Accident Severity Prediction

Fahad Thasim

August,2020

Introduction

Problem Statement

Reducing the number of collisions has become one of the important public safety challenges around the world. Millions of humans lives are lost each year due to car crashes. Prediction of the severity and understanding the key factor that contribute to these car accidents becomes vital for many governments across the world. Thus, it enables in creation of optimized traffic systems, safer routes and better transport infrastructure. Hence, the significance of developing a severity prediction model.

Data Acquisition and Cleaning

Data sources

The crash data used for the project is of the dataset provided in Kaggle here, which comprises of information regarding the collisions from the year 2016 and collected from multiple data sources. The dataset comprises of all types of collisions including motor vehicle, bicycle collisions etc. It also captures various information about the weather condition, location of the incident, etc. Details on the dataset is given here.

Using this information, we can see how attributes like weather conditions, location of the incident impact the severity of the accidents.

Data cleaning

Data was collected from multiple sources and combined into one. There were many missing data in the dataset. The dataset comprises of information on collisions from 2016 to 2020. Since, the data on 2020 is limited, these were excluded in certain analysis.

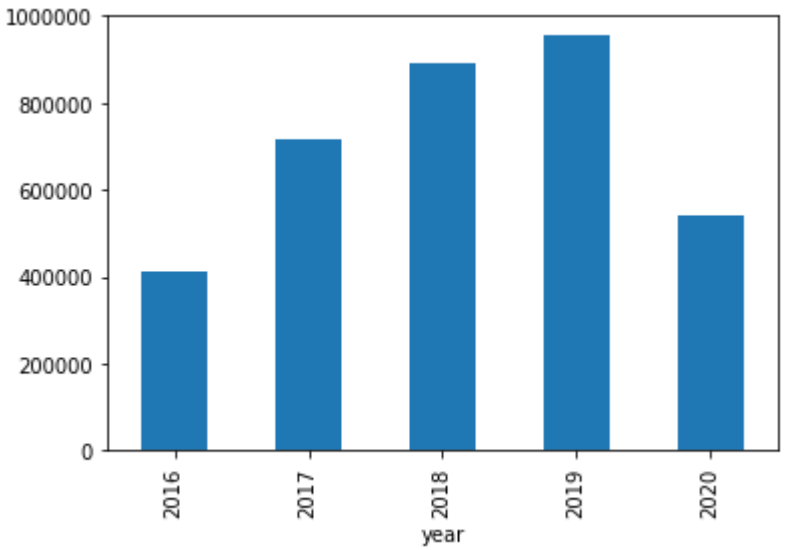


Figure 1: Year wise distribution

Columns on weather information also contained many missing values. Columns like temperature and pressure had a range of values, hence these were replaced with a mean. While other data like humidity, wind chill, visibility, wind speed,wind participation were replaced with zero to fill the null values.

Exploratory Data Analysis

Distribution of the data

We see that the majority of the data of the accident are of few major states in the country. This can be seen through the choropleth shown below.

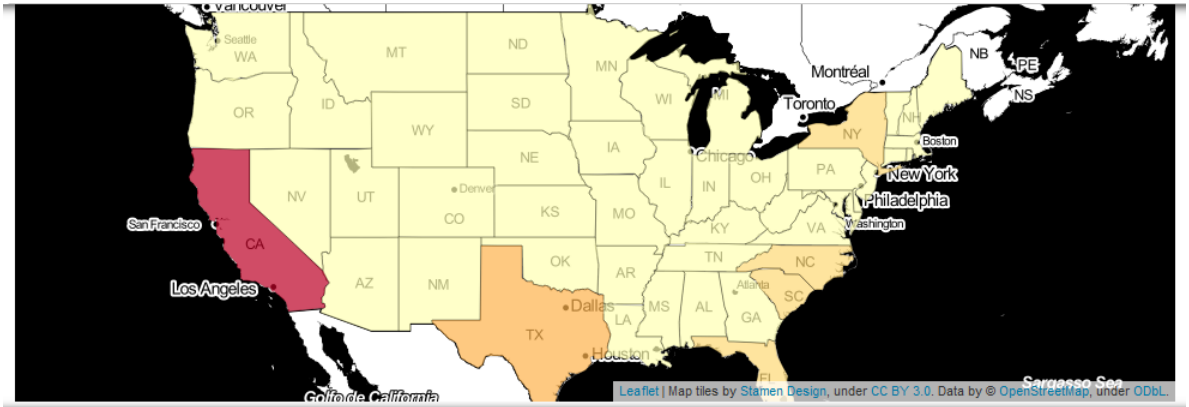


Figure 2: Choropleth of state wise accidents

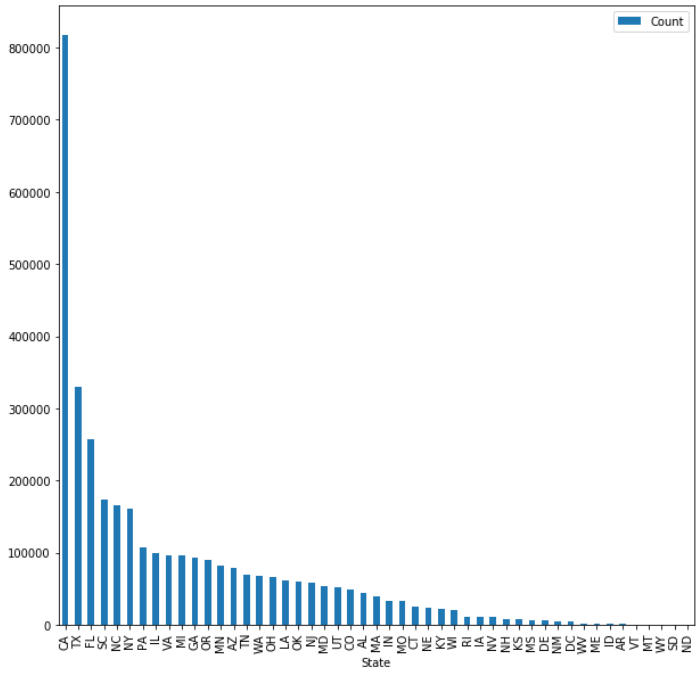
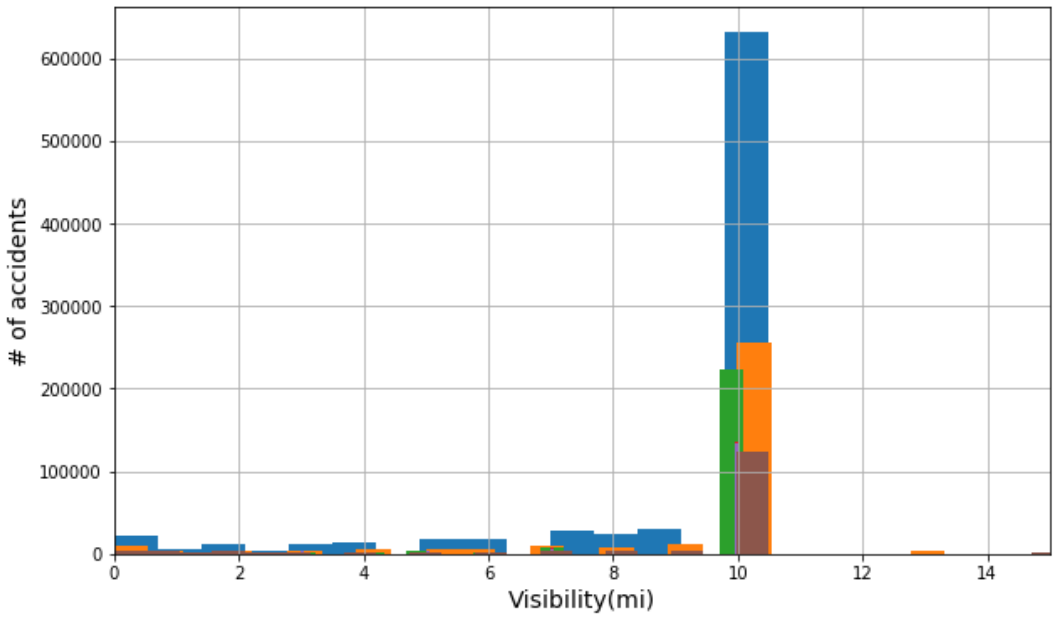


Figure 3: State-wise Accident Count

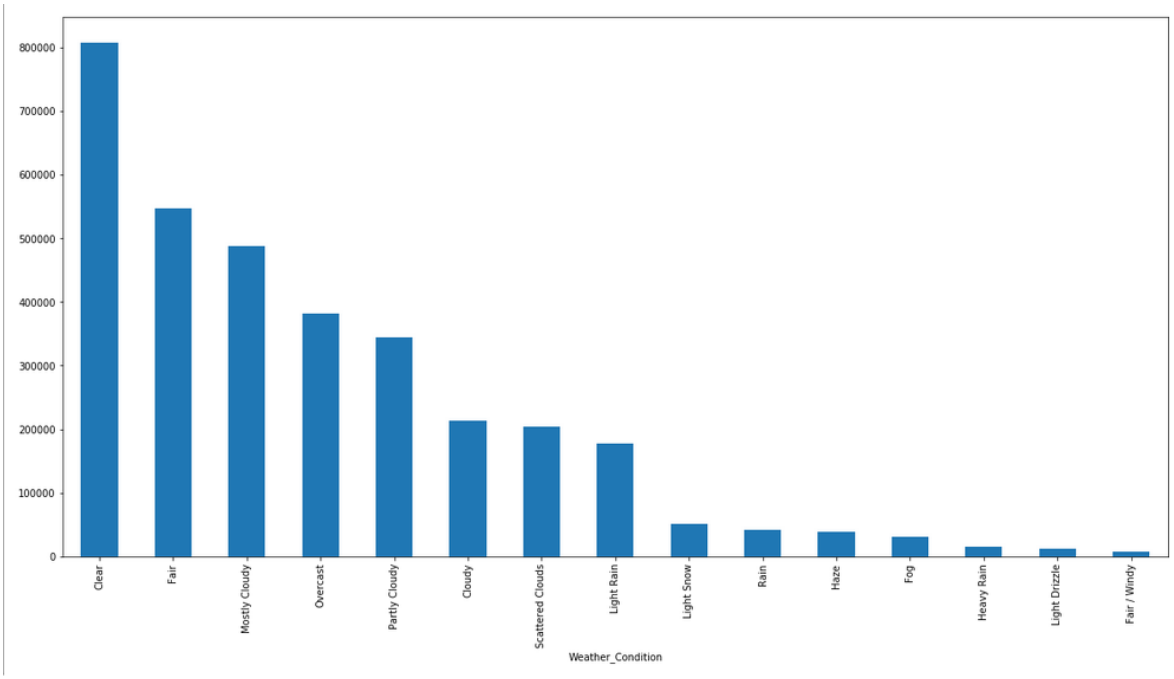
Relationship between accident and visibility

I hypothesized that the a majority of accident was affected by the visibility, but the distribution of data shows that the majority of the accident occurred when there was sufficient visibility. Hence, we can conclude that the visibility does not impact the accident



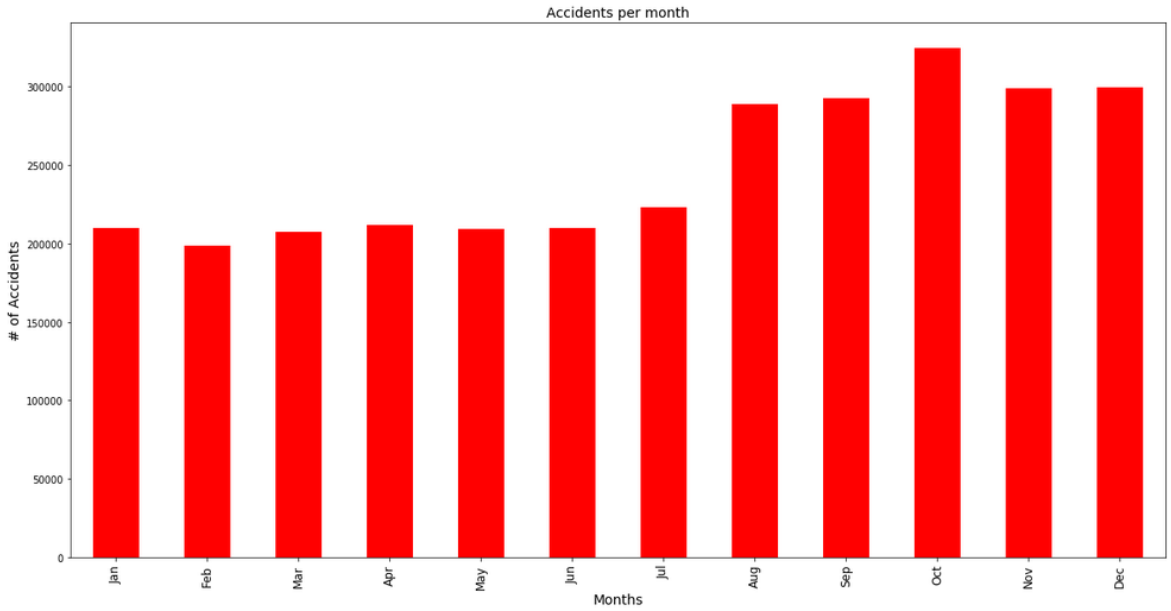
Relationship between accident and weather

Data seems to be equally distributed across all weather conditions and is balanced in the distribution.



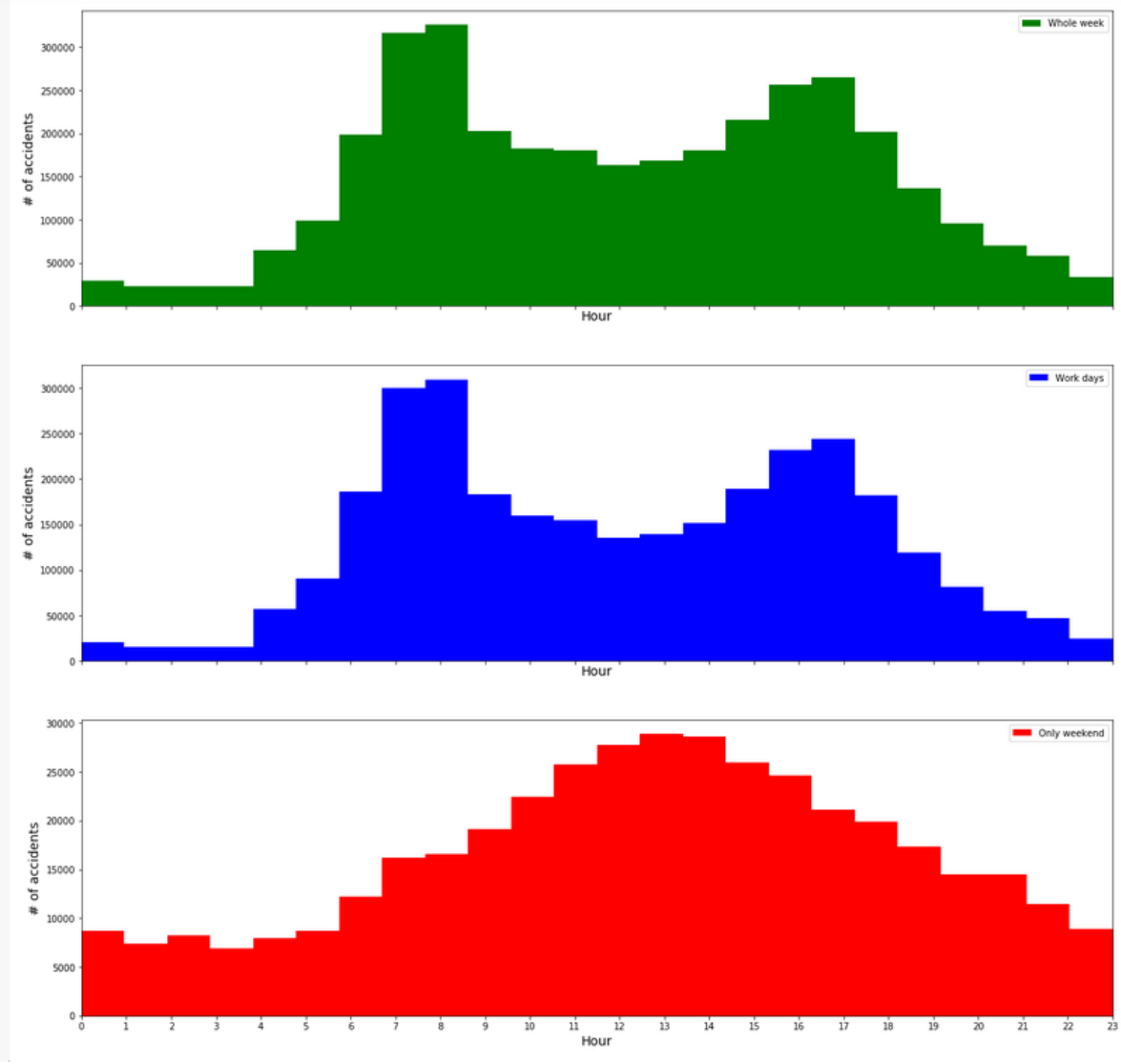
Month wise Distribution of the data

Data seems to be balanced and the accident data seems to increase in the second half of the year.



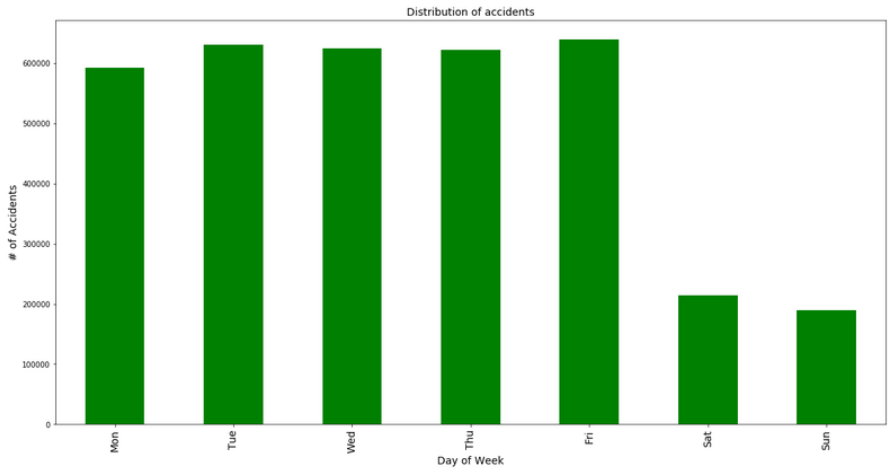
Accident Hour wise distribution

Majority of the accident distribution coincide with the office timings peaking around 9 and 5 during the weekdays and the around noon in case of weekends.



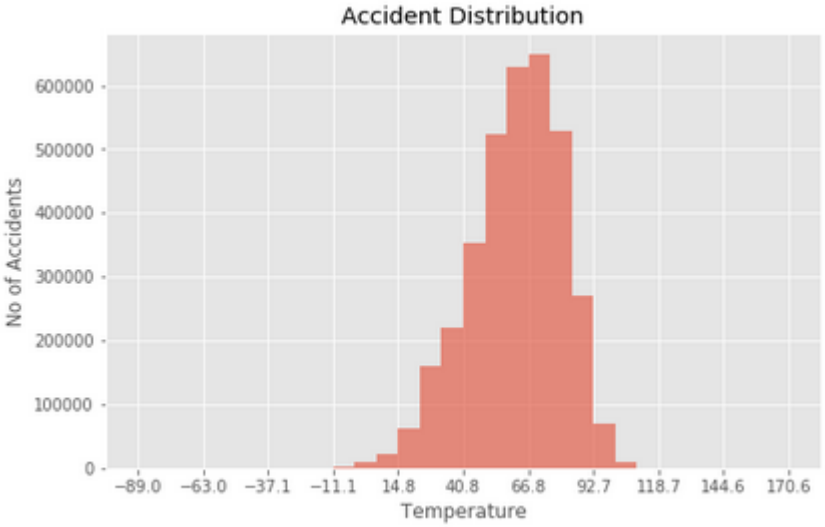
Weekday Distribution

Majority of the accident distribution occurs during the weekdays, compared to weekends.



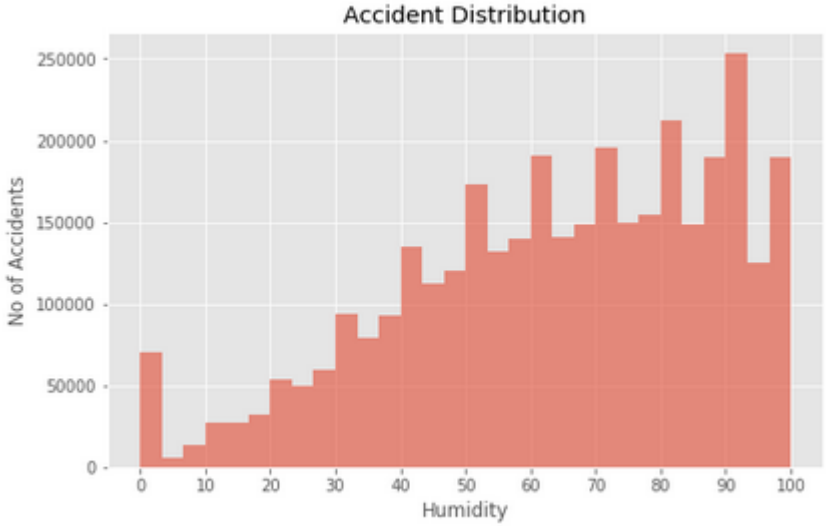
Relationship with temperature

Accident data appears to be normally distribution with mean around at 66 degree Farenheit.



Relationship with humidity

Accident count appears to increase with the increase in humidity. Showing a relationship between the two.



Predictive Modelling

We are using the classification model under supervised machine learning algorithms to predict the severity for the accident data. We have used couple of models for classifying the accidents.

Models

We have used logistic regression and decision tree classifier to classify the accident severity. Other models, like SVM and KNN algorithms showed slow performance due to large dataset. Hence, analysis was not possible with the two and compare the performance of all the 4 algorithms.

Performance of Classification Models

	Logistic Regression	Decision Tree
Log Loss	0.7603603252166787	
Accuracy	0.6755742940752238	0.6757598577876138

Conclusion

In this study, I have analysed the distribution of the accident data and various factors that influence the accidents. These were then used to predict the severity of the accident. I have used the classification models using Logistic Regression and Decision Tree Classifier for the classification. The models was obtaining ~68% accuracy and there is still scope for improvement. With more data on these accidents, the model could be trained to improve its accuracy and help organization in creating better transport infrastructure.