Car Accident Severity Analysis

US Countrywide Traffic Accident Dataset (2016 - 2020)

The Project aim to understand the factors which play a role in the severity of accidents using Machine Learning Models.

Submitted by: Abhinand G K

September 2020

Car Accident Severity Analysis

1. Introduction

1.1. *Background*

Traffic accidents are a significant source of deaths, injuries, property damage, and a major concern for public health and traffic safety. Accidents are also a major cause of traffic congestion and delay. Effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency. The effective treatment of road accidents and thus the enhancement of road safety is a major concern to societies due to the losses in human lives and the economic and social costs. Tremendous efforts have been dedicated by transportation researchers and practitioners to improve road safety.

1.2 *Problem*

The world as a whole suffer due to car accidents. Accurate predictions of severity can provide crucial information for emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures. Previous years countrywide car accident dataset can be used to determining severity prediction.

1.3 *Interest*

Obviously, Government would be very interested in accurate prediction of accident Severity, because effective management of accident is crucial to mitigating accident impacts and improving traffic safety and transportation system efficiency.

2. Data acquisition and cleaning

2.1 *Data sources*

U.S countrywide car accident dataset data can be found in Kaggle datasets. Here countrywide car accident dataset, which covers 49 states of the USA is used for the analysis. 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.

2.2 *Data cleaning*

Data downloaded from kaggle were combined into one table. There were 3513637 accident records in this dataset. There were some missing data, because of lack of record keeping. The severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic. The columns End_Lat, End_Lng contains 74% null values so decided to drop the columns. Description include many distinct values decided to drop that column also. And there is many non-relevant columns in the dataset and they are analysed and removed. After removing the unwanted columns the rows contains nan value also removed. The time format is changed from string to time stamp and duration of the incident also calculated.

2.3 *Feature Selection*

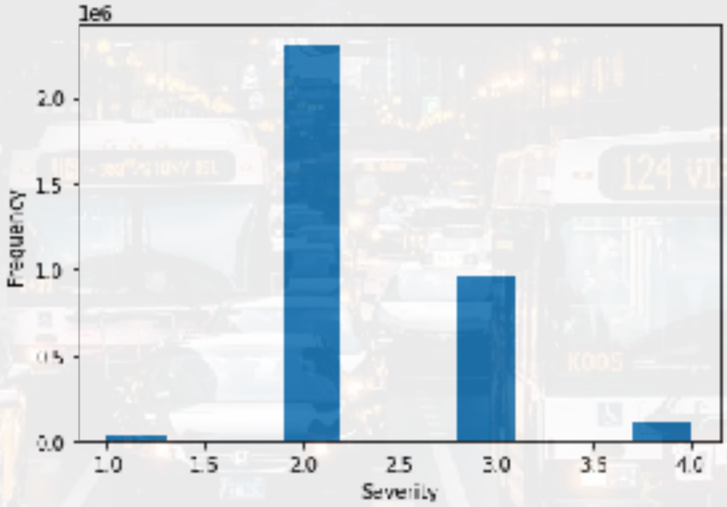
After data Cleaning there were 3414252 samples and 38 features in the data. Upon examining the meaning of each feature, it was clear that there was some redundancy in the features. From these highly correlated features, only one was kept, others were dropped from the dataset. After all, 23 features were selected.

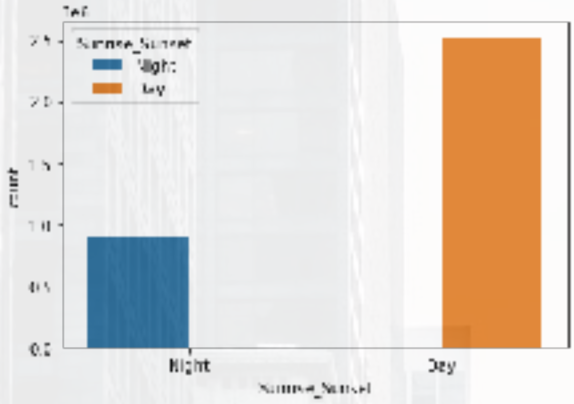
3. Exploratory Data Analysis

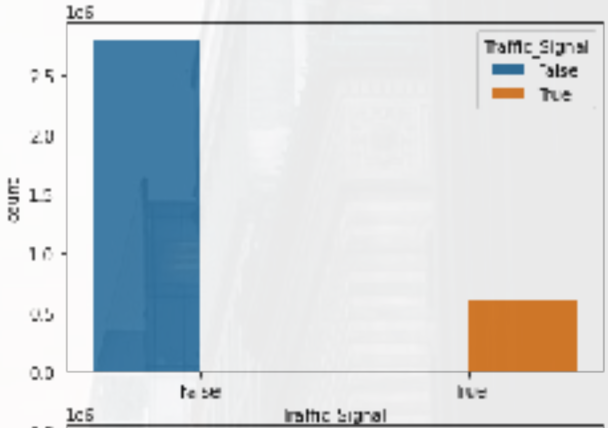
3.1 *Calculation of target variable*

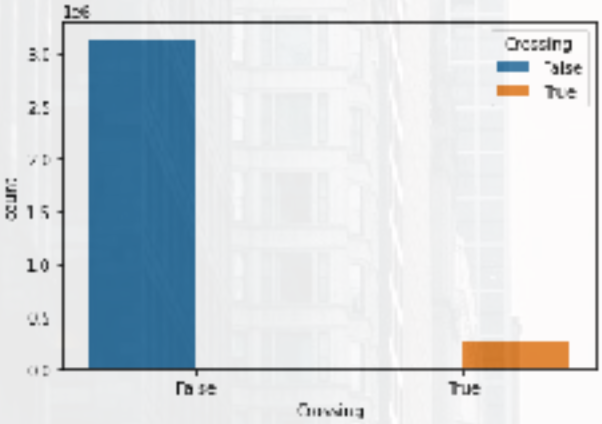
The characteristics data set contains information on the time, place, and weather and lighting conditions and type of intersection where it occurred. An initial analysis of the data was performed for the selection of the most relevant features for this problem, reducing the size of the dataset and avoiding redundancy. With this process the number of features was reduced from 38 to 20. Incident time duration was not a feature in the dataset, and had to be calculated. I chose to calculate the difference of start time and end time in minutes as the target variable. This is done by the time format is changed from string to time stamp and difference calculated.

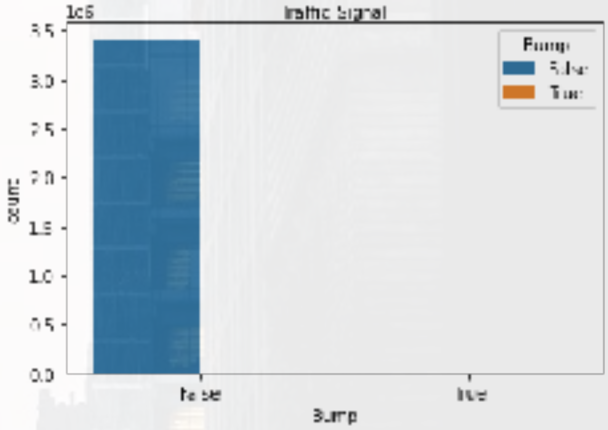
Graphs



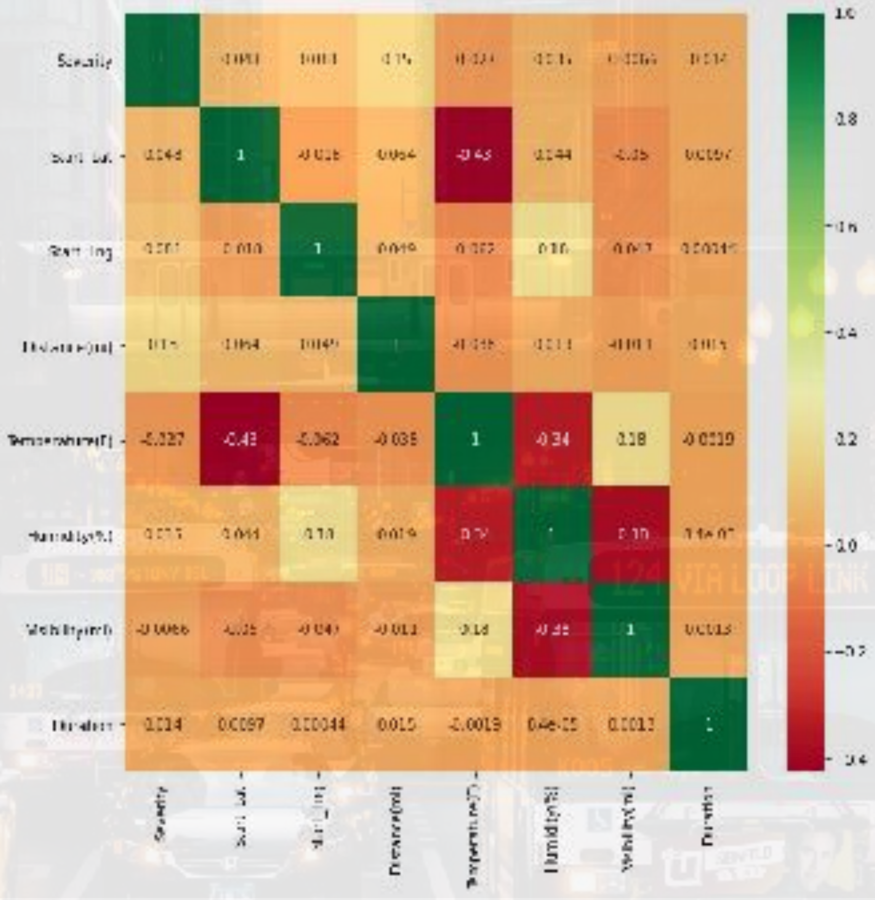








Correlation Heat map



3.1 *Description*

From the characteristics dataset the following features are selected

Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
Distance(mi)	The length of the road extent affected by the accident.
State	Shows the state in address field.
Temperature(F)	Shows the temperature (in Fahrenheit).
Visibility(mi)	Shows visibility (in miles).
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
Amenity	A POI annotation which indicates presence of amenity in a nearby location
Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
Crossing	A POI annotation which indicates presence of crossing in a nearby location.
Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
Junction	A POI annotation which indicates presence of junction in a nearby location.
No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
Railway	A POI annotation which indicates presence of railway in a nearby location.
Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.

Station	A POI annotation which indicates presence of station in a nearby location.
Stop	A POI annotation which indicates presence of stop in a nearby location.
Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location
Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.

In addition, one new features were crafted, to perform a duration analysis of the accident severity.

For categorical features the features are encoded using a one-hot encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array.

4. Predictive Modelling

Different classification algorithms have been tuned and built for the prediction of the level of accident severity. These algorithms provided a supervised learning approach predicting with certain accuracy and f1 score. These two properties have been compared in order to determine the best suited algorithm for his problem. Here preprocessing is done by standard scaler tool from scikitlearn library. After that the sample data is split into train and test data set in the ratio 80:20.

4.1 *K-Nearest Neighbour*

The K-nearest neighbours (KNN) algorithm is a type of supervised machine learning algorithms. KNN is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. Here KNN is implemented with Python's Scikit-Learn library.

```
f1 score: 0.6720916706047082
Accuracy score: 0.67088
```

4.2 *Decision Trees*

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

f1 score: 0.67519080668298
Accuracy score: 0.6754

4.3 *Logistic Regression*

Logistic regression is a fundamental classification technique. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, and it's convenient for you to interpret the results. Although it's essentially a method for binary classification, here it is applied to multi-class problems as one verses other method.

f1 score: 0.6720916706047082
Accuracy score: 0.6774341492949696

4.4 *Support Vector Machine (SVN)*

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. The advantages of support vector machines are Effective in high dimensional spaces, Still effective in cases where number of dimensions is greater than the number of samples, Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient, Versatile: different Kernel functions can be specified for the decision function.

f1 score: 0.6774341492949696
Accuracy_score: 0.67319

5. Result

The metrics used to compare the accuracy of the models are the f1-score and accuracy score. This table reports the results of the evaluation of each mode

Algorithm	f1-score	Accuracy score
K-Nearest Neighbour(K=7)	0.6720916706047082	0.67088
Decision Trees	0.67519080668298	0.6754
Logistic Regression	0.6720916706047082	0.67088
Support Vector Machine (SVN)	0.6774341492949696	0.67207

6. Conclusions

In this study, I analyzed the relationship between various features and Traffic Accident Severity. Here classification models is used to predict Traffic Accident Severity. These models can be very useful effective in management of accident, mitigating accident impacts and improving traffic safety accidents and thus the enhancement of road safety. It is a major concern to societies due to the losses in human lives and the economic and social costs. It could help identify the severity and it help to improve road transportation system efficiency.

7. Future directions

I was able to achieve 67% of accuracy in the classification problem. However, there was still significant variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing date of incident, month, weekend or not, place, street, etc. And here only 15% of total dataset is used due to computational problems. More data, especially data of different types, would help improve model performances significantly.

Reference

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019).
- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.