Prognosis And Analysis Of Accidents And Severity

1st N Meghana Reddy
Computer science and engineering
PES UNIVERSITY
Bangalore, India
meghana.narpala@gmail.com

2nd Harshapriya C R
Computer science and engineering
PES UNIVERSITY
Bangalore, India
harshapriya112@gmail.com

3rd Priyanka G
Computer science and engineering
PES UNIVERSITY
Bangalore, India
priyankagopi86@gmail.com

***Abstract*—Reducing traffic accidents is one of the most crucial public safety challenges, so analysis and prediction of road accidents have been a trending topic of research in most recent years. The enterprising approach to deal with traffic safety problems is to focus on avoiding probable unsafe routes. For the fruitful implementation of this approach, accident prediction and severity prognosis are critical. The motto of the project is to examine and study data of 50 states in the US and to inform the general public and Government officials about trends and possible causes of traffic accidents and what could be done to reduce them**

***Index Terms*—Accident Prediction, Severity Estimation, Supervised Learning,US Accidents**

I. Introduction

Traffic and Road accidents are one the major issue in most of the countries. Around 1.35 million lives are lost per year because of accidents. The latest report on road safety given by ' World Health Organization (WHO) ' says 65 percent of deaths due to accidents are from high-income nations like the USA. This costs nearly 3 percent of GDP in most of the countries. This is one of the most impactful non-natural leading causes of death. Approximately, 1,72,500 road accidents occur each day across the country. Thousands of pedestrians, children, animals, and travelers are losing their lives which includes not only physical damage and property damage but also emotional damage. Reducing road mishaps is an essential public safety challenge all over the world. Therefore, accident analysis has been a subject of much research in recent decades. So this paper is about the analysis of accidents and prediction of severity. Over the years, various studies have utilized various data-sets that are either private or not easily accessible. Regardless of efforts, results were not that effective. The main thing about this data-set is it is publicly available. By employing the US-Accidents data-set and through an extensive set of experiments and processing millions of records of accident cases across nearly 50 cities to draw out conclusions that could help government and to find major factors that affect the severity of the accident and predict severity by developing various models of supervised learning like Random forests classifier, decision trees classifiers, and logistic regression.

II. Literature Survey

The most relevant predecessor work that we were comparable to ours was about the unfortunate/lamentable intermediary entangled in the accidents that happen on the road which we know is increasing day by day. The vehicles catastrophes is a crucial communal protection matter with the utmost exploration dedicated to research and forecast of the well known infrequent incident. Almost all the papers that we have read so far sight to interrogate the accidents that take place on road. Extensive exploration is performed based on the data types of accidents give rise to the sexuality of which person was driving throughout the dusk and the dawn time with their steering incident and most catastrophes occur due to drunk driving and contravention superior to accidents. Descriptive analysis, single and multivariate regression analysis is conveyed out to high point enormous exposed characteristic like 'age', 'sex', 'circumstance', 'breach kind' etc to recognize their separate and cumulative results on catastrophe. The most common indistinguishable thing that we read from the papers was the catastrophes are likely to happen during the nights than during the day as the majority common basis is shallow clarity. The analysis that is mostly used to train and develop the predictive models are regression analysis,deep Neural Network, casualty analysis, random forests, decision trees, logistic regression, and random forest. The prediction helps us to recognize the universal study here is fragmented into few parts:

- To visualize the present catastrophes circumstances in US like how any number of accidents happening each year, whether the catastrophes that is taking place is increasing or decreasing, the severeness of the catastrophes and outline.
- Discourse the mixture to the public as well as state government, of how to decrease these catastrophes and what all solutions can be executed by each state established on decisive elements.

The main area we seek to solve is :

- To find major elements that influences accident severity.
- To build a model that can accurately predict accident seriousness. To be particular for a given catastrophe without complete details about itself like vehicle type or driver attribute our model can predict the likelihood of this catastrophe being a severe one
- Model should predict severity of likelihood.

Limitations are:

- Only this data-set is publicly accessible while others are restricted/inaccessible.

