Predicting Severity of an accident

Devmani Dhiman

October 12, 2020

https://www.linkedin.com/post/edit/6722449218474192896/

1. Introduction

1.1 Background

According to Wikipedia, there were more than 273 million road motor vehicles in the USA (till 2018). Over 76% of the American drives to work alone, which makes USA one of the most densely populated with vehicles. The Increasing number of vehicles increase the chances of the Road fatalities, on an average nearly every year there are 6 million car accidents and more than 90 people die in car accidents every day. Around 3 million people in US are injured every year in car accidents. These crashes result in nearly 6% fatality. Main causes of these accidents are:

a. Driver under alcohol influence

b. Distracted Driving

c. Speeding

d. Reckless Driving etc.,

Apart from these sometimes-environmental conditions also makes it’s difficult to drive and increases the chances of an accident. What if there is a way to predict the severity of an accident. This information would be crucial if we predict the severity of an accident, we can possibly mitigate the actual losses.

1.2 Problem

Data that might help in predicting the accident severity might include road condition, street lights, Light condition, weather condition, time of the day, month or year. This project aims to predict the severity of an accident based on these data.

1.3 Interest

Everyone who drives (nearly everyone drives or use some sort of transportation) would be interested in knowing how likely is the occurrence of an accident on a particular day.

2. Data Acquisition and cleaning

The dataset is available as comma-separated values (CSV) files, KML files, and ESRI shapefiles that can be downloaded from the Seattle Open GeoData Portal. The data is also available from RESTful API services in formats such as GeoJSON.

2.1 Loading the dataset

We download the dataset to our project directory and take a look at the data types and the dimensionality of the data. We can see that the dataset contains 194,673 records and 40 fields.

The metadata of the dataset can be found from the website of the Seattle Department of Transportation. On reading the dataset summary, we can determine the description of each of the fields and their possible values.

The data contains several categorical fields and corresponding descriptions that could help us in further analysis. We make an attempt at understanding the data in terms of the fields that we shall take into account for later stages of model building.

The X and Y fields denote the longitude and latitude of the collisions. We can visualize the first few non-null collisions on a map.

