Predicting the Effect of Accidents on Traffic

1. Introduction

1.1 Background

Road vehicle accidents are a global problem that effect lives in a profusion of ways like increase in traffic, causing harm to the people involved in the accident, damage to vehicles etc. Accidents are a result of a number of factors including poor road infrastructure and management, unenforced traffic laws, unsafe road user behaviour.

1.2 Problem

Data that might contribute to determining effect on traffic include information about the accidents that have occurred, like the temperature, the visibility, day/night time, weather conditions etc. that all contribute to affect the intensity of traffic.

2. Data acquisition and cleaning

2.1 Data sources

All the data required to solve the problem is provided in a single Kaggle dataset here: https://www.kaggle.com/sobhanmoosavi/us-accidents. This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to June 2020, using two APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset. This provides a large dataset but 50,000 entries were used which were enough to evaluate valuable results.

2.2 Data Cleaning

Data downloaded from the above link required cleaning. There were missing values, outliers but normalisation was not needed.

Since only 50,000 entries were needed, the dataset was shuffled first then the first 50,000 entries were chosen rest of the entries were discarded.
A few rows did not have values for some of the columns, hence all the rows having data points with null values were removed as null values adversely affect the prediction model.

First, the dataset had column containing the date and time of the accident. However, the date of the accident was useless, so was the minute and second the accident occurred. Only the hour of the day the accident occurred was required hence it was extracted and put into a new column and the old column consisting of date and time was removed.

Second, a column contained the zipcode of the area the accident occurred in but some of the entries included more than one zipcode. So the zipcode that occurred first in the entry was extracted and other zipcodes that followed were discarded.

Third, the feature containing the weather condition had too many classifications of the weather condition, for example some columns contained more than one weather condition, like, 'fair/clear'. Since the weather conditions were pretty much similar every weather condition after a '/' were discarded since they add no value to us. Followed by this, the unique weather conditions in the dataset showed clearly that they could be segregated in two distinct categories, clear weather or unclear weather. All the rows having 'fair' or 'clear' were considered as clear weather conditions, all the other rows having conditions like 'overcast', 'heavy rain', 'fog' etc were considered unclear.

2.3 Feature Selection

After data cleaning, the dataset still had 49 features. Many of these features were useless in helping solve the problem.

All the features that were extracted from the dataset were:

- **Severity:** This contained values (1, 2, 3, 4); 1 being mild traffic and 4 being heavy traffic, this was the target variable.
- **County:** It contained the county the accident occurred in.
- **State:** Contained the state initials, for example NY for New York
- **Zipcode:** Contained zipcode of the location of accident
- **Temperature:** Contained the temperature in Fahrenheit of the day during the accident
- **Humidity:** Contained the humidity during the accident
- **Visibility:** Contained the visibility in miles during the day of the accident
- **Weather condition:** Contained the weather condition (that we formatted to be clear/unclear)
- **Sunrise/sunset:** contained whether it was day or night during the time of the accident.
- **Time:** Contained the hour of the day the accident occurred.

3. Exploratory Data Analysis

3.1 Relationship between Temperature and Visibility

A scatter plot was made between the two to see the different visibilities for different temperatures in the dataset. The scatter plot gives an insight of how dense values are in some range and how some values are extreme and are clear outliers that would not help in making a good model.

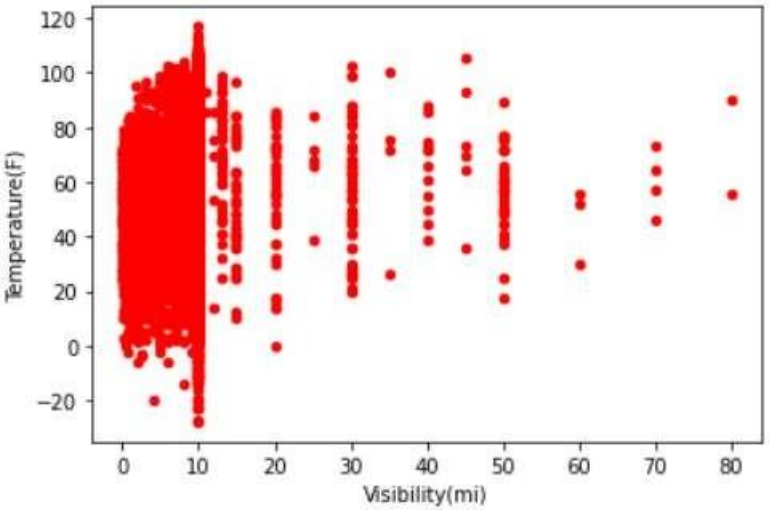

Figure 1: Relationship between Temperature(F) and Visibility(mi)

The scatterplot clearly indicates a lot of outliers present in the feature set Visibility. Most of the values are 10 or under, hence all the values above 10 were discarded, and the scatterplot was made again.

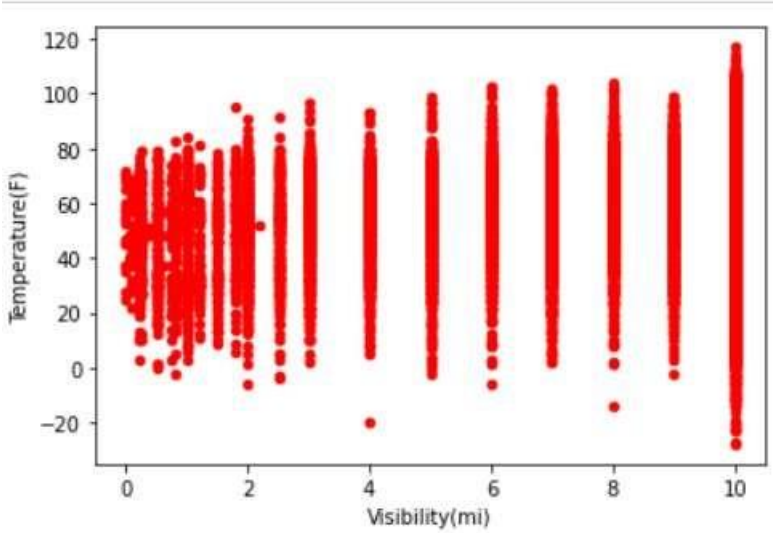

Figure 2: Relationship between Temperature(F) and Visibility(mi) after outlier removal

Hence the feature is homogenous now with no outliers.

3.2 Visualizing the Temperature feature

A boxplot was made for the temperature feature. A boxplot visualisation helps in extracting a lot of insights of a dataset. It helps in identifying the outliers, the median and the different quartiles.

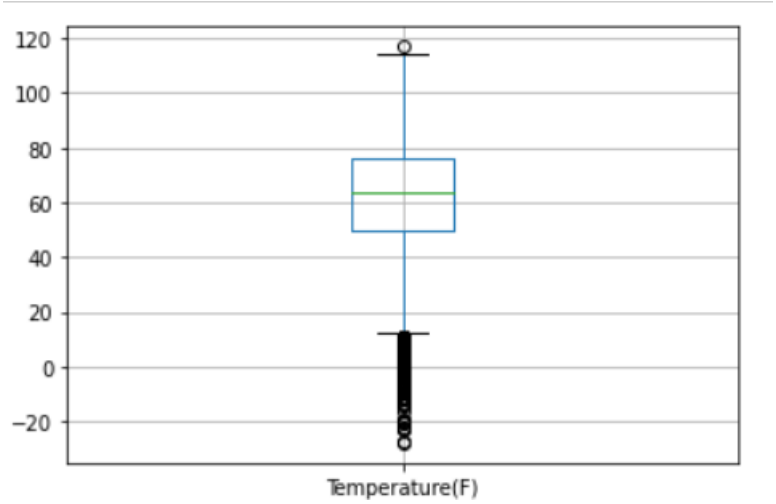

Figure 3: Boxplot of Temperature feature

It is clearly visible that it has a lot of outliers. All the values are mainly in the range from 40 to 90 Fahrenheit. So, all the other values that are out of this range are removed.

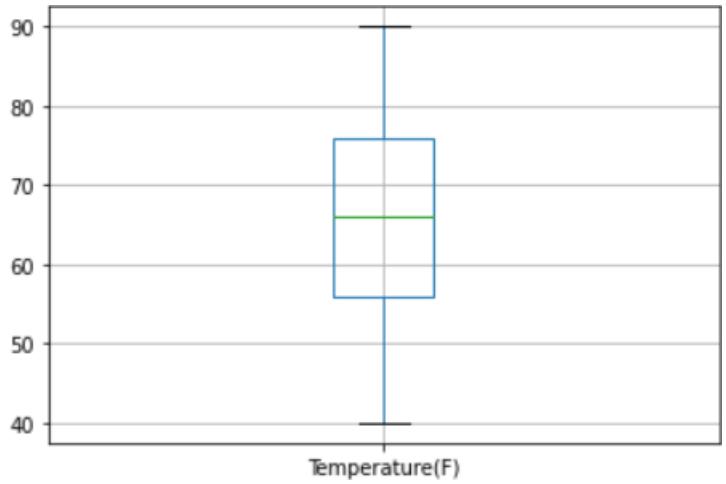

Figure 4: Boxplot of Temperature feature after removing outliers

The boxplot was plotted again after removing the outliers and it can be clearly seen from the figure that the data is homogenous now.

3.3 Visualizing the Humidity feature

Similarly, a boxplot was plotted for the humidity feature.

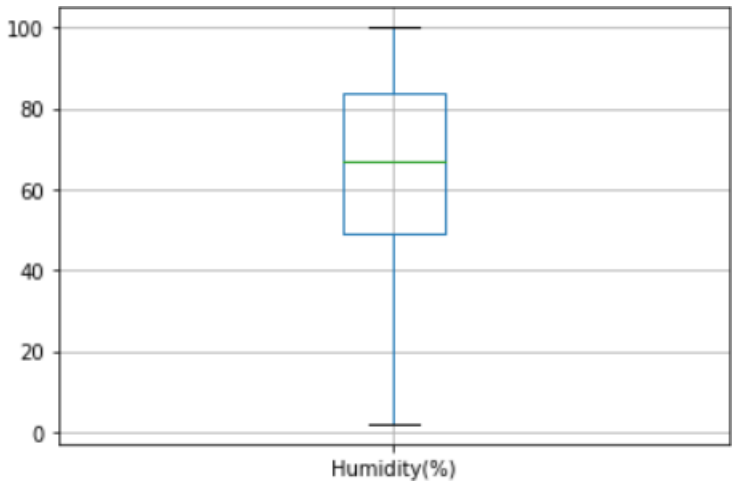

Figure 5: Boxplot for Humidity feature

It has no outliers since the values could only be between 0 and 100, but at the same time from the above boxplot we could roughly make the observation that most of the values are above 40. Hence all the rows having humidity below 40% were removed.

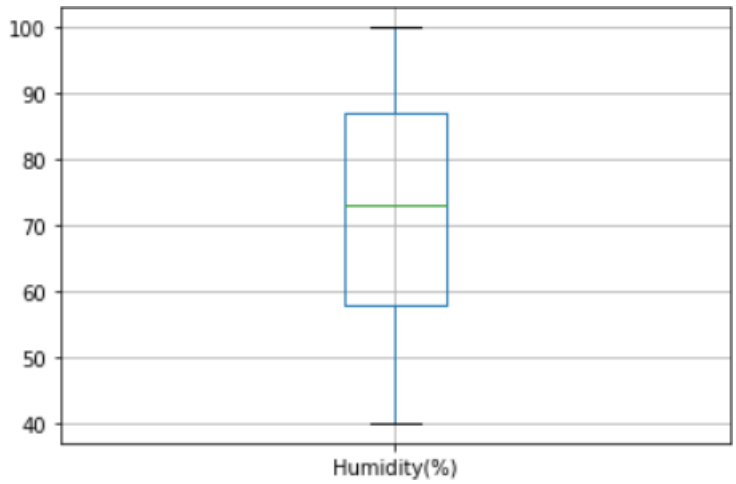

Figure 6: Boxplot for Humidity feature after formatting

After formatting the boxplot was plotted again and the feature looked homogenous and good enough to use.

4. Predictive Modelling

The problem is a classification problem. There are many ways to go about a classification problem, logistic regression, SVM, Decision trees, Nearest Neighbour (KNN). For this model, K Nearest Neighbour was used as it is a straightforward algorithm that gives good and efficient results.

4.1 K Nearest Neighbour Algorithm

4.1.1 Feature Selection for Nearest Neighbour Algorithm

The dataset was formatted in a way such that the nearest neighbour algorithm yielded the best results. To use the algorithm the scikit-learn library was imported, and to use this library the dataset was converted to a Numpy array.

The dataset was divided on the basis of features, to describe the target feature and the dependent features. Severity feature was the target feature, Temperature(F), Humidity (%), Visibility(mi) and Weather Condition were the features that were dependent.

4.1.2 Train/test Split for Nearest Neighbour Algorithm

The dataset was divided into 2 parts: 80% of the data was selected to train the model and 20% of the data was left out for testing the out of sample accuracy of the model.

4.1.3 Defining the 'k' parameter

'k' is the number of neighbours we consider around a given data point to compare and evaluate the prediction, and it changes based on different scenarios, too low value of 'k' ends up underfitting, too high ends up overfitting, hence an optimum value of 'k' was calculated by a function and the 'k' which gave the best results was chosen, **k=8.**

After getting this value, KNN algorithm was used to produce results.

5. Conclusions

In this study I analysed the relationship between the various aspects like weather conditions, time of the day, visibility etc when a road accident occurs and how severe the accident will be. I built a classification model that takes in all these factors affecting the severity of traffic due to an accident and predicts how severe the traffic will be.

This model and ideas of similar basis will be immensely useful for map navigation services, which will predict the severity of an accident on the user's route based on the features available and re-navigate the user as necessary.

6. Future Directions

The model had a ~70% in sample accuracy and a ~67% out of sample accuracy. I think the model can use some improvements like: having more information about the intensity of the effect of different features.

Other data like the type of vehicle, two-wheeler/four-wheeler, data from other countries if optimized, could bring significant improvements to the model.