127964
         Test set Accuracy:  0.6560565870910698
```

Logistic Regression

While Linear Regression is suited for estimating continuous values (e.g. estimating house price), it is not the best tool for predicting the class of an observed data point. In order to estimate the class of a data point, we need some sort of guidance on what the most probable class for that data point would be. Logistic regression is a good option when the predictor variable can only be two values. In our case, severity can only be classified as 1 or 2, so logistic regression is a good model to use.

First, we need to set up the test and training sets. Then we import LogisticRegression and train the model. Then we make predictions with the test set. Lastly, we import the jaccard_similarity_score to get the accuracy of our model.

The accuracy of this model is almost 70%, definitely an improvement on our K-nearest neighbor model.

```
In [33]: from sklearn.metrics import jaccard_similarity_score
         jaccard_similarity_score(y_test, yhat)
Out[33]: 0.6957461440220061
```

4. Results

After data analysis and modeling, we've determined the most important features that impact accident severity and can predict accident severity with 70% accuracy. The most important features in accident severity are Junction Type, Collision Type, Number of Persons involved, and the accident description.

The most interesting observation is that bad weather did not increase the likelihood that an accident would be severity 2.

Our K-Nearest neighbor model produced only 66% accuracy. This was a good model choice given the characteristics of the dataset. However, the accuracy was not as high as would be acceptable. Perhaps the accuracy in this model was lower because the different attributes used in predication are not necessarily related to one another. Therefore, it may be hard for the algorithm to compile the similarity measure.

Our logistic regression model produced 70% accuracy. This is certainly an improvement and sufficient accuracy for the first pass at the model. Perhaps this model had better accuracy because the target variable was binary.

5. Discussion

The insights gained from this model is that accidents involving pedestrians and cyclists at intersections and midblock are the most severe in terms of personal injury. For our first responders, this means that in order to reduce accident damage, first responders should be concentrated in locations where pedestrians and cyclists are common - like city centers, school zones, city parks, etc. In addition, first responders should be positioned in places that reduce drive times to these locations.

6. Conclusion

In conclusion, this was a successful first attempt at predicting accident severity using machine learning classification models.

To take this analysis farther, we would like to do the following things.

First, rerun the analysis using weather, road conditions, and light conditions as the only variable. The low impact these variables had on accident severity was interesting and it would be worthwhile to isolate that variable and understand what, if any, impact it has on accidents.

Second, we would want to run the same analysis as above but with a dataset for which the records contain all information in totality. A lot of the records used were missing some values, either weather related, geographical, or other. It would be interesting to see how the models performed on a data set with 100% available information.