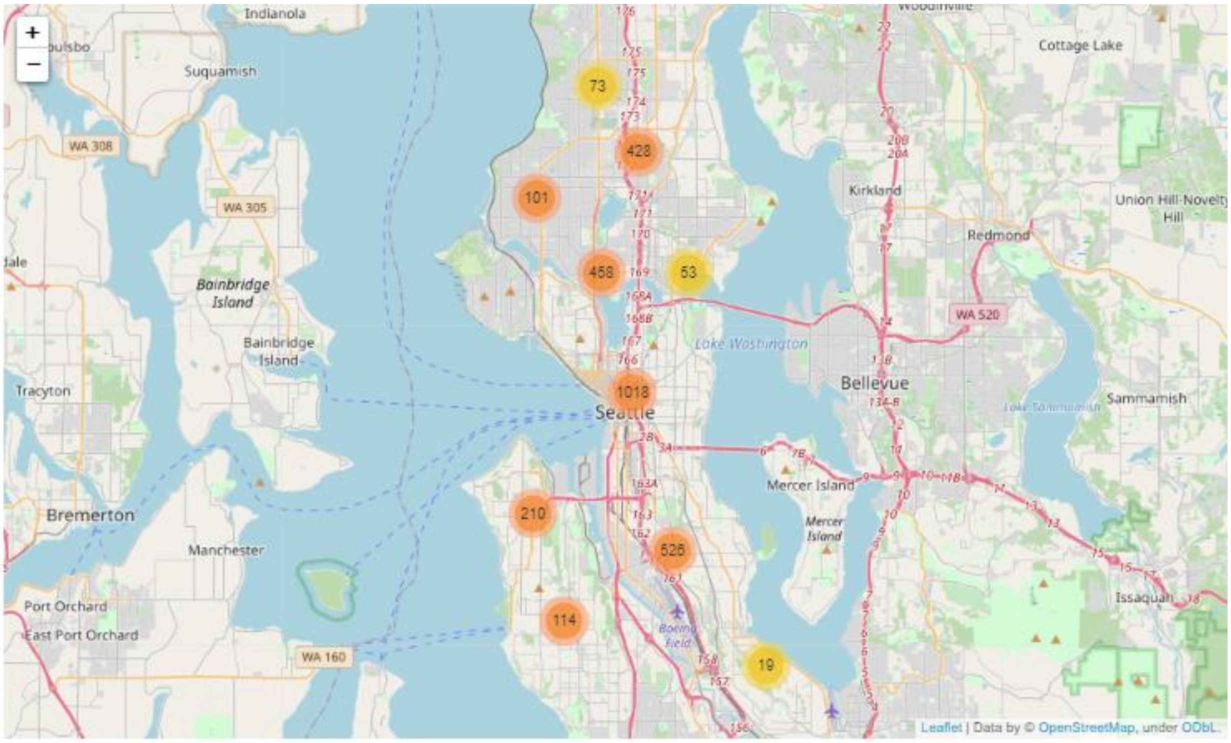


Fig 1. Map of Seattle city with accident data across the city

2.2 Data Cleaning

The dataset used for this project is based on car accidents which have taken place within the city of *Seattle, Washington* from the year *2004* to *2020*. This data is regarding the severity of each car accidents along with the time and conditions under which each accident occurred. The data set used for this project can be found on Seattle website. The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury). The dataset has 194673 rows and 38 features.

In order to deal with the issue of columns having a variation in frequency, arrays were made for each column which were encoded according to the original column and had equal proportion of elements as the original column. Then the arrays were imposed on the original columns in the positions which had *Other* and *Unknown* in them. This entire process of cleaning data led to a loss of almost 5000 rows which had redundant data, whereas other rows with unknown values were filled earlier.

2.3 Feature Selection

After data cleaning, there were 169957 rows and 38 features (columns) in the dataset. Upon examining it was clear that there are some redundancies in the features. For e.g. 'INCDTTM' and 'INCDATE' both features contain the data about the data and time of the accident. The difference between the two is that 'INCDATE' feature only contains the date of the accident and 'INCDTTM' contains the date as well as time data of the accident.

There were other features that had no effect on the severity of the accident. Those features are

- 'OBJECTID'
- 'INCKEY'
- 'COLDETKEY'
- 'REPORTNO'
- 'STATUS'
- 'ADDRTYPE'
- 'INTKEY'
- 'LOCATION'
- 'EXCEPTRSNCODE'
- 'EXCEPTRSNDESC'
- 'SEVERITYDESC'
- 'COLLISIONTYPE'
- 'SDOT_COLCODE'
- 'SDOT_COLDESC'

- 'SDOTCOLNUM'
- 'ST_COLCODE'
- 'ST_COLDESC'
- 'SEGLANEKEY'
- 'CROSSWALKKEY'
- 'HITPARKEDCAR'

These features are either unique identifier for the dataset set such as INTKEY, OBJECTID, etc. or are the description of the codes used for severity of an accident example SDOT_COLCODE, SDOT_COLDESC, etc.

The INCDTTM feature was used to analyze the number of accidents over the years but later dropped. On analyzing the data using this feature we found out that the number of accidents decreased over the years. Correlation also showed that there is some relation between the features (Pearson correlation coefficient > 0.3). For example, PERSONCOUNT was related to VEHCOUNT as this seems logical too.

The features that were selected for the prediction are:

- X
- Y
- PERSONCOUNT
- PEDCOUNT
- PEDCYLCOUNT
- VEHCOUNT
- UNDERINFL
- ROADCOND
- LIGHTCOND
- WEATHER
- SPEEDING

3. Exploratory Data Analysis

The first step I took in EDA was to plot the map of the Seattle city using Folium library. The plotted map shows that most of the accidents occurred in the center of the city, more than 50% of the accidents happened in the center of the city. This seems logical as center of the city attracts large population, which increases the chances of having an accident.

3.1 Number of accidents over the year

The dataset we have had data from year 2004 to 2020. On plotting the graph, we see that the total number of accidents have gone down over the years. To quantify this data, we can approximate the % decrease in the accidents over the year. The data shows that there is approximately 30% decrease in the number of accidents over the years.

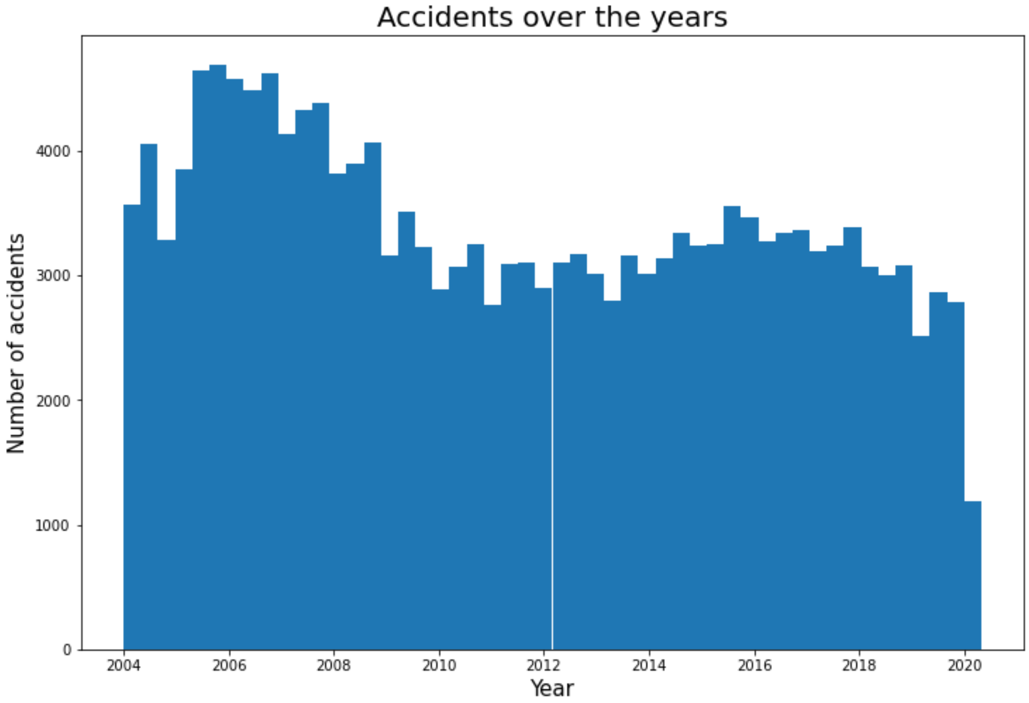


Fig: Number of accidents from year 2004-2020

3.2 Effect of weather conditions

There are total 9 weather condition that are available in the dataset. I used Seaborn library for plotting as it is simple and easy to use. The bar graph (using count plot) shows us that more than 90% of the accident happened in either “Clear”, “Raining” or “Overcast” weather conditions, and approximately 70% of those accidents occurred in “Clear” weather condition. Clear weather condition is generally considered good for travelling but our analysis points towards the other direction. This means that there are other contributing factors which governs the severity of an accident. This relationship is depicted below.

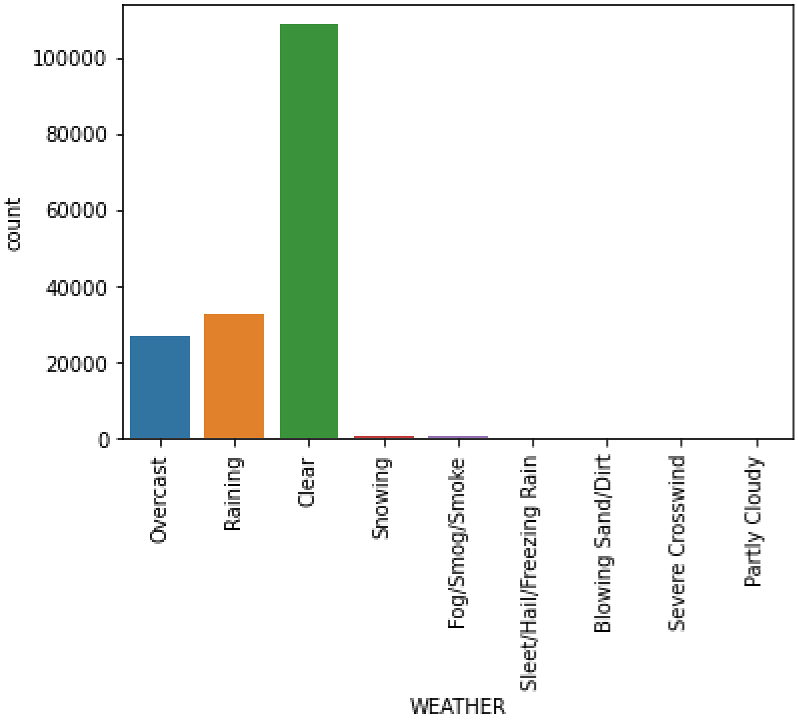


Fig: Accident count vs weather conditions

3.3 Conditions of Roads

I hypothesized that road conditions can play a vital role in predicting the severity of an accident. If the road is wet due to rain then there are high chances of an accident as the friction between the tires and the road decreases which make your vehicle slip/skid on the road causing an accident to happen. On Examining the graph, we can infer that “Dry” and “Wet” Road conditions are major reason for the accident. But we should not jump to conclusion before analyzing the other features of the dataset. The bar graph is shown below.

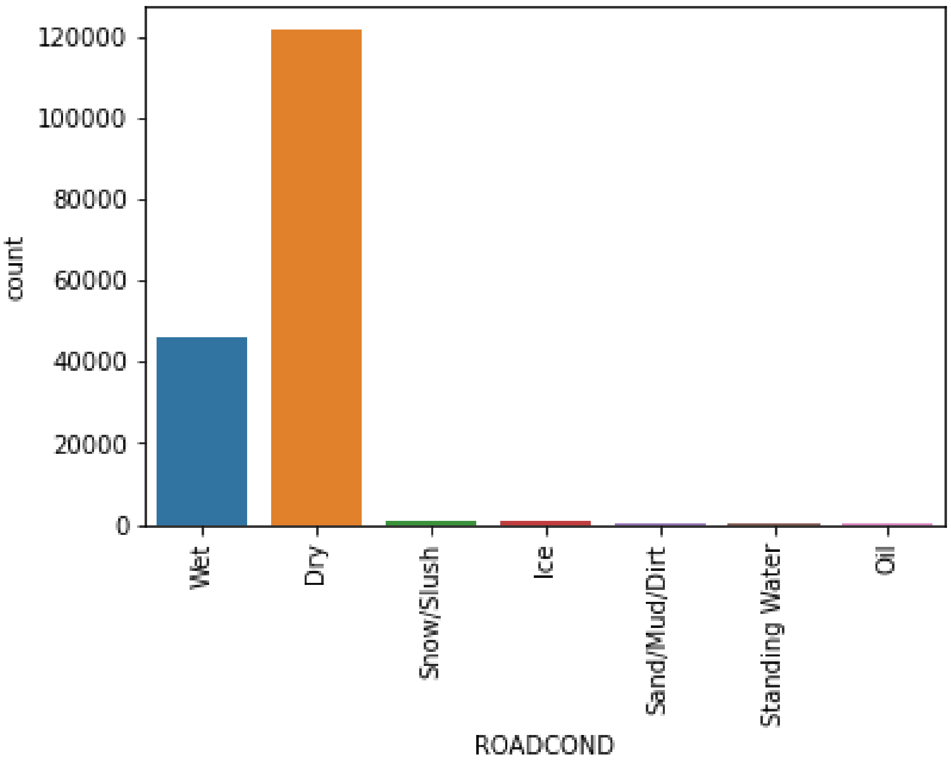


Fig: Number of accidents vs Road Condition

3.4 Effect of Light Condition

Similar to Road condition and weather, Light condition also play an important role in determination of accident severity. Light condition is one of the factors that influence the chances of an accident. The graph shows us that most of the accidents happen during either "Daylight" or when it's" Dark". The graph below shows us the count of accident in different light condition.

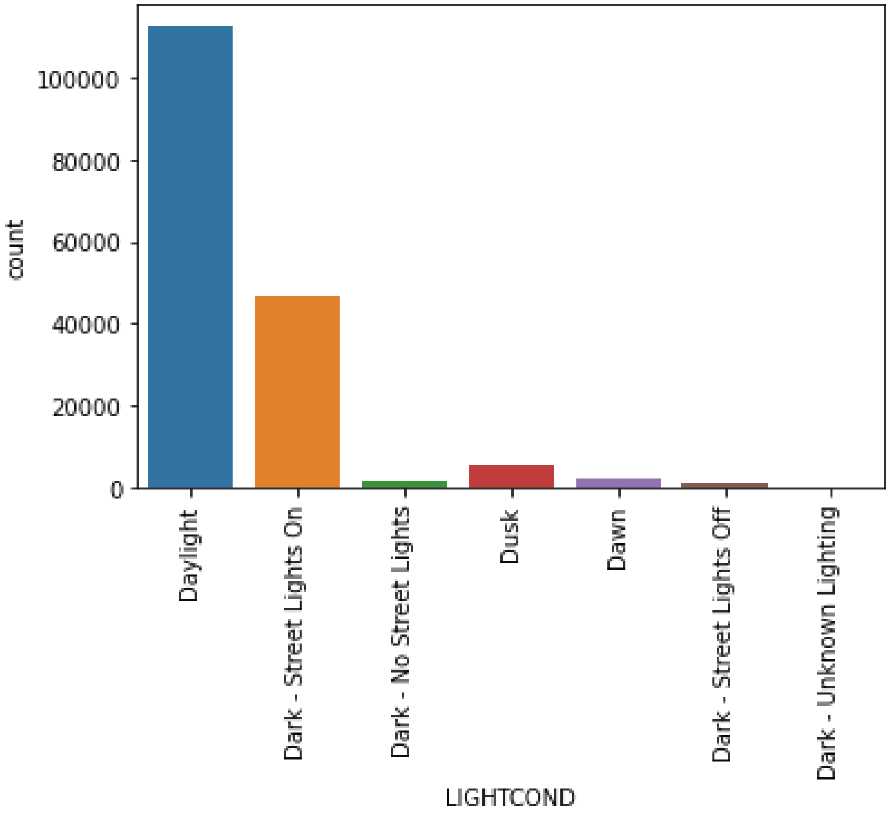


Fig: Light Condition vs number of accidents

We can infer that after Daylight, most of the accident occur in Dark light condition.

3.5 Some other graphs used for analyzing the data

1. Under Influence of any drugs/alcohol

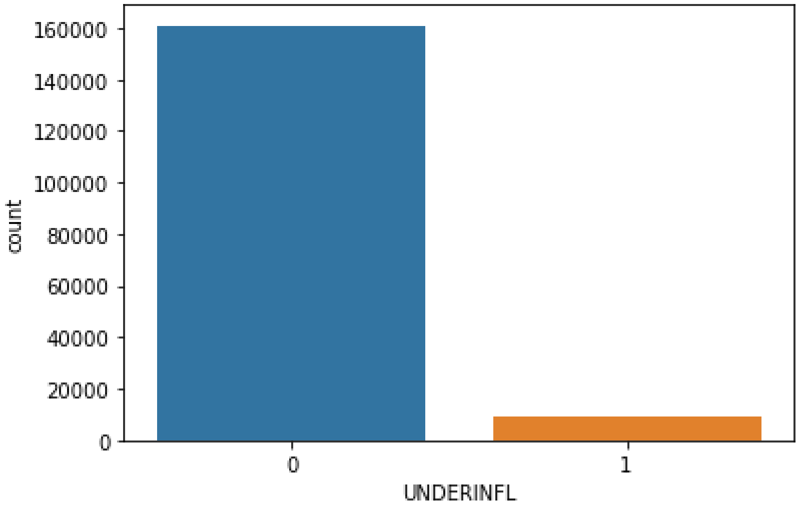


Fig: Driver under influence vs No. Of accidents

2. Vehicle Count vs Person count

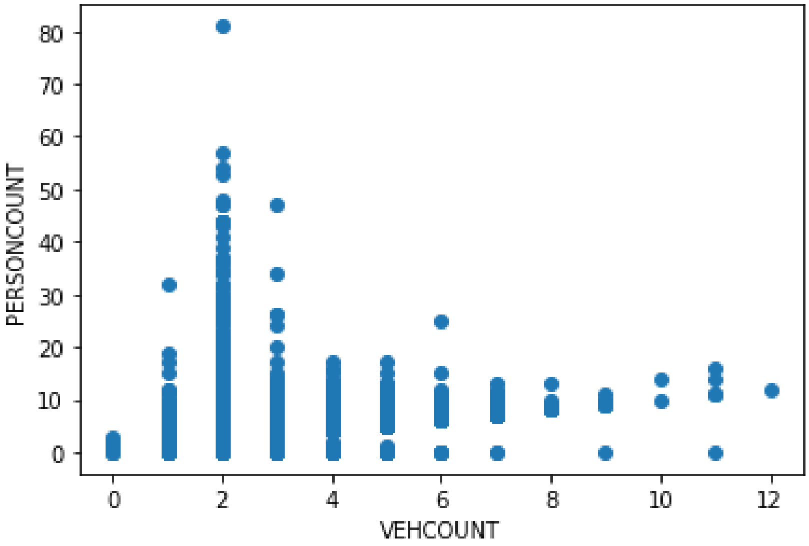
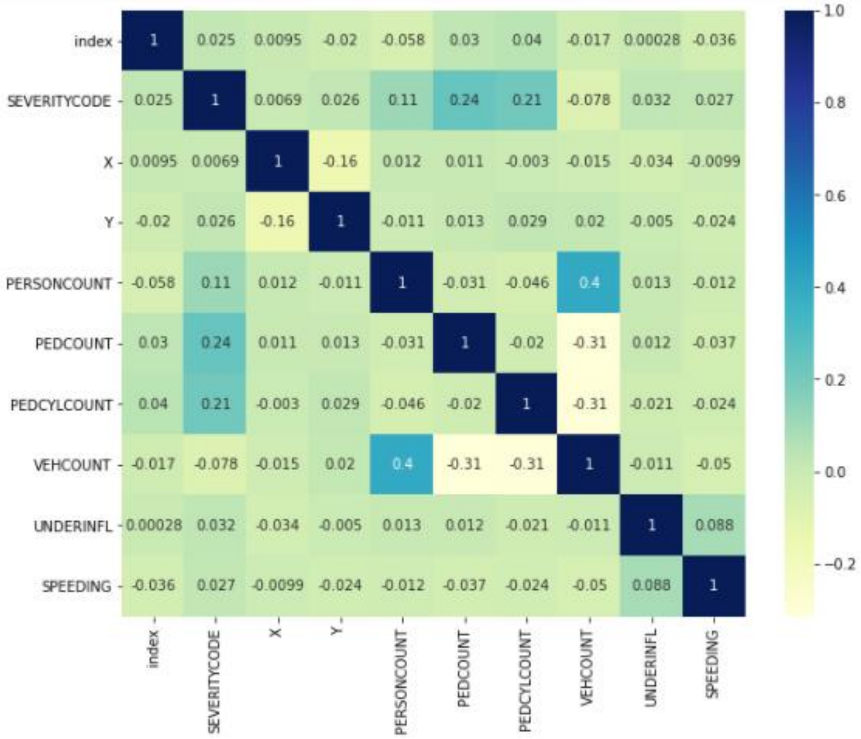


Fig: Scatter plot between Vehicle count and person count

3. Heatmap of Correlation between variables



	index	SEVERITYCODE	X	Y	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	UNDERINFL	SPEEDING
index	1	0.025	0.0095	-0.02	-0.058	0.03	0.04	-0.017	0.00028	-0.036
SEVERITYCODE	0.025	1	0.0069	0.026	0.11	0.24	0.21	-0.078	0.032	0.027
X	0.0095	0.0069	1	-0.16	0.012	0.011	-0.003	-0.015	-0.034	-0.0099
Y	-0.02	0.026	-0.16	1	-0.011	0.013	0.029	0.02	-0.005	-0.024
PERSONCOUNT	-0.058	0.11	0.012	-0.011	1	-0.031	-0.046	0.4	0.013	-0.012
PEDCOUNT	0.03	0.24	0.011	0.013	-0.031	1	-0.02	-0.31	0.012	-0.037
PEDCYLCOUNT	0.04	0.21	-0.003	0.029	-0.046	-0.02	1	-0.31	-0.021	-0.024
VEHCOUNT	-0.017	-0.078	-0.015	0.02	0.4	-0.31	-0.31	1	-0.011	-0.05
UNDERINFL	0.00028	0.032	-0.034	-0.005	0.013	0.012	-0.021	-0.011	1	0.088
SPEEDING	-0.036	0.027	-0.0099	-0.024	-0.012	-0.037	-0.024	-0.05	0.088	1

Fig: Heatmap of correlation data

The correlation graph shows us that features such as "PEDCOUNT" (Number of Pedestrian involved), PERSONCOUNT (Number of people involved in accident) and PEDCYCLCOUNT (Number of bicycles involved in collision) have a moderate correlation with SEVERITYCODE. So, these variables can help us in predicting the severity of an accident more accurately.

Another Correlation matrix was used to find out more insights between different variables that are used in the prediction. With help of this matrix we could figure out that RoadCond_Dry and Weather_Clear have high correlation. So we can use one variable to guess the value (if missing)

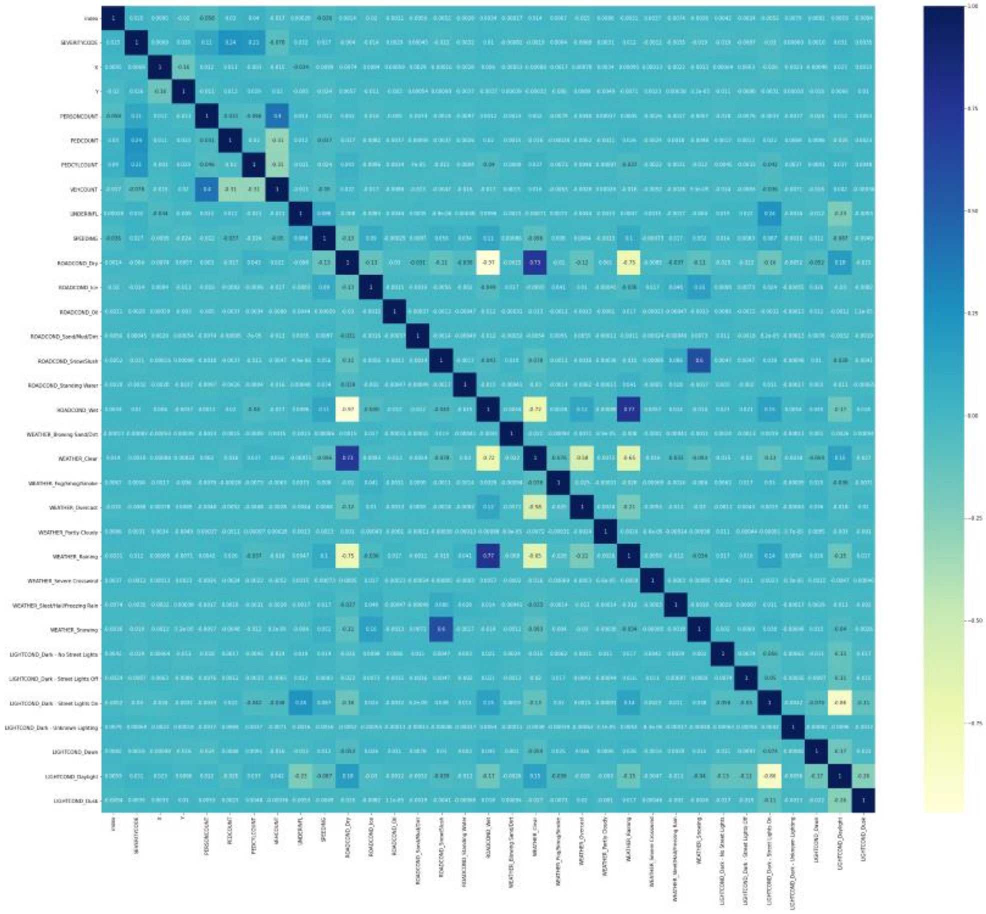


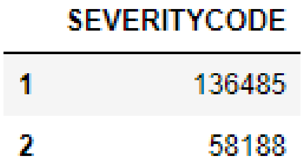
Fig: Correlation matrix after One Hot Encoding

4. Predictive Modeling

There are 2 types of models, Classification and regression. The Regression task is used to find out a continuous value of the target variable. In case of Classification, it is used when we need to classify the target feature into different categories. Our job was to classify the severity of an accident based on different features available in the dataset, so I chose Classification model to classify the severity of an accident.

4.1 Classification models

The application of classification model is straightforward. Our target is divided into 2 categories namely “1” and “2”. “1” representing “property damage” and “2” representing “injuries”. The distribution of the 2 severities is shown below:



	SEVERITYCODE
1	136485
2	58188

I used to methods for predicting the target variable namely, Logistic Regression and Decision Tree. Classification matrix was used to find the models ability to correctly predict the target variable.

Both the method achieved a decent accuracy and had similar classification matrix. Logistic Regression achieved an accuracy of 73% where the Random Forest Classifier achieved an accuracy of 70%. The classification score of both the algorithms is listed below.

Algorithm	Target Variable	Precision	Recall	F1-score
Logistic Regression	1	0.72	0.98	0.83
	2	0.82	0.21	0.34
Random Forest	1	0.73	0.87	0.80
	2	0.57	0.35	0.43

These results are due to the fact that the dataset is highly imbalanced. Due to the imbalance in the input dataset our model is doing well in predicting the Severity code = 1 as there are more data points with respect to this code.

5. Conclusion

In this study, I analyzed the relationship between the severity of an accident and different environment condition that may have impact on the severity of that accident. I analyzed that the severity of an accident is not dependent on a single feature, rather is dependent on more than one feature. The feature that were given in the dataset were not able to accurately predict the severity of an accident due to the following reasons:

1. Imbalanced dataset

2. Insufficient data in the dataset

   There were only 2 types of severity code present, whereas in metadata file there are a total of 4 different severity code present. Given the fact that this data is collected from 2004 there are slight chances that accident of severity = 3, 2b never occurred. The accuracy of the model can be increased if we get the sufficient data for all the categories.

6. Future Scope

The accuracy from both the models is decent and there are chances of improvement. I think the models could use more improvements on capturing more features and traits of the vehicle. For example, the vehicle that has suffered an accident may have some fault in its mechanical system that lead to an accident of high severity. The weight the vehicle is carrying, overweight vehicles tends to have more chances of accidents rather than normal weight vehicles. More data, especially data of different types, would help improve model performances significantly. Models in this study mainly focused on environmental features. However, car conditions, and other factor might also contribute to a severity of an accident. These interactions data are obviously more difficult to extract and quantify, but if optimized, could bring significant improvements to the models.