- The seriousness in this data is "a manifestation of the result the catastrophe has on traffic" preferably than injury seriousness that has been already read by enormous articles.
- The accident data might not be completely available

III. Proposed Solution

The main objective of the model is recognizing the key factors that are influencing the severity of the accident and correctly predicting the accident severity. The likelihood of the accident being a severe one is predicted by the model. The creation of the data-set is done with the advanced real-time traffic accident prediction that can be used for additional prediction of severe accidents in real-time.Patterns of how serious accidents can happen and the crucial factor, we should be able to implement acquainted steps and finer allocation of economic and human resources. The analysis and results showed notable improvements to predict severe accident events.Data cleaning was first accomplished to detect and handle corrupt or missing records. Exploratory Data Analysis and feature engineering were then done for most of the features. Finally, Random Forest Classifier, Logistic regression, and Decision Tree Classifier were used to develop the severity prediction model. The final model achieved 99.0 percent test accuracy on re sampled data.

A. Data Sources And Collection :

There are over 3 million records in this data set consisting the information about accidents that happened between February 2016 to December 2019. Data set creators collected streaming traffic data from "Map Quest Traffic API" and "Microsoft Bing Map Traffic API". This API live-stream the traffic events like accidents, congestion on the road recorded by a group of entities - traffic sensors, law enforcement agencies, traffic cameras, US and state departments of transportation, and within the road-networks.

B. Overview of the data set :

- Details about the features in the data set :
  There are 12 Traffic attributes, 9 Address attributes, 11 Weather attributes, 13 Point of interest features, and 4 periods of day features. Total Accidents - 2,243,939 , accidents captured by Map Quest 1,702,565 that is 75.9 percent, Bing captured 516,762 accidents that is 23 percent.
  The top States Where accidents occurred more are California (485K), Texas (238K), Florida (177K), North Carolina (109K), New York (106K).

C. Data cleaning :

While the data set has over 3 million records, it is not yet ready to use for analysis. US-accidents data set hold in total, which is a little bit too much. We removed some of the columns and combined them into one. Such as the features containing points of interest are merged into one. There are many irregular data in the data set like Null records, Date time type format, missing data, multiple columns representing the same data. To label all these irregularities in data, data cleaning is mandatory and is the most important step. Data cleaning was first done to find out and handle missing records, incomplete data, required features for analysis.

- Feature Selection :
  Sorting out to the point features that contribute to the prediction of risk was a major task as it plays a major role in prediction accuracy.
- Useless features :
  Feature ID doesn't impart any useful information about accidents. 'End_Time', 'Distance', 'TMC', 'Duration, 'End_Lng', 'End_Lat' and start location can only be collected only after the accident has happened and thus these features cannot be predictors for severity of accident prediction. Data set creators have already extracted the POI features for 'Description'.When it comes to categorical features, 'Turning_Loop', 'Country' have only one class thus they are dropped. And then 'Start_Time' was mapped to 'Year', 'Month', 'Minute'(in a day), 'Weekday', 'Day' (in a year), and 'Hour'.
  Considering the percentage of the count of missing values in each feature, missing values have been dropped from the data set and for some of the features where only some part of the values were filled. Features having more than 60 percent of missing data, were dropped.

D. Exploratory data analysis :

Only the data of accidents that happened after February 2019 and the accidents which were reported by MapQuest was finally used in exploratory data analysis and modeling. Hence the irrelevant features, values can be eliminated to the greatest amount. Based on the exploration analysis done in the accidents with severity level 4 are much more severe than accidents of other levels, between which the severance is far from definite. Hence, we focused on severity level 4 accidents and reorganized the levels of severity into level 4 against other levels. Severity specifies the effect of the accident. Severity 3 and 4 accidents involve mostly the causalities, severity 2 involves the accidents having wounds. Severity 1 and 2 accidents were reported essentially due to the insurance claims.

E. Model :

Data is split into Train and test data. After re-sampling and splitting the data, the training data size is about 64000, the data size of the test data is about 16000. Features are standardized based on unit variance. Different models were evaluated on their ability to predict the severity of accident prediction.

A notable number of earlier studies influenced regression-based models to perform accident prediction. Therefore, we implemented logistic regression as a suitable baseline to perform our binary classification task. Therefore, our first model used to predict the severity of accidents is the Logistic regression model.

Logistic regression is a statistical model and a supervised learning model that uses the logistic function to model a

binary dependent variable. It outputs the probabilities for classification problems with two possible outcomes. Logistic regression yielded decent results, where the training accuracy on the training data resulted is 91.7%, and test accuracy is 92.1%.
The confusion matrix compared the actual outcomes to the predicted outcomes predicted from the logistic regression model. For evaluating the classification models used, we have used accuracy as our evaluation metric.

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + False\ positive + True\ negative + Fasle\ negative}$$

Fig. 1. Evaluation metric

Then we employed a decision tree classifier which is a supervised learning model. Decision trees are effective in performing multi-class classification on a data set. On training the data using a decision tree classifier, the training accuracy was improved to 100% and test accuracy was improved to 99.7%. The result yielded from the decision tree classifier is nearly perfect. Important feature plot was plotted after training and testing the data using decision trees. The plot showed that apparent patterns of Spatio-temporal accidents are the most useful features to predict severity.
Out of them, street frequency is more dominant than other features. Some other features like a traffic signal, interstate highway are also important. Besides these Spatio-temporal features, weather features like wind speed, humidity, pressure, temperature and are also very important.

Finally, Random forest was employed, which is also a supervised learning model. It is an ensemble learning method used for classification, regression, and also used for other tasks. They work by constructing multiple decision trees at the time of training the data and outputs the class which is taken by considering mode or average/mean prediction of individual decision trees. The random forest classifier model attained 99.6% train accuracy and 99.5% test accuracy. The results are good at prediction and can be used without any further doubts.

IV. Experimental Results

We have tested a huge data of various records of accidents occurred in states of United nations. Millions of records of accident cases data are used to train Machine learning models. Supervised learning models like Logistic Regression, Decision Trees, Random forests are used to predict the severity of accidents and find major factors that affect accident severity.Highest accuracy i.e., 99.7 percent is gained with Decision tree classifier.

A. Logistic Regression

'Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable'.
In regression analysis, logistic regression(or logit regression) is used to estimate the parameters of a model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".
Here , Logistic regression was employed as a baseline to perform binary classification task.Various attributes like temperature,longitude and latitude coordinates,weather conditions are taken into account for prediction of accident severity using linear regression model.Train accuracy gained by this model is 91.7 percent and test accuracy achieved is also 92.1 percent.

TABLE I
Logistic Regression

Train / Test data	Accuracy Gained in percent
Train Accuracy	91.7
Test Accuracy	92.1

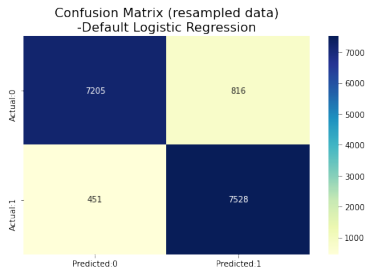


Fig. 2. Confusion Matrix for Logistic Regression

B. Decision tree

A Decision tree is one of the most widely used machine learning algorithm that splits the data into subsets.The splitting procedure starts with a binary splits and continues until no further partitions can be made. The main aim of Decision tree is to fit the given training data in as small as possible tree.The logic behind the goal of achieving smallest possible tree is the sense that uncomplicated ,simple answer is chosen over the detailed and complicated explanation.Achieving smallest decision trees also helps in finding results much faster than larger ones."It is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility".

Here we used Decision tree model to predict the seriousness of catastrophe.Nearly 30 features which include Accident zone details,Location attributes like longitude and latitude, various weather conditions like humidity,temperature,wind chill etc are used to predict the severity of catastrophe.A bar plot of feature importance is also provided in EDA section. Decision Tree classifier,on successful training with data, train accuracy achieved is 100 percent.The test accuracy obtained by Decision

tree is 99.7 percent which is better than logistic regression classifier

TABLE II
DECISION TREE

Train / Test data	Accuracy Gained in percent
Train Accuracy	100
Test Accuracy	99.7

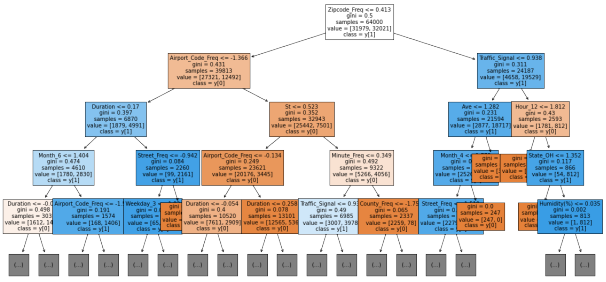


Fig. 3. Decision Tree

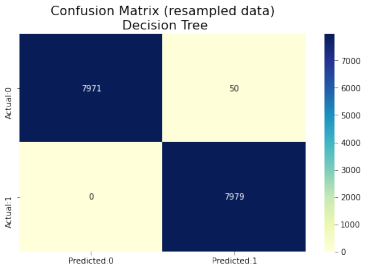


Fig. 4. Confusion Matrix of Decision Tree

C. Random Forest

Another Supervised learning model is 'Random Forest'.These are often referred as an ensemble method of learning for classification, regression and other tasks that work by constructing a multitude of decision trees at training time and returning the class that is the mode or mean of each tree Random forests are most widely used as "black box" models in businesses, because they produce reasonable predictions across a wide range of data with minimum requirements in software tools like scikit.

Random Forest is another model we used for US traffic and road Accidents analysis and Prediction.Same features which are used for decision tree model are used for random forest classifier whereas the order of feature importance changes when it comes to random forests.99.6 percent of training accuracy is reached and we have gained 99.5 percent if test accuracy.

Key Findings

- Find major factors that influence Accident Severity.
- Road Accident severity can be predicted accurately when we have only limited features also

TABLE III
RANDOM FOREST

Train / Test data	Accuracy Gained in percent
Train Accuracy	99.6
Test Accuracy	99.5

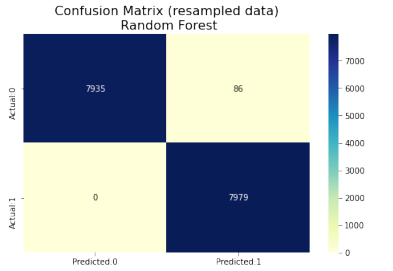


Fig. 5. Confusion Matrix of Random Forests

- Serious accidents are more likely to happen at areas having more accidents whereas for larger areas,less severity
- There is a 2 percent chance for a catastrophe to be a serious one when it occurred on National Highway .This is nearly 2.3 times of the mean and greater than all other street types.
- If a road accident occurs at junction or signal where people will be present around,it is less likely to be severe .When the accident occurs at outskirts or deserted area it is more likely to be severe.
- After deep analyzing of data, Weather attributes like temperature, pressure , wind chill , humidity etc.
- Time also plays important role in prediction of seriousness.

V. CONCLUSION

Thousands of lives are lost due to road accidents.There is an urge to stop these catastrophes.By knowing the main causes for accident severity. Government can take actions accordingly to overcome the loopholes in infrastructure and management.So interpretation and prediction of Severity of accident is essential.There are vast number of researches going on this topic.But most of the studies use small-scale data sets, and some of them are not readily available and are not publicly accessible to researchers. Due to this, they are not able to apply the research for real-time predictions. We used a publicly accessible nationwide traffic accident data set, named US-Accidents. By using this data, we built some of the classification models. Our prediction model incorporated logistic regression, decision trees, random forest. The Decision tree gave the best accuracy out of all other models. These models used various features such as time information, weather data, points-of-interest, traffic events, address, the period of the day to predict the accident severity.Additionally, there were different categories of data attributes, and we studied which features among them help in predicting the severity of traffic accidents and found that points-of-interest, traffic events, and time showing and has

significant value. Henceforth, in the future, we plan to include other publicly accessible and available sources of data such as annual traffic reports and demographic information for the task of real-time prediction of traffic accident severity.

Team Contribution

Harshapriya C R - 1774 : Literature survey on a paper about Accident Risk Prediction based on Heterogeneous Sparse Data.Worked on logistic regression model , pre-processing of data is handled and took part in preparing IEEE format paper.

N.Meghana Reddy - 1272 : Literature survey on a paper about Analysis of US accidents and Solutions.Worked on Random Forest model , Exploratory data Analysis is handled and took part in preparing IEEE format paper.

Priyanka G -1797 : Literature survey on a paper about Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis.Worked on Decision Tree model , Exploratory data Analysis is handled and took part in preparing IEEE format paper

REFERENCES

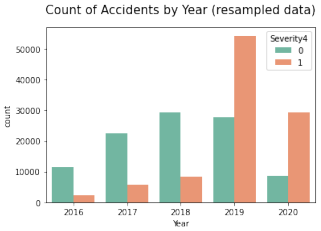
[1] Imran Ashraf, Soojung Hur, Muhammad Shafiq and Yongwan Park, “Catastrophic factors involved in road accidents: Underlying causes and descriptive analysis,”, October 9 2019.

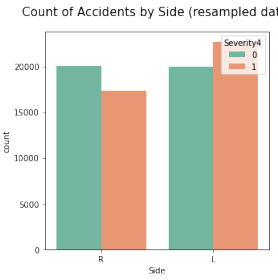
[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath, “A Countrywide Traffic Accident Dataset,”, 19 Sep 2019.

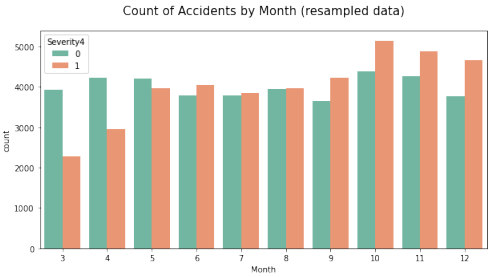
[3] Swapnil Kisan Nikam , “Analysis of US Accidents and solutions,” 3-2020.

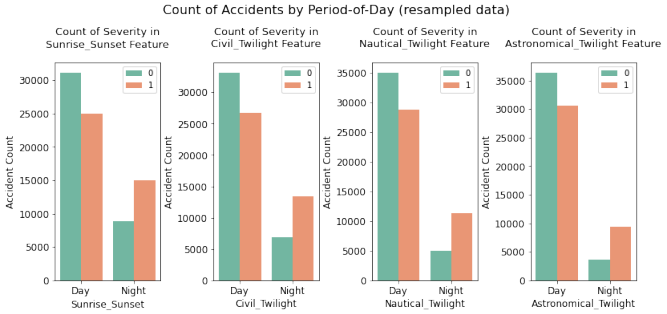
VI. APPENDIX

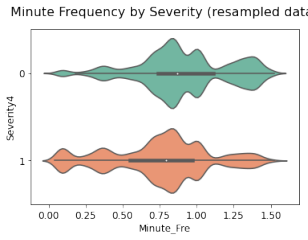
Visualizations

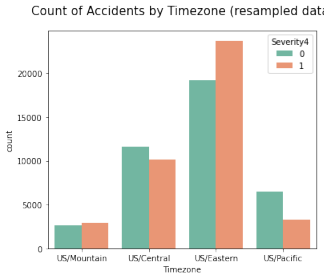


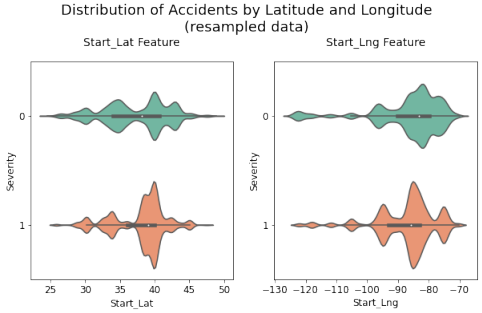


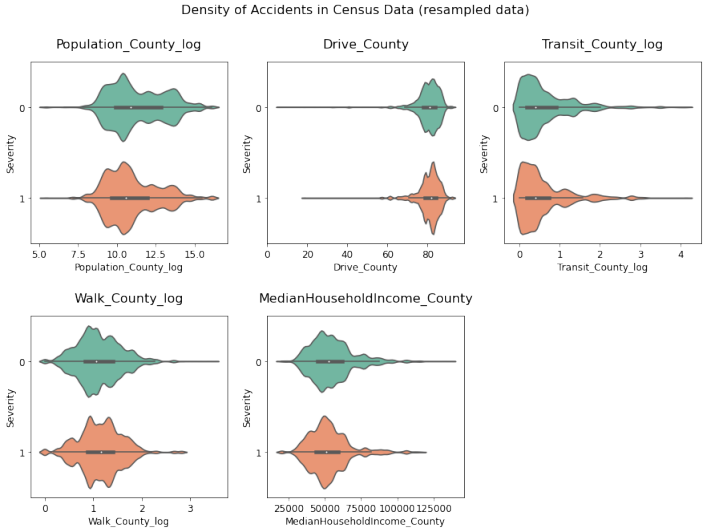


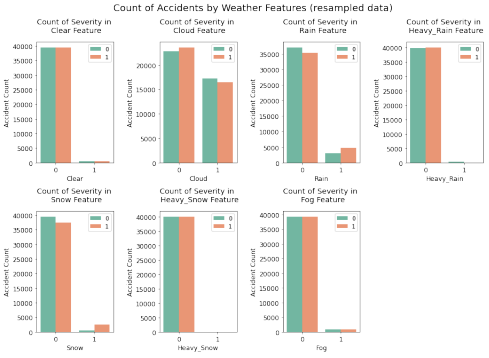


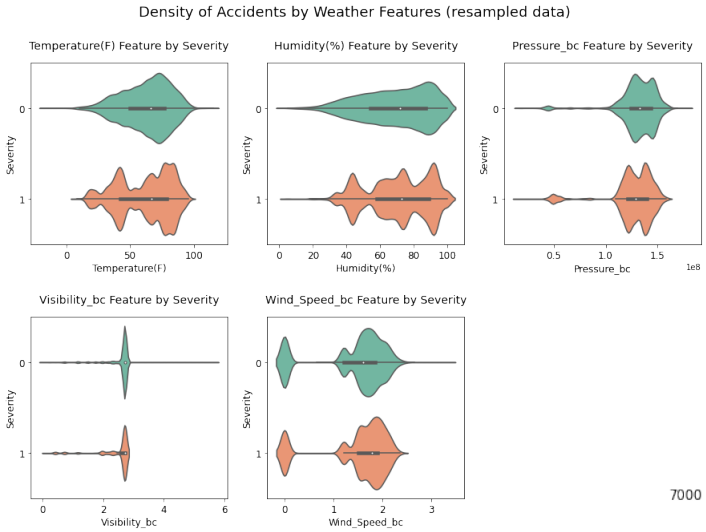


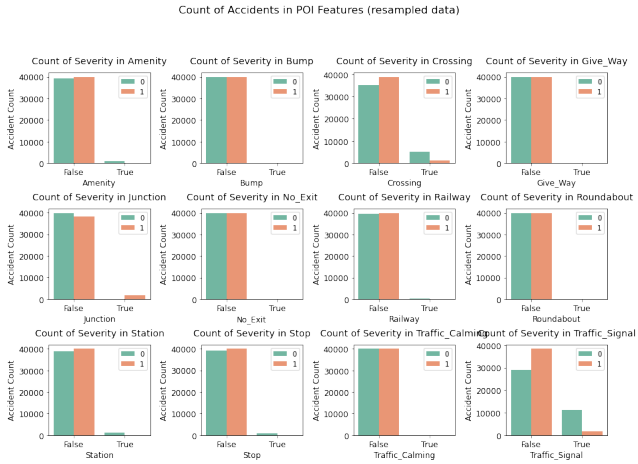


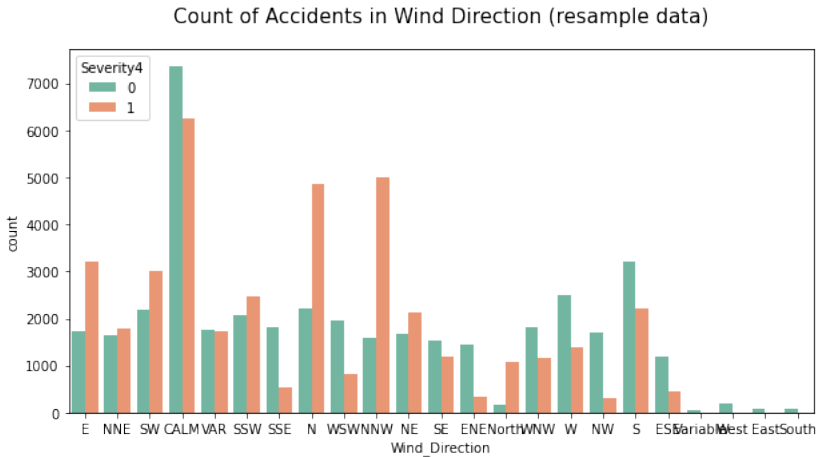


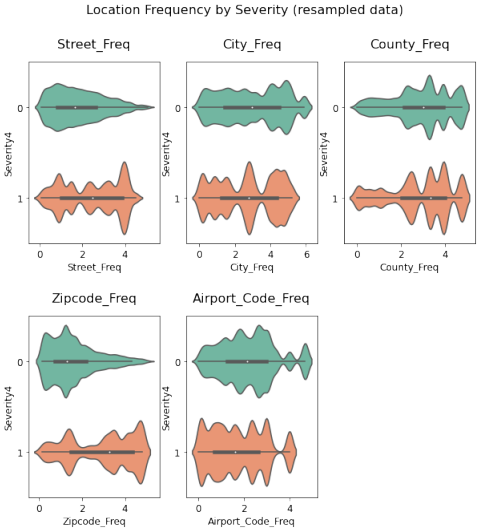

Density of Accidents in Census Data (resampled data)
Population_County_log
Drive_County
Transit_County_log
Walk_County_log
MedianHouseholdIncome_County
Severity
Population_County_log
Drive_County
Transit_County_log
Walk_County_log
MedianHouseholdIncome_County

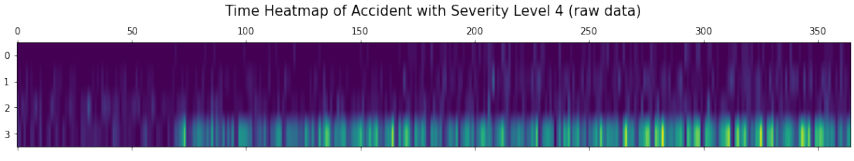

Count of Accidents by Weather Features (resampled data)
Count of Severity in Clear Feature
Count of Severity in Cloud Feature
Count of Severity in Rain Feature
Count of Severity in Heavy_Rain Feature
Count of Severity in Snow Feature
Count of Severity in Heavy_Snow Feature
Count of Severity in Fog Feature
Accident Count
Clear
Cloud
Rain
Heavy_Rain
Snow
Heavy_Snow
Fog

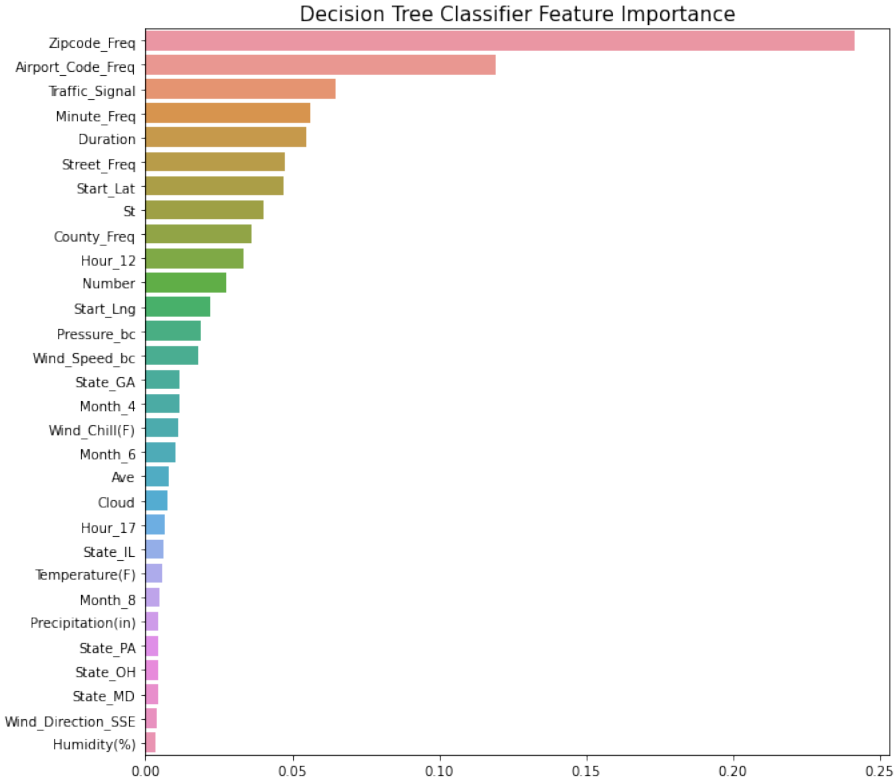

Density of Accidents by Weather Features (resampled data)
Temperature(F) Feature by Severity
Humidity(%) Feature by Severity
Pressure_bc Feature by Severity
Visibility_bc Feature by Severity
Wind_Speed_bc Feature by Severity
Severity
Temperature(F)
Humidity(%)
Pressure_bc
1e8
Visibility_bc
Wind_Speed_bc

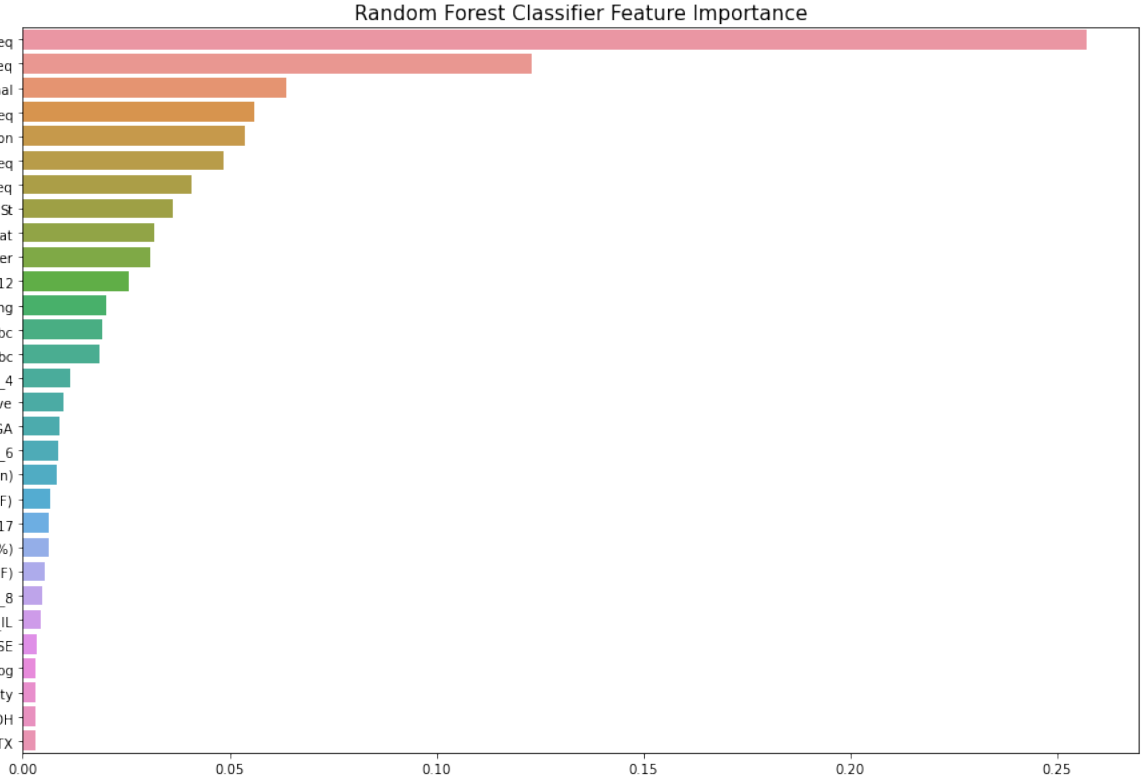

Count of Accidents in POI Features (resampled data)
Count of Severity in Amenity
Count of Severity in Bump
Count of Severity in Crossing
Count of Severity in Give_Way
Count of Severity in Junction
Count of Severity in No_Exit
Count of Severity in Railway
Count of Severity in Roundabout
Count of Severity in Station
Count of Severity in Stop
Count of Severity in Traffic_Calming
Count of Severity in Traffic_Signal
Accident Count
False
True
Amenity
Bump
Crossing
Give_Way
Junction
No_Exit
Railway
Roundabout
Station
Stop
Traffic_Calming
Traffic_Signal


Count of Accidents in Wind Direction (resample data)
Severity4
0
1
count
E
NNE
SW
CALM
VAR
SSW
SSE
N
WSW
NNW
NE
SE
ENE
North
WNW
W
NW
S
ESE
Variable
West
East
South
Wind_Direction


Location Frequency by Severity (resampled data)
Street_Freq
City_Freq
County_Freq
Zipcode_Freq
Airport_Code_Freq
Severity4
Street_Freq
City_Freq
County_Freq
Zipcode_Freq
Airport_Code_Freq


Time Heatmap of Accident with Severity Level 4 (raw data)


Decision Tree Classifier Feature Importance
Zipcode_Freq
Airport_Code_Freq
Traffic_Signal
Minute_Freq
Duration
Street_Freq
Start_Lat
St
County_Freq
Hour_12
Number
Start_Lng
Pressure_bc
Wind_Speed_bc
State_GA
Month_4
Wind_Chill(F)
Month_6
Ave
Cloud
Hour_17
State_IL
Temperature(F)
Month_8
Precipitation(in)
State_PA
State_OH
State_MD
Wind_Direction_SSE
Humidity(%)
0.00
0.05
0.10
0.15
0.20
0.25


Random Forest Classifier Feature Importance
0.00
0.05
0.10
0.15
0.20
0.25