Analysis on the Impact of Weather and Road Conditions on traffic accidents in USA

Rahul Kata
PES1201802018
katarahul@gmail.com

Pavan A
PES1201800157
pavanappanna7@gmail.com

Harsha Kulkarni
PES1201802000
kulkarniharsha14@gmail.com

Abstract — Road accidents have become very common nowadays. Many people don't follow traffic rules but that isn't the only reason, effects of weather and climatic changes are also significant. So, for safe transportation, careful analysis of road traffic data is very critical to discover the features that affect the traffic accidents, the most. Different models such as logistic regression, random forest, are used to best fit the data, so as to predict as accurately as possible. classification model was built by logistic regression and clusters are formed by K-means clustering algorithm and the non-linearities are handled by random forest by exploiting the correlation between the features of data points. The first and foremost job in data analysis is data pre-processing. Handling the missing values, normalization, removing the curse of dimensionality is done prior to analysis. Suggestions for safety driving based on the convicted features and deducted rules, are made.

I. Introduction

With the exponential increase of vehicles every year, the fatal accidents also increased considerably. Climatic conditions such as temperature, humidity, pressure, wind speed etc, have considerable effects on those accidents. Appropriