IBM APPLIED DATA SCIENCE CAPSTONE PROJECT

Car Accident Severity

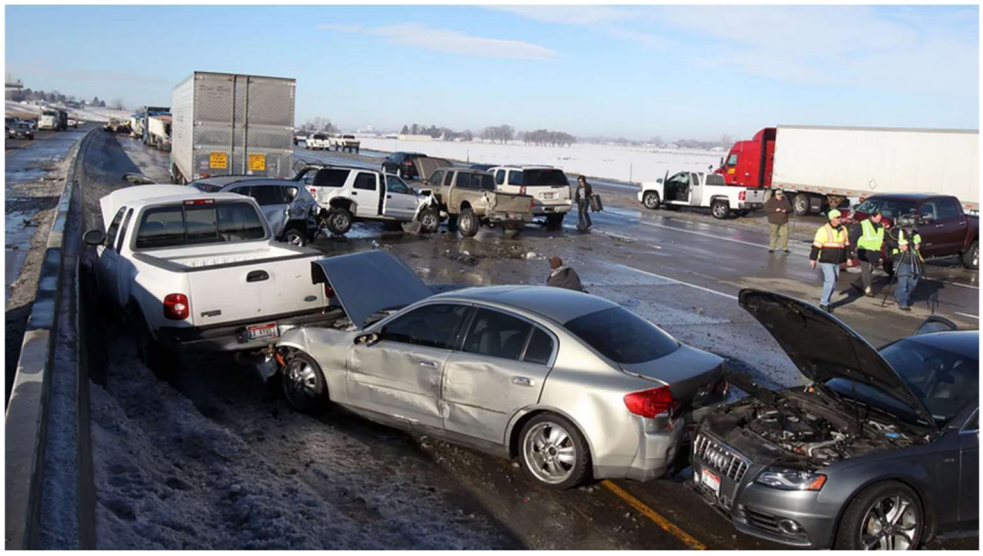
Help the Indian Government to prevent accidents...

1. Introduction/Business Problem

1.1 Background

Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.



According to World Health Organization Fact sheet details

- 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years
- More than half of all road traffic deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists.

1.2 Problem

Road accidents are serious concern for majority of nations around the world. The purpose of this project is to predict the severity of any road, which will play a crucial factor for various Government Departments/Authorities like Police, R&B and Transport to take proactive precautionary measures.

1.3 Interest

Of course! Road accidents cab be prevented. The prediction aim for sustainable development, has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2021. Others, who are interested to reduce the accident impact, claims and to improve the Road safety such as Insurers, Organizations and Public Persons may also be interested.

2. Data Description

2.1 Data Source

The data come from “Kaggle” the world's largest data science community and also offers public data platform, where they have been published.

The dataset comprises of two csv files and one xls file:

Accident_train.csv: the training dataset

Accident_test.csv: the test dataset

AttributeLevelsDescription.xls

The dataset contains 8849 records and 17 columns in Accident train and 1549 records and 17 columns in Accident_test, which includes the data attributes such as Collision Reference Number, Policing area, Collision Severity etc... For good description of each attribute, you can refer to the Attribute Levels Description. Some or all can be used to train the model.

2.2 Data Pre-processing

To solve the problem our dataset should have balanced labels. Our dataset is very imbalanced with data corresponding to 3: Slight injury collisions is 89.23%, 2: Serious injury collision is 9.52% and 1: Fatal injury collision is 1.25%. In Data pre-processing, the data set is imputed by replacing NAN and missing values with most frequent values or median of the corresponding columns such as Junction_Detail, Junction_Control, Road_Surface_Conditions and Ped_Crossing_PC. All categorical variables are labeled by integers using Label encoder for each column.

Collision_Severity

1: Fatal injury collision, 2: Serious injury collision and 3: Slight injury collision

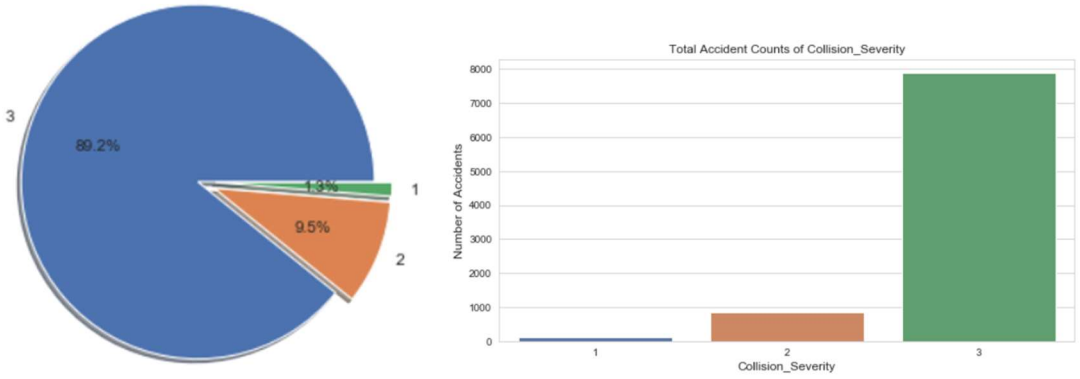


Fig 1: Accident occurrences of Collision Severity

The data is visualized for correlation. Negatively correlated features are selected to be dropped. Feature importance is plotted to visualize and only features with high importance are taken into consideration for accident severity.

2.3 Feature Selection

After data pre-processing, simply by reading the definition of each feature from the pdf mentioned above, we can see that there are some features whose values are derived from other features. There are mixed types of data, drop Collision Reference No. column, which is only Identity. For selecting the best features, functions from sklearn library is used. After all, finally 15 features were selected.

3. Methodology

Exploratory Data Analysis

3.1 Examining the Environmental conditions

First I chose to examine the lighting conditions, road surface conditions, and weather conditions in which each accident occurred, and their timing on the train dataset.

<u>Lighting, Road and Weather Conditions:</u>

All of these variables showed a similar distribution (see Figure 1), with one category accounting for 25–63% of observations, a secondary category making up 19–32%, and the remainder covered by uncommon conditions. This is unsurprising as, although data on weather conditions were not available, we know that in the Northern Ireland it

is generally dry and clear, but sometimes rainy, and more driving is done during daylight hours.

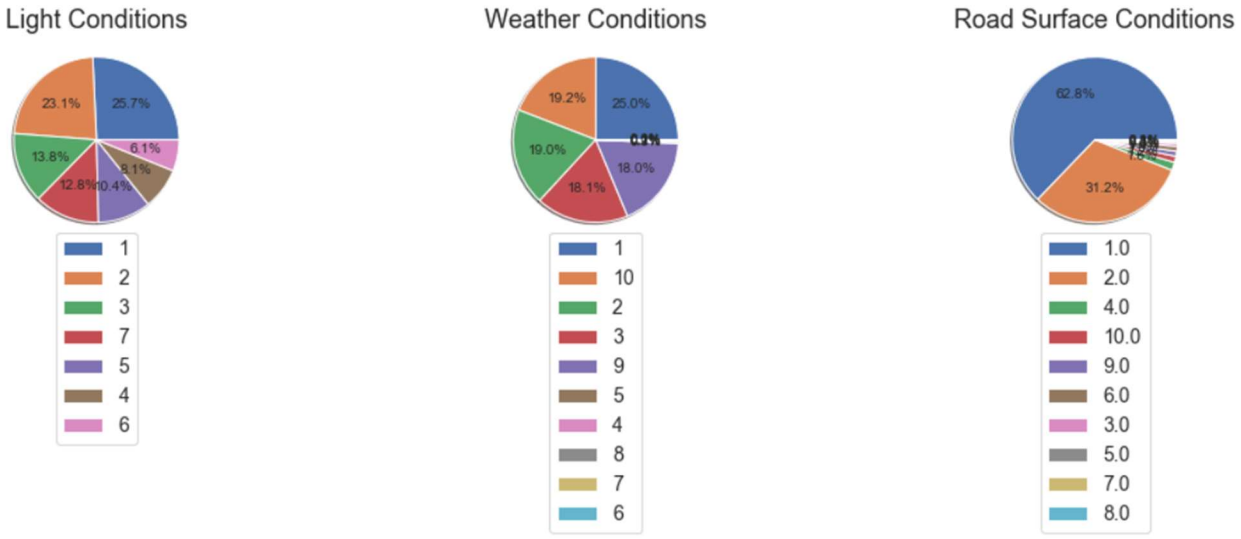


Fig 2: Accident Occurrence of Environmental Conditions

Lighting Conditions 1 : Daylight : street lights present,2 : Daylight : no street lighting,3 : Daylight : street lighting unknown,4 : Darkness : street lights present and lit,5 : Darkness : street lights present but unlit,6 : Darkness : no street lighting and 7 : Darkness : street lighting unknown

Weather Conditions 1 : Fine without high winds,2 : Raining without high winds,3 : Snowing without high winds,4 : Fine with high winds,5 : Raining with high winds,6 : Snowing with high winds,7 : Fog or mist - if hazard,8 : Strong sun (glaring),9 : Other and 10 : Unknown

Road_Surface_Conditions 1 : Dry, 2 : Wet / damp,3 : Snow,4 : Frost / ice,5 : Flood,6 : Oil,7 : Mud,8 : Leaves,9 : Slippery (after dry spell) and 10 : Other

3.2 Calculate the Accident Distribution across Time:

The next thing I wanted to look for was a seasonal pattern. Grouping incidents by date produced the plot shown in Figure 2. Although the variance is slightly higher in winter months, there is very little change in the overall rate throughout the year (less than 5% standard deviation in monthly mean), as indicated by the dashed regression line. Taking in to account the change in traffic volume by month, the standard deviation is still only 8%.

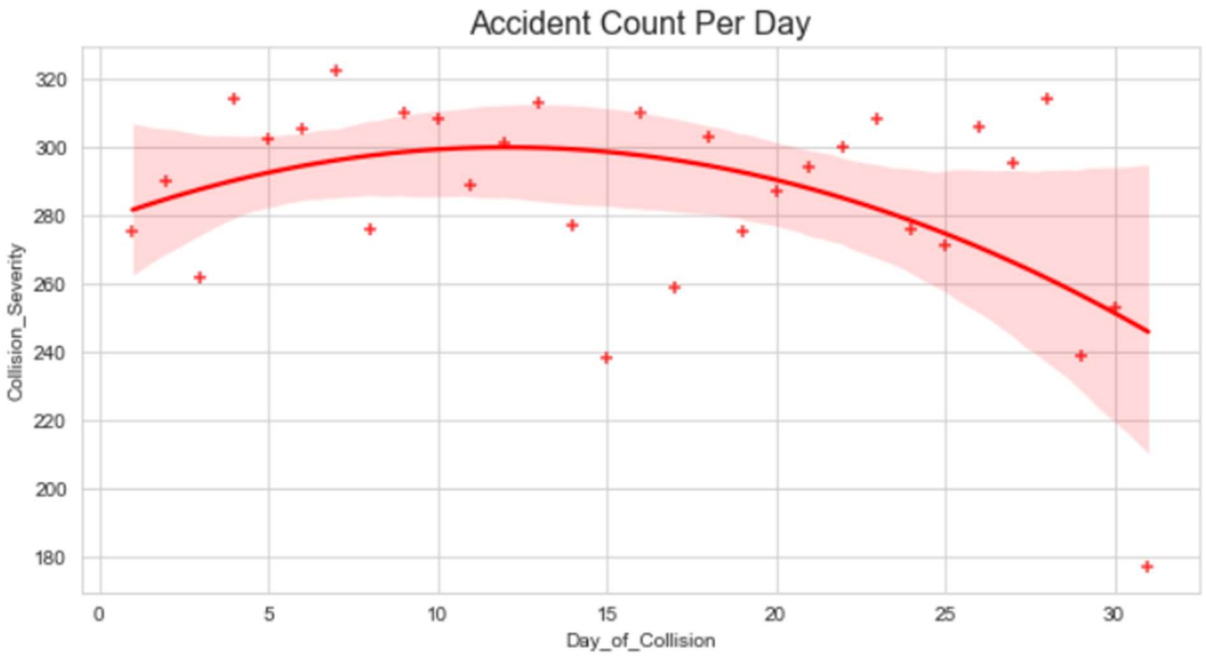


Fig 3: Day wise Distribution of Accidents throughout the year

One point of interest is the one outlier in late December. The day of the year with the lowest number of crashes: New Year's Eve.

Grouping incident by time of day showed unsurprising peaks during the morning and evening rush hours, and a far lower rate during the night than the day.

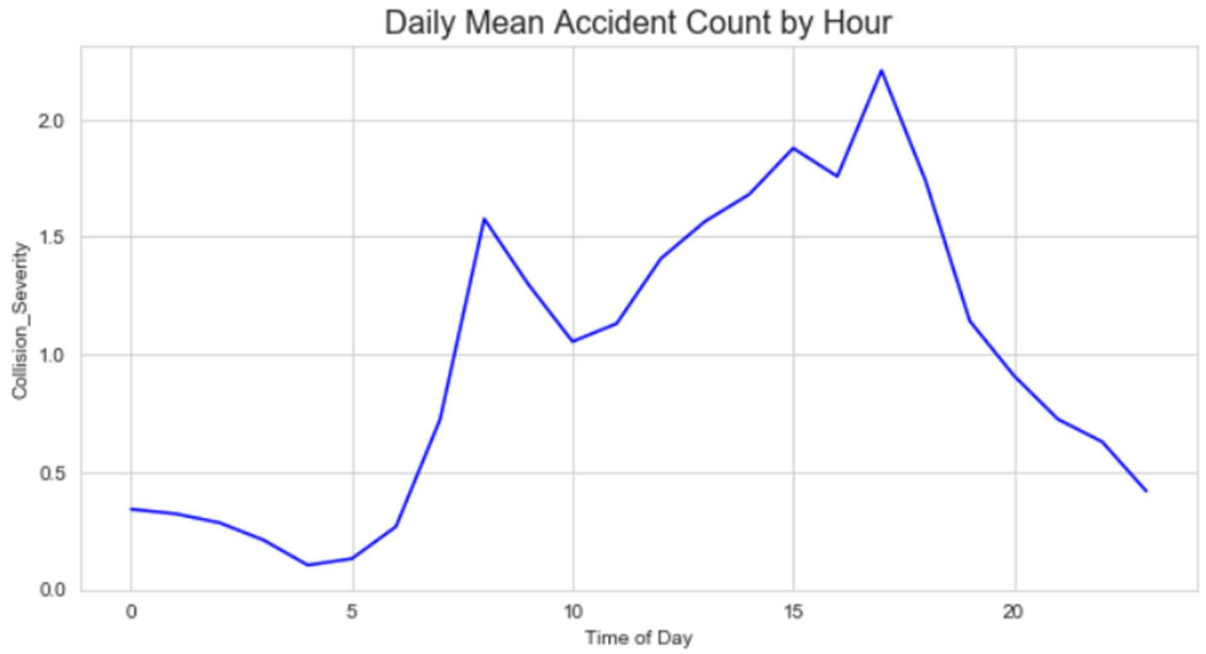


Fig 4: Hourly Distribution of Accidents throughout the day

3.3 Calculating the Accidents on Weekday and in Policing Area

Grouping incident by Week day showed high counts on Friday and low counts on Sunday because of holiday on that day.

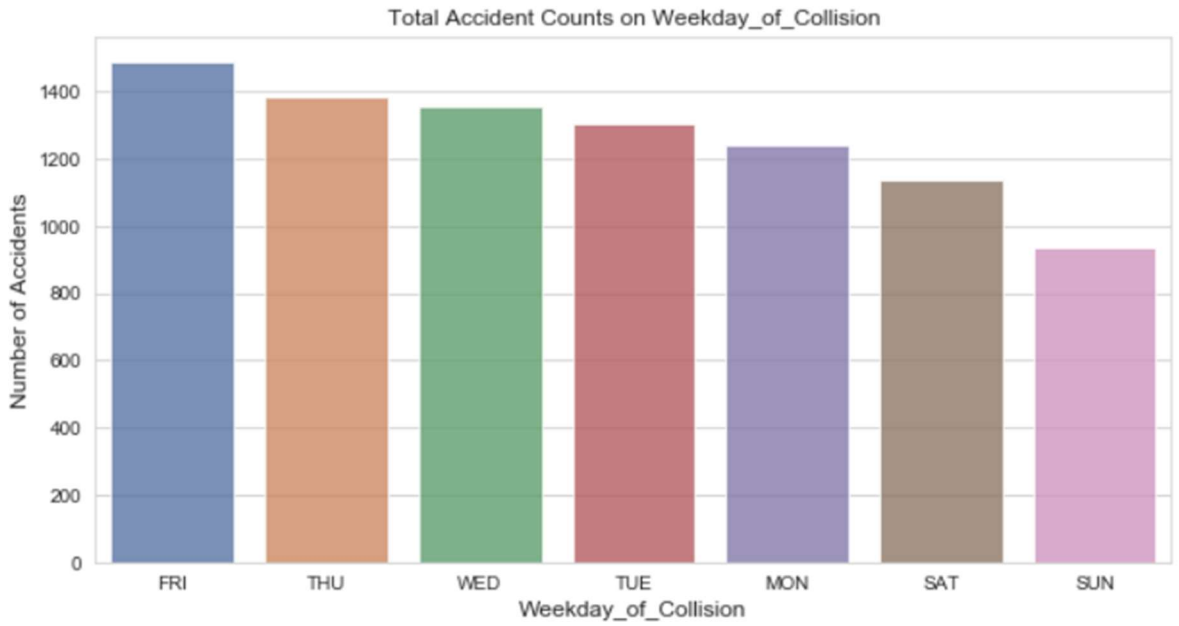


Fig 5: Accidents Occurrences of Weekday

Grouping incident by Policing Area depicted that BELC: Belfast City, had the highest accidents. It is the capital and largest city of Northern Ireland with inhabitants of 0.28 million.

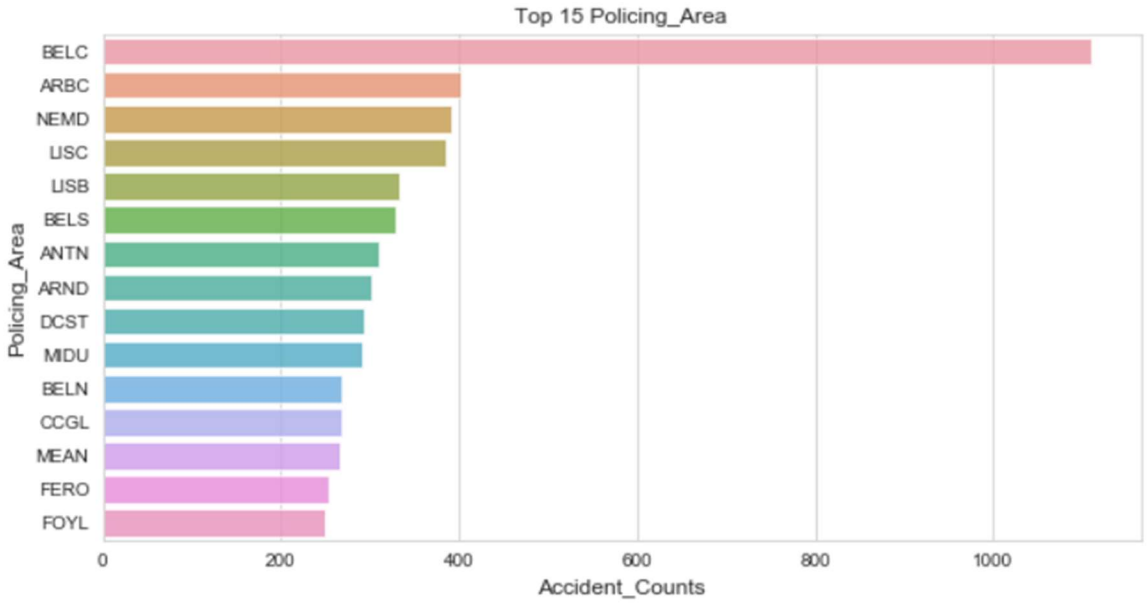


Fig 6: Accidents occurrences in Top 15 Cities

3.4 Correlation Matrix

The next thing I wanted to quickly check correlations among columns is by visualizing the correlation matrix as a heat map. Dark blue means positive, light pale blue means negative. The stronger the color, the larger the correlation magnitude. The Collision severity has a positive correlation with Weather conditions and negative correlation with Speed limit.

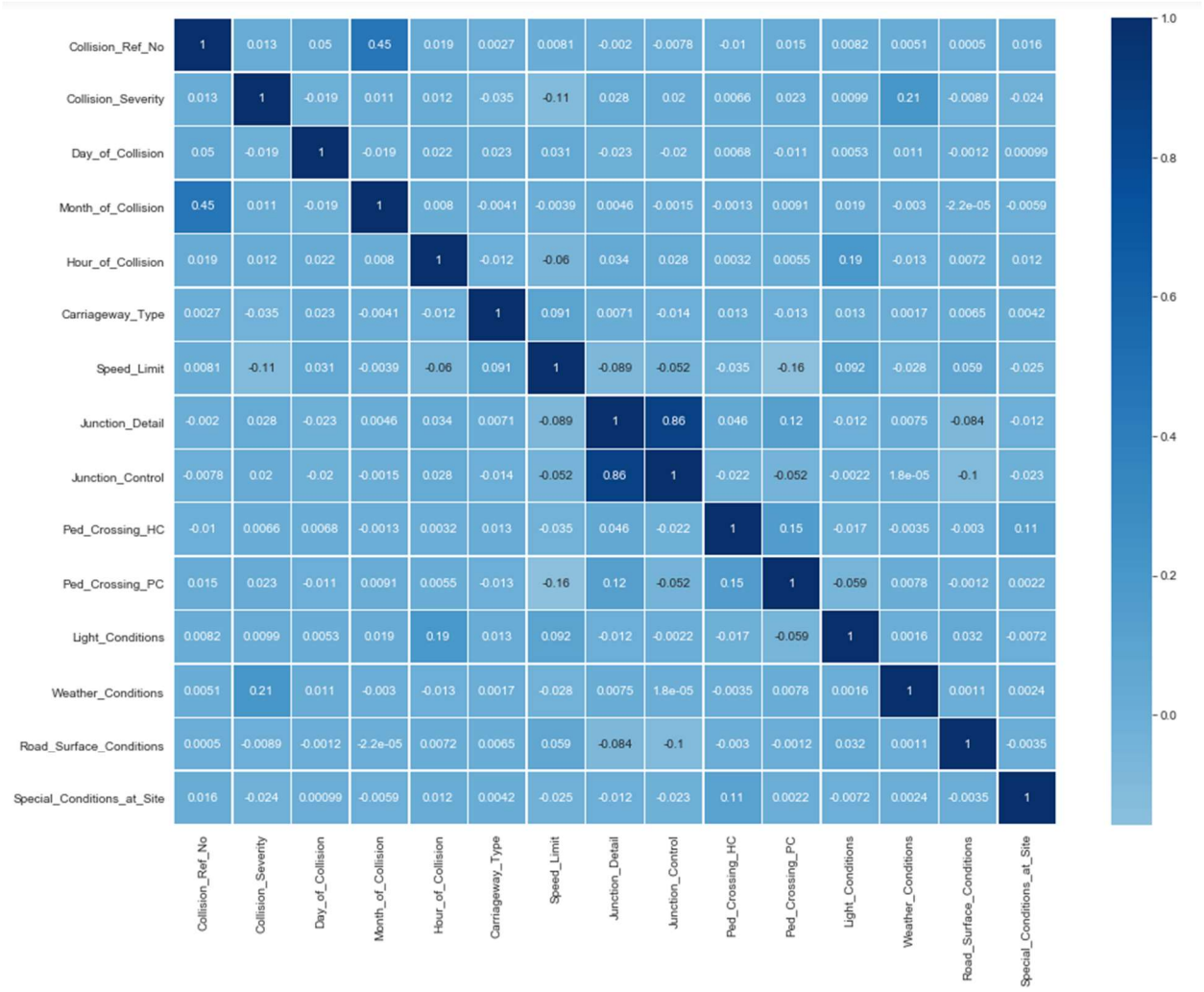


Fig 7: Correlation Matrix

4. Predictive Modeling

4.1 Machine Learning Models

Before model building, I have transformed all the categorical features using Label encoding method and normalized the datasets. For modeling I have spilt the dataset into 80-20 ratio using train test split method. I.e. 80% as train set and 20% as test set.

Five classification algorithms were used and evaluated to predict the accident severity. Algorithms considered for classification were XGBoost classifier, Random Forest classifier, Support Vector Machine classifier, Logistic Regression and Gaussian Naïve Bayes classifier.

The first thing came to mind is that severity is based on different decisions like Weather conditions and their timing. So I tried ensemble methods as the data is very imbalanced.

XGBoost Classifier: As a part of trying out different ensemble methods, I selected one of the most popular distributed gradient boosting technique. This performs faster compared with other algorithms due to parallel tree boosting method. The prediction score of this algorithm is 88.64%.

Random Forest Classifier: Another ensemble method classifier, which is a meta estimator that fits a number of decision tree classifier various sub samples of the dataset and uses averaging to improve the predictive accuracy and control over fitting. The prediction score of this algorithm is 88.42%.

Support Vector Machine Classifier: SVMs are based on the idea of finding a hyperplane that divides the data set into two classes. For larger training dataset SVM training time is high. The prediction score of this algorithm is 88.59%.

Logistic Regression: This model is basic and popular for solving classification problems. Logistics regression uses sigmoid function to deal with outliers. Class weight parameters sets the weights for imbalanced classes by adjusting weight inversely proportional to class frequency. The prediction score of this algorithm is 88.59%.

Gaussian Naïve Bayes: This model is probabilistic frame work for solving classification problems. This model did not perform well because of dependent features. The prediction score of this algorithm is 21.41%.

Model	Score	Count
XGBoost	88.64	{3: 1664, 2: 104, 1: 2}
Support Vector Machines	88.59	{3: 1770}
Logistic Regression	88.59	{3: 1770}
Random Forest	88.42	{3: 1740, 2: 29, 1: 1}
Naive Bayes	21.41	{1: 1365, 2: 43, 3: 362}

Fig 8: Report Table

5. Discussion of Results

The expectation, going in to this investigation, was that driving in the dark and in poor weather was dangerous, and that this would be evidenced by a pattern of higher crash rates during the winter, and in areas which experience generally colder and wetter conditions (the north and west of the country), but this is not supported by the results. There is no significant change in the rate of traffic accidents throughout the year; the monthly crash rate varies very little, and December and June have been shown to have very similar hourly patterns of incidents. The analysis also found no connection between the proportion of accidents occurring in different lighting conditions and the rate or severity of those accidents.

6. Discussion of recommendations

Road traffic injuries can be prevented. Governments need to take action to address road safety in a holistic manner. This requires involvement from multiple sectors such as transport, police, health, education, and actions that address the safety of roads, vehicles, and road users.

Effective interventions include designing safer infrastructure and incorporating road safety features into land-use and transport planning, improving the safety features of vehicles, improving post-crash care for victims of road crashes, setting and enforcing laws relating to key risks, and raising public awareness

7. Conclusion

Although the results have been somewhat surprising, the objectives of the investigation were met, but there are still many areas of the data which could be investigated further, such as how junction layout or vehicle type relate to collision rates in different conditions.