Data Mining to Analyze and Prevent Vehicular Crashes

Dhruv Shetty, Jean Nei, Sameer Desai, Sushmita Chaudhary, Bhaskar Vemuri

April 2020

ABSTRACT

In 2018, there were 33,654 fatal motor vehicle crashes in the United States alone which resulted in 36,560 deaths.[1] More specifically there were 11.2 deaths per 100,000 people and 1.13 deaths per 100 million miles travelled.[1] Consequently, road safety is a top priority in many places with governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety. One of those technologies is big data, and it's already generating positive results and saving lives.

The aim of this project is to perform analysis on this data based on various attributes, like severity of the accident, timezone, time taken to clear the traffic, variation with respect to the state, weather conditions and place of occurrence. We also introduce association rules to analyze the factors which are related to each other and causing the accident.

This project aims to apply clustering techniques on various attributes like severity, side and city to choose the 'k' value using both SSQ and Gap statistic techniques. After performing the analysis, this paper would do several supervised and unsupervised learning techniques like Logistic regression, Decision trees, SVM and Neural Networks and perform a comparative study of those methods based on the accuracy.

I. INTRODUCTION

One of the important public safety challenges is reducing traffic accidents. Given the sheer volume of the occurrence of such incidents, it seems imperative that the data collected be analyzed in an effort to introduce measures to avert future incidents.

In the year 2018, the Bureau of Transportation Statistics (BTS) reported 273.6 million registered vehicles in the United States[3]. With an estimated population of 327.2 million in the same year, around 85% of the population owned vehicles (passenger cars, motorcycles, trucks, buses and other vehicles) and this statistic has been increasing ever since[4]. A high vehicular density naturally implies a proportionally high number of traffic accidents - fatal or otherwise.

IIHS (**Insurance Institute for Highway Safety**) reported an alarming 33560 deaths from 33654 vehicle crashes in 2018. Although the overall per capita death rate decreased by 3% over the previous year (2017), the numbers are still high and occur due to a variety of reasons in like seat belt usage, speeding, phone usage, alcohol consumption, weather conditions etc[5].

Motor vehicle crash deaths per 100,000 people by type, 1975-2018													
Year		Passenger vehicle occupants		Pedestrians		Motorcyclists		Bicyclists		Large truck occupants		All motor vehicle deaths*	
	Population	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate	Number	Rate
2015	321,418,820	22,741	7.1	5,495	1.7	5,026	1.6	828	0.3	598	0.2	35,485	11.0
2016	323,405,935	23,957	7.4	6,080	1.9	5,337	1.7	848	0.3	662	0.2	37,806	11.7
2017	325,719,178	23,866	7.3	6,075	1.9	5,229	1.6	800	0.2	680	0.2	37,473	11.5
2018	327,167,434	22,891	7.0	6,283	1.9	4,985	1.5	854	0.3	678	0.2	36,560	11.2
*Total includes other and/or unknowns													

***Yearly Snapshot for motor vehicle deaths per 100000 people by type[2].*

A lot of research has been conducted into identifying factors contributing to the cause of traffic related accidents. The dataset[9] that we are working with covers 49 states of the United States, with 3 million records (2974335 rows) collected over a time period of 3-4 years. Specifically, the dates range from February 2016 to December 2019 and have been collected real-time through various entities such as law enforcement agencies, traffic cameras and sensors etc. This data is characterized by 49 attributes detailing each of the three million tuples. Each tuple represents a traffic accident involving at least one vehicular entity.

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
TMC	2246264.0	207.831632	20.329586	200.000000	201.000000	201.000000	201.000000	4.060000e+02
Severity	2974335.0	2.360190	0.541473	1.000000	2.000000	2.000000	3.000000	4.000000e+00
Start_Lat	2974335.0	36.493605	4.918849	24.555269	33.550402	35.849689	40.370260	4.900220e+01
Start_Lng	2974335.0	-95.426254	17.218806	-124.623833	-117.291985	-90.250832	-80.918915	-6.711317e+01
End_Lat	728071.0	37.580871	5.004757	24.570110	33.957554	37.903670	41.372630	4.907500e+01
End_Lng	728071.0	-99.976032	18.416647	-124.497829	-118.286610	-96.631690	-82.323850	-6.710924e+01
Distance(mi)	2974335.0	0.285565	1.548392	0.000000	0.000000	0.000000	0.010000	3.336300e+02
Number	1056730.0	5837.003544	15159.278074	0.000000	837.000000	2717.000000	7000.000000	9.999997e+06
Temperature(F)	2918272.0	62.351203	18.788549	-77.800000	50.000000	64.400000	76.000000	1.706000e+02
Wind_Chill(F)	1121712.0	51.326849	25.191271	-65.900000	32.000000	54.000000	73.000000	1.150000e+02
Humidity(%)	2915162.0	65.405416	22.556763	1.000000	49.000000	67.000000	84.000000	1.000000e+02
Pressure(in)	2926193.0	29.831895	0.721381	0.000000	29.820000	29.980000	30.110000	3.304000e+01
Visibility(mi)	2908644.0	9.150770	2.892114	0.000000	10.000000	10.000000	10.000000	1.400000e+02
Wind_Speed(mph)	2533495.0	8.298064	5.138546	0.000000	4.600000	7.000000	10.400000	8.228000e+02
Precipitation(in)	975977.0	0.020495	0.235770	0.000000	0.000000	0.000000	0.000000	2.500000e+01

*** Description of a partial subset of the main database*[9]

II. METHODOLOGY

The methodology implemented follows a traditional data analysis structure:

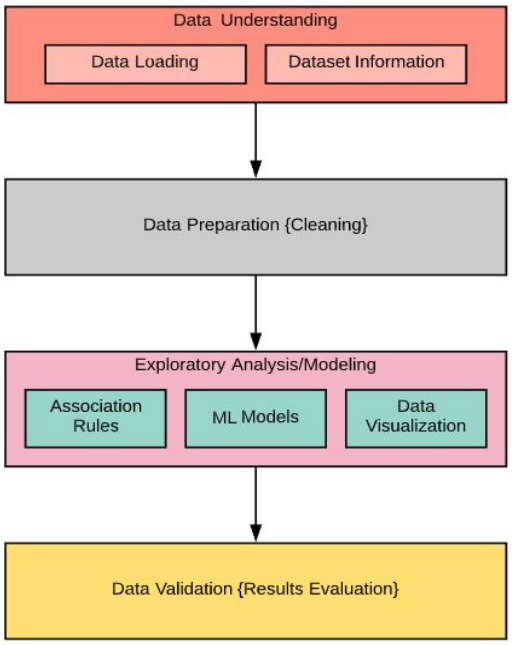

Figure 1. Workflow

Figure 1 establishes the process flow of the work conducted into the analysis. Our code is broken down into four main segments - each altering the dataset to provide detailed analysis of it before we reduce it to the required dimensions.

While the visualizations provide us a general understanding of the attributes of the dataset and how they relate with each other, the models/analysis run on the refined datasets give us insight on what factors contribute significantly to severity of accidents. These factors can then be weighed into the 4 E's - Enforcement, Engineering, Emergency Medical Services and Education. If high risk factors for each area are readily available, changes can be incorporated into local traffic regulations, construction of road infrastructure such as signage and road structure, well paved emergency routes and raising awareness in the general public.

Since our dataset contains close to 3 million records, we segment the data based on factors that are relevant to effectively ***create association rules*** and ***generate models.*** This is done to reduce performance time as generating models takes quite a chunk of time to execute. Our data is broken down by state, zip code, and county. We prefer using Massachusetts and have selected Middlesex county which includes locations like Somerville, Cambridge etc. The reason for starting off with Massachusetts is because we are all located here and find it best to analyze a

territory we are all familiar with. By doing this we are able to reduce our target dataset to about 30 thousand records, and even further when we segment by county, which results in a total of about 11 thousand records.

III. CODE

Our python notebook code is divided into 5 different sections.

1. **Dataset Info**

   This section of the notebook presents a basic idea about the columns we used in the dataset and sample data points. There are 3 million records (rows) with 49 attributes (columns) - most of which can be clubbed together to identify a common factor. For e.g. Start_Lng/Start_Lat/End_Lng/End_Lat, Street, Side, City, County, State, Zipcode and Country are all a representation of location. Similarly, windchill, humidity, pressure, visibility and other such attributes all relate to the weather conditions at the time of the accident. Some of the non-integer data such as TMC (Traffic Message Channel) provide the description of the accident which can be clubbed with the severity of the accident. A more detailed account of the dataset can be obtained here[0].

2. **Data Visualizations**

   We used mainly Seaborn and Matplotlib packages to perform visualization for this dataset. Seaborn's heatmap function was used to plot the rectangular data as a color encoded matrix. The features used were TMC, Severity, Distance, Temperature, Wind_Chill, Humidity , Pressure and Visibility. Matplotlib was extensively used to generate Bar Graph. This library gave us a lot of out of the box API's which were very simple to use. More specifically the Matplotlib's pyplot API which has convenience functions called subplots which was the utility wrapper that helped us in creating common layouts of subplots in a single call.

   We also used the Figure class which is part of Matplot.figure module which acted as a top level container for all plot elements. To generate a WordCloud, we processed the feature to gather all text as a single 'String'. That processing was performed inside the processText method.

3. **Association Rules and Analysis**

   Before running the association rules we did some preprocessing of the data. We wanted to focus on the point of interest (all of the boolean columns), severity, and if it was day or night time, so we dropped everything else that did not match this criteria. Next, we had to

transform the boolean columns to a usable format that can be interpreted since the association rules works with string representations. Specifically we transformed the rows that had true values to the string representation of the column, e.g. column: ***Cross_Walk, row: true*** would become ***Cross_Walk, row: +Cross_Walk***, the plus represents it was present. Anything that was false was omitted before running the association rules because the false values were dense and we would not be able to derive anything meaningful since the results would be noisy. Next we chose a support value of 2% and a confidence value of 50%. We chose 2% percent as a support value after some trials at 50%, 25%, 10% and 5% because it allowed for us to see more granular correlations. We chose a confidence value of 50% because we wanted to focus on strong correlations.

4. Machine Learning Models

Before running the models, we cleaned the data by removing the outliers. The first part of this process was to drop the rows with negative time duration. The next step was to find out the outliers within the time_duration column which deviate significantly from the mean. We filled such outliers with median values.

Our aim was to predict the Severity of the accident based on all the other input parameters. For the input feature list, we excluded the boolean type of columns because - they were discrete and our dataset had a lot of null values for the boolean types of columns. So we selected 33 total features out of the original 49. The following are the attributes we used for predicting the models.

TMC	County	Wind_Direction	No_Exit	Traffic_Signal
Severity	State	Weather_Condition	Railway	Turning_Loop
Start_Lng	Timezone	Amenity	Roundabout	Sunrise_Sunset
Start_Lat	Temperature(F)	Bump	Station	Hour
Distance(mi)	Humidity(%)	Crossing	Stop	Weekday
Side	Pressure(in)	Give_Way	Traffic_Calming	Time_Duration(min)
City	Visibility(mi)	Junction		

In order to run the models, we selected two Counties in the United States - Dallas in Texas and Middlesex in Massachusetts. The reason we selected Dallas was because it

ranked among the top 5 most accident prone counties in the US and we wanted to compare it with Middlesex, where most of us are living right now. We randomly split the data into training and test datasets with 80% going into the training set and 20% for testing.

We generated models based upon 5 different classification algorithms and fed our data to get the accuracy score.

a. **Logistic Regression:** We used the sklearn Logistic Regression classifier with liblinear solver with the maximum iterations set to the default of 100.
b. **K-Nearest Neighbors:** For this section, we used sklearn KNeighborsClassifier with the number of neighbors at 6.
c. **Decision Tree:** We used sklearn Decision Tree classifier using both gini and entropy criterion to measure the quality of a split. We tuned the maximum depth of the tree as 8 and measured with both the information criterion on both the counties.
d. **Random Forest:** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.
e. **Gaussian Naive Bayes:** The Gaussian Naive Bayes assumes that all the features are following the Gaussian distribution or Normal Distribution. We have used the sklearn GaussianNB module which implements the Gaussian Naive Bayes algorithm for classification.
f. **Neural Networks:** We defined a model which would predict the accuracy of the accident based upon the description of the accident. For that, we have initially tokenized the words in the 'Description' column and considered only those descriptions which have a frequency of at least 100. Keras Text tokenization utility class was used to vectorize the text corpus.

   We trained the model using categorical_crossentropy as the loss function on softmax activation with each batch_size of 64 for 15 epochs.

IV. RESULTS

The following are results from the analysis conducted:

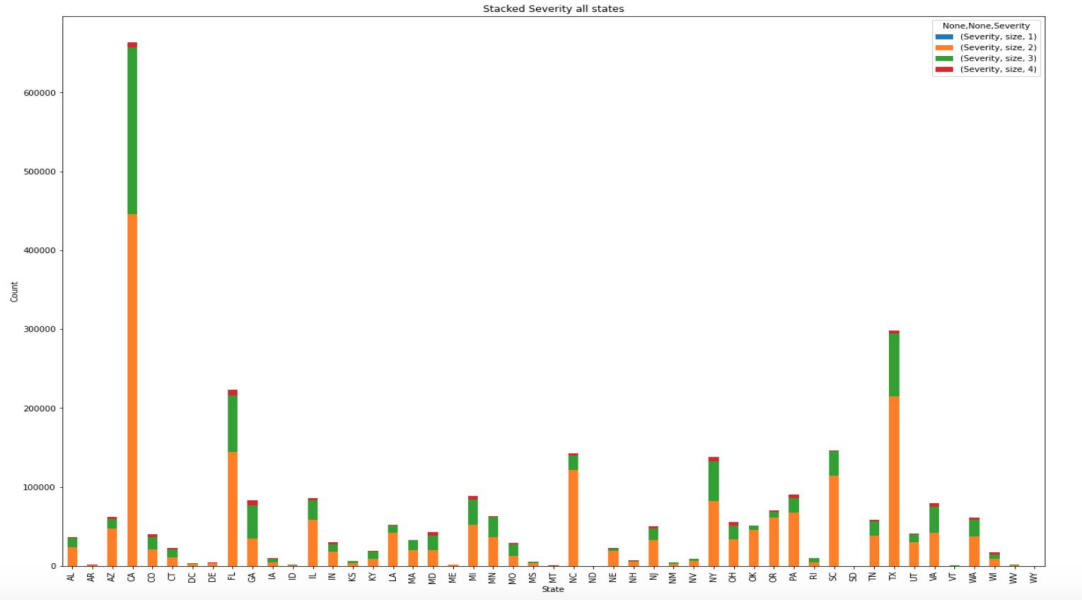

Figure 2: Number of accidents by state with severity

Stacked Bar Graphs:

Figure 2 shows top states with the most accidents - California, Texas and Florida respectively. States with the least number of accidents are Wyoming, North Dakota, and South Dakota.

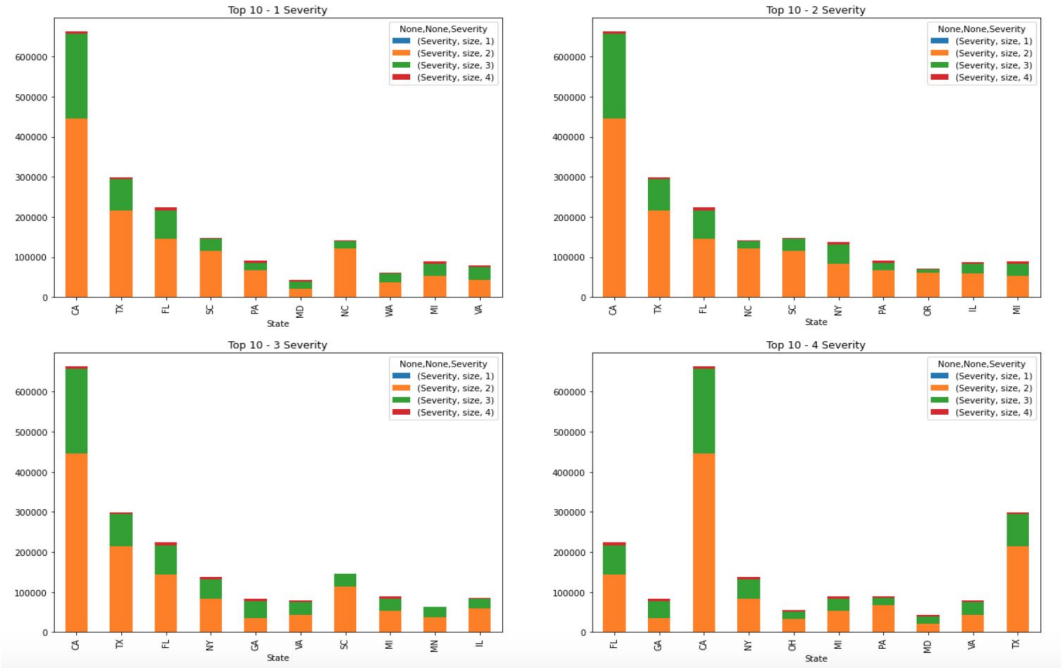

Figure 3: Top 10 states ranked by severity

Breaking down the data in Figure 3, we list the top 10 states for each severity category. A severity of 1 represents a short delay and a severity of 4 represents a significant delay. Looking at the figure above you can see the top 10 states for each severity category. California is ranked number one in accidents with severities 1, 2 and 3. It is interesting to note that Georgia has significantly less accidents but has the most accidents with a 4 severity and Florida has the most in this category. What is also interesting to note is that Texas is ranked top 3 in each severity 1, 2, and 3, but they are last in severity 4. Florida is ranked top 3 in each severity category.

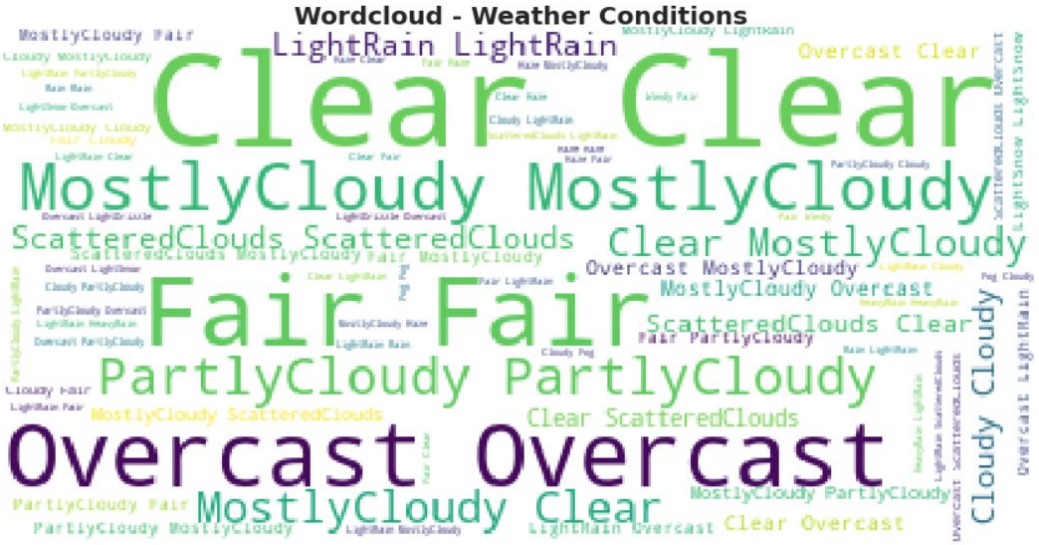

Figure 4: Weather conditions word cloud

Word Cloud:

The word cloud in Figure 4 is generated on the basis of weather conditions. We observe that most accidents happen when the weather conditions are - Clear, Fair and Overcast. This seems correct because there is always a tendency to overspeed when the weather conditions are clear and roads become slippery when it is overcast. This also indicates that drivers tend to have a rather cautious approach to driving when the weather conditions are not suitable.

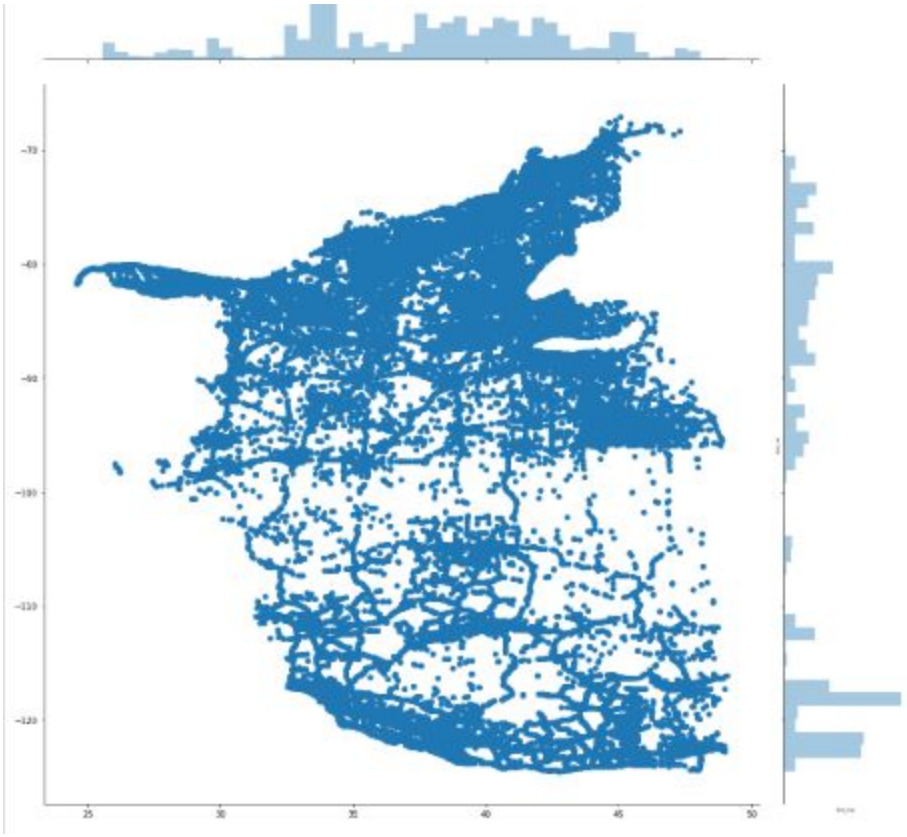

Figure 5: Joint Plot

Joint Plot:

From the visualizations performed above we inferred that California has had the maximum number of accidents. We were curious to confirm this by performing EDA on the dataset. Seaborn's joint plot was perfect fit for our analysis and we created joint plots for latitudes and longitudes recorded for the start and end of the accidents. It is observed for the joint plots most accidents happen with latitudes 35 - 45.

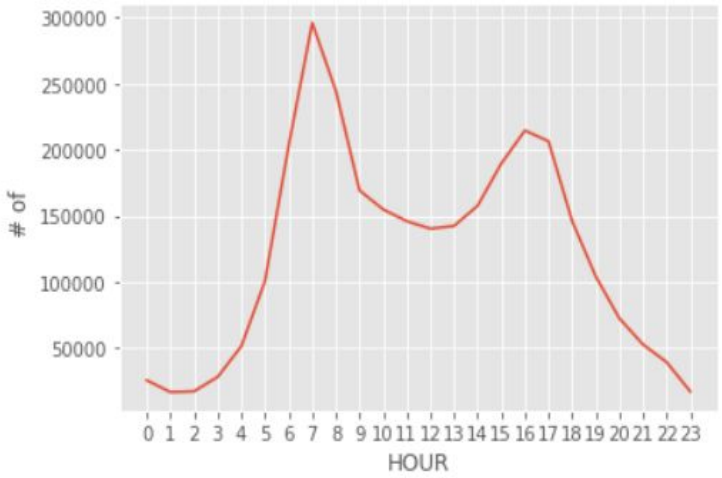

Figure 6: Number of accidents vs time of day

Time plot:

It's very easy to guess that morning and late afternoon are peak travel hours where people commute to workplaces, schools and colleges. To visualize this we plotted a graph between the hour and number of accidents (Fig. 6). It’s definitely not surprising to see that two peaks are at the 7 am and 4 - 5 pm slots.

Vertical Bar Plots:

What time of the year did most accidents happen ? How did it impact the traffic?

The busiest travel period in the USA is towards the end of the year. There is a good probability that most accidents happen during this period.We observed from the plot of number of accidents vs (week and year) .

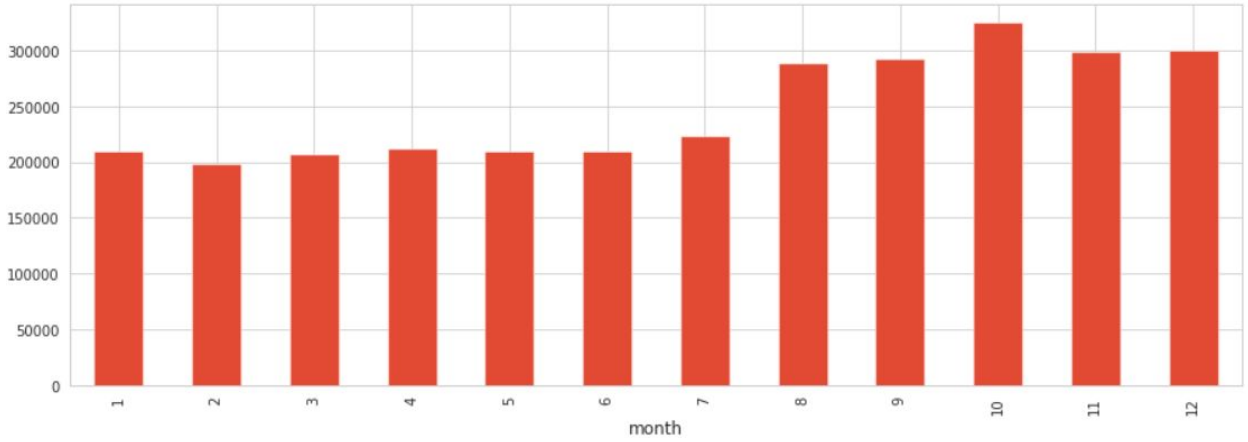

Figure 7: Plot of months versus accidents

We see that from the past three years most accidents have happened during Week 48-52 which is around Thanksgiving week. Also we found out that during weeks 48-52, the accidents had higher impact on the traffic and resulted in longer delays.

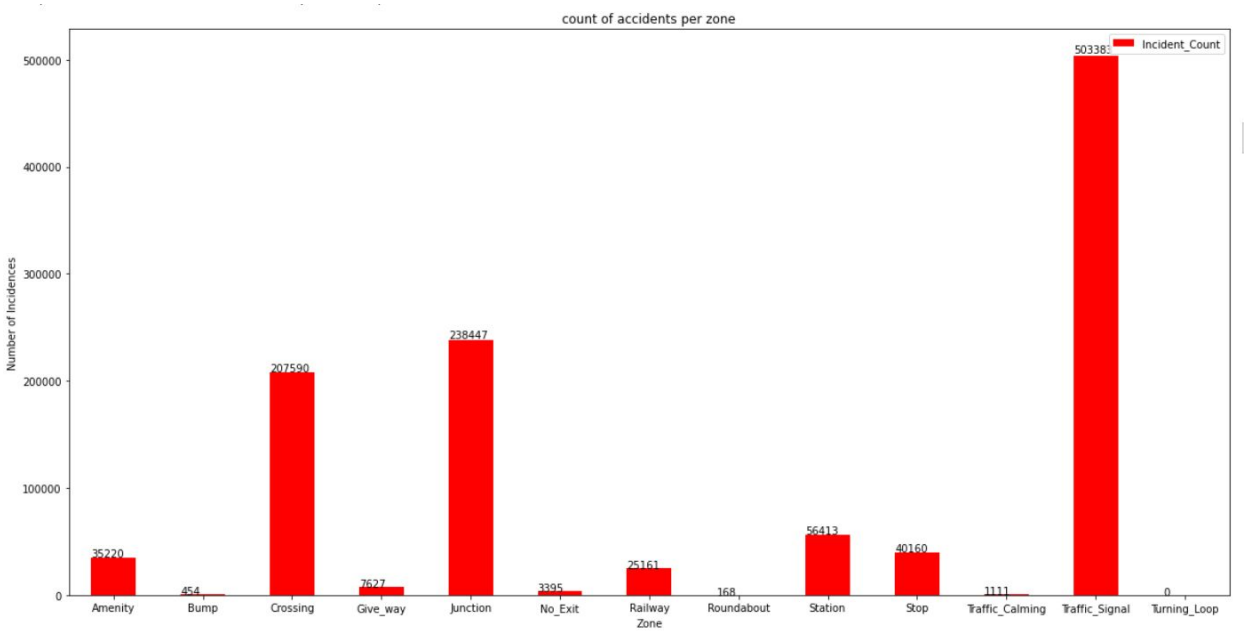

Figure 8: Statistics for accidents based on Zone

In Figure 8, reducing the dataset to zones of accident and then summing the boolean data for each zone shows that the three highest accident riddled zones are traffic signal (503383), junctions (238447) and crossing (207590). The lowest three zones of accident are turning loops (0), roundabouts (168) and bumps (454).

X	Severity	Confidence	Support
+Crossing, Day	2	0.85	0.041
Scattered Clouds, Day	2	0.595	0.048
+Crossing, +Traffic_Signal	2	0.864	0.025
Day, Light Rain	2	0.62	0.036
Cloudy	2	0.59	0.021
+Crossing, Day, +Traffic_Signal	2	0.865	0.021
+Junction, Night	3	0.653	0.021
+Junction	3	0.622	0.089
+Junction, Clear	3	0.596	0.024
+Junction, Day	3	0.613	0.069

Table 1 : Association Rules Middlesex County Massachsusetts

Table 1 describes an overview of some association rules obtained from Middlesex County in Massachusetts. These were obtained using a support threshold 2% of and confidence threshold of 50%.

Results from Models:

a. **Logistic Regression:** In this model, we observed an accuracy score of 76.8% for Dallas County dataset and 83.8% for Middlesex County. Although the number of data samples are greater for the former one, we could see that the latter is more accurate by around 7%. The reason for this could be that the distribution of the data in the feature space was different for both of them.

b. **K-Nearest Neighbors:** We observed that for the same number of neighbors, the accuracy score on Dallas County data was 68.3% compared to that of Middlesex which was 58.3%. As we see KNN is a non-parametric model, whereas LR is a parametric model, Logistic Regression outperforms K-Nearest Neighbors in accuracy.

   For optimizing the number of neighbors while using the KNN model, we plotted the accuracy vs number of neighbors with the range of neighbors from 1 to 9. The 'knee' for the ideal number of neighbors for both the Counties is 2.

c. Decision Tree: It was observed that the Entropy score of accuracy on Dallas county was 79.9% and using the gini criterion, it was 79.5%. On the other hand, for Middlesex county, we observed the entropy score of accuracy as 78.1% and gini accuracy score as 78.2%. The decision tree classifier is performed similarly on both the datasets.

d. **Random Forest:** We observed an accuracy score of 85.7% using the Random Forest Classifier on the Dallas dataset using the number of trees in the forest as 100. After running this model, we viewed the feature importances which give a sense of which of the variables have the most effect in these models. We found the top 20 important features and selected those which have an importance factor greater than 0.03. By running the Random forest classifier again only based on these important features, the accuracy improved by 4% - with a score of 90%.

   Running a similar analysis on the Middlesex dataset, we got an accuracy score of 89.4% on the full featured data and by selecting only important features, the accuracy improved to 89.8%.

   By comparing the important features on both the counties, 9 out of 10 features were commonly impacting the accident. Those features are:

   1. Start_Lng
   2. Start_Lat

3. Temperature(F)
4. Humidity(%)
5. Pressure(in)
6. Traffic_Signal
7. Hour
8. Time_Duration(min)
9. Side_R

e. **Gaussian Naive Bayes:** For Dallas county, we observe an accuracy score of 62% compared to Middlesex county which has 55.6%. Since Gaussian Naive Bayes assumes that the data is distributed normally, we could observe that Dallas county has a better distribution than Middlesex county due to greater data samples by a factor greater than 3 times. This takes to the basic principle that as the data samples increase, the it would tend towards the Gaussian distribution.

f. **Neural Networks:** We were able to obtain a model which predicts the severity of the accident based upon just the description obtained from the reports with an accuracy of around 90%.

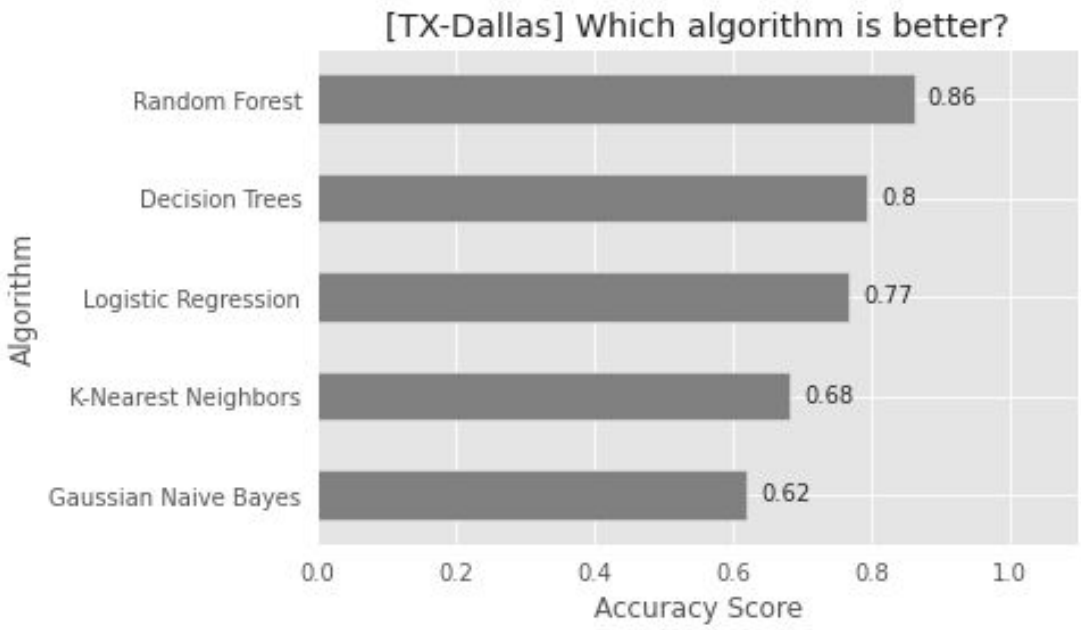

Figure 9: ML model accuracy score comparison for Dallas, Texas

We can see that random forest performed the best for Dallas, Texas at 86% and Naive Bayes the worst at around 62%. The other models also did fairly well. Random forest, being an extension of decision trees, does better than it - which can be observed here.

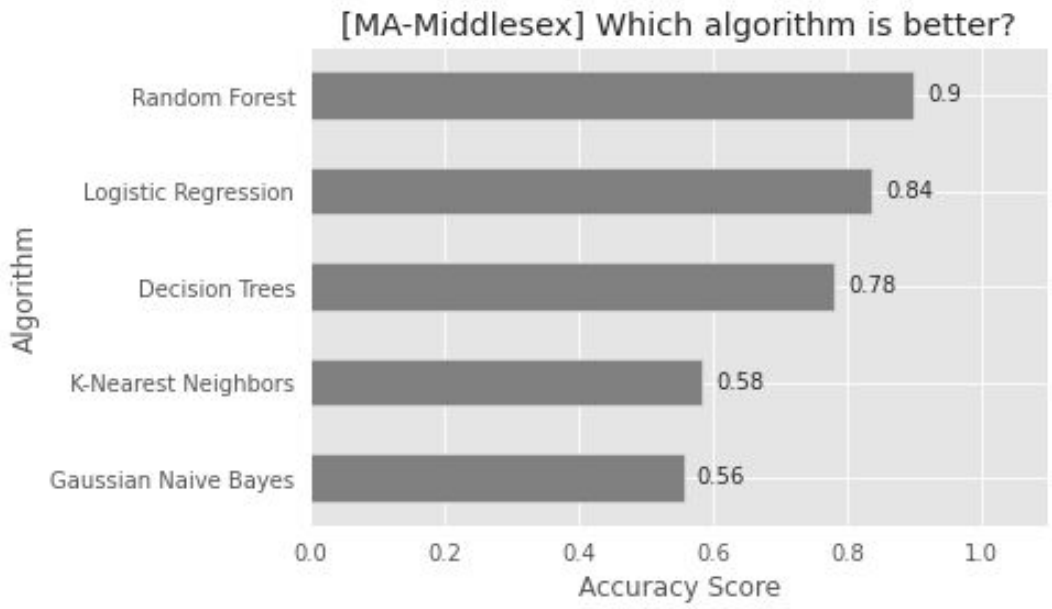

Figure 9: ML model accuracy score comparison for Middlesex, Massachusetts

We can see that random forest performed the best for Middlesex, Massachusetts as well at 89% and Naive Bayes the worst at around 56%. The models did fairly well apart from KNN and Naive Bayes.

V. DISCUSSION

From the bar plot for zone based accidents (Fig. 8), it feels instinctual to think that most of the traffic accidents happen either at traffic signals because of the multi-directional nature of the zone or at intersections. Turns out, this estimate is quite correct as around 503383 accidents occurred at traffic signals and another 238447 accidents occurred at junctions (intersections). The third largest zone of accidents is at crossings. Since accidents at traffic signals, intersections and crossings all involve some form of speeding and the entities involved are often moving in different directions (passengers and many vehicles cross perpendicular to oncoming vehicles at crossings), these zones seem to be a hotspot for traffic accidents.

Next we discuss the results for our association analysis. The main focus of this analysis was to determine if there were correlations between severity and point of interest of the accident. Based on the results in **table 1**, we can see there is a correlation between the severity of the accident and where the accidents occured. For example we can see when there is a junction, the severity of accidents tends to be higher, i.e. 3, than it is when a junction is not present, i.e. 2. In addition to this we can see if there is a traffic signal present the severity tends to be lower. These results make sense because traffic lights are designed to control the flow of traffic and with the presence of them it makes sense that traffic is not significantly hampered, whereas at junctions it is more difficult to control the flow of traffic. What is interesting about this is that there is a very

high confidence between traffic signals and cross walks and the severity, which correlates to a severity of 2. This potentially means that individuals are either being more reckless at junctions or these areas are more difficult to navigate and not as straightforward as areas with traffic signals / cross walks.

Another important feature we decided to ponder upon is weather conditions. On an average weather conditions result in 20 percent of accidents. We first plotted a word cloud to see the weather conditions that resulted in most number of accidents. From the word cloud we could infer that most number of accidents have taken place when the sky is clear. Total of 808171 accidents happened when the weather was clear. And least number of accidents, around 34314 accidents happened when the weather was hazy. One inference we could think that there is a general tendency to overspeed during the clear weather conditions . And people reduce the speed and are extra cautious during wet and hazy conditions.

Comparing the top features which are causes of accidents across both the counties, we observe that the geographic location in terms of latitude and longitude plays a major role in the cause of the accidents. This might be because of the lack of stringent state road laws, most of the accidents are happening in the states of California and Texas. The next important factors which are causing the accidents - temperature, humidity and pressure are all weather related. Because the weather conditions in both these states are not too extreme compared to the other states, people tend to overspeed which is the main reason for the accidents. By imposing strict laws in such states, most of the accidents could be prevented.

VI. FUTURE WORK

This study is limited by the computing power available. Running models on a cumulative 3 million records is time consuming and does not bode well on standard systems. A future alternative would be to host data on AWS and use Apache Spark to cluster data. Traffic accidents are one of the major public safety issues prevailing in every country right now. While our investigation is limited to two counties, the study can potentially be extended to other counties in multiple states. We also want to apply Deep learning techniques to this dataset.

While this study provides a concrete analysis of the recorded traffic accidents, the application of it would be in moulding the 4 E’s (Enforcement, Education, Engineering, EMS) around the observations inferred. With these inferences, we can form new traffic regulations and related policies, create road infrastructure shaped to avoid accidents and inculcate awareness of road rules in the general public - all with the intention of reducing risk to human and other (animal) lives. The additional insight from research like this can assist agencies in creating data driven policies.

VII. CONCLUSION

Throughout this study, the dataset has been viewed as a pool of factors that could potentially clue us in to what are the most meaningful attributes with the most impact on reduction of vehicle accidents . Since our attributes are not of the same type, their importance is weighed via associations (for the bool type data) and through learning models for other types (float/int). Additionally, running neural nets provides us with the severity of accidents based on their description.

From all these analyses, we observe that there are about 10 attributes that primarily impact the severity of accidents and that most of these accidents occur in the most ideal conditions - clear weather and daytime hours. The chances of a motor accident occurring at a traffic junction in southern states is much higher given that these ideal weather conditions are present for a longer duration through the day and year. The results of the study allow us to narrow down the field of focus in traffic accidents to these southern states. Introducing different traffic regulations with more creative road infrastructure designs can reduce these accidents. One proposal would be to elongate the islands at 4 way crossings by a small measure that would enforce cars to slow down before turning. Additionally, speed breakers can be positioned before pedestrian crossings to slow down vehicle speeds. Creating routes to incorporate Emergency Medical Services and contributing to raising road safety awareness could also be targeted in a very specific manner by analysing the results provided in this study.

Although many factors such as mobile phone usage while driving, drug and alcohol consumption, and police officer versus driver's perspective of the incident are under-reported, many recommendations can be drawn based on the results of the overall analysis of data and the county-versus-county comparison.

REFERENCES

[0] "US-Accidents: A Countrywide Traffic Accident Dataset." *Sobhan Moosavi*, 31 Dec. 2019, https://smoosavi.org/datasets/us_accidents.

[1] "Fatality Facts 2018: State by State." *IIHS*, Available: www.iihs.org/topics/fatality-statistics/detail/state-by-state.

[2] "Fatality Facts 2018: Yearly snapshot," *IIHS*. [Online]. Available: https://www.iihs.org/topics/fatality-statistics/detail/yearly-snapshot.

[3] Statista Research Department, "Number of cars in U.S.," *Statista*, 12-Mar-2020. [Online]. Available: https://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990/.

[4] U.S. Department of State Bureau of International Information Programs, "Does Everyone in America Own a Car?", Available: https://photos.state.gov/libraries/cambodia/30486/Publications/everyone_in_america_own_a_car.pdf.

[5] "Find Out What Are The Leading Causes For Most Car Accidents | Markets Insider," *Business Insider*. [Online]. Available: https://markets.businessinsider.com/news/stocks/find-out-what-are-the-leading-causes-for-most-car-accidents-1027930348.

[6] "Fatality Facts 2018: State by State." *IIHS*, https://www.iihs.org/topics/fatality-statistics/detail/state-by-state/

[7] Statista Research Department. "Number of Cars in U.S." *Statista*, 12 Mar. 2020, https://www.statista.com/statistics/183505/number-of-vehicles-in-the-united-states-since-1990/.

[8] "3.2.4.3.1. Sklearn.ensemble.RandomForestClassifier¶." *Scikit*, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[9] Moosavi, et al. "A Countrywide Traffic Accident Dataset." *ArXiv.org*, 12 June 2019, https://arxiv.org/abs/1906.05409.

[10] US Department of Transportation https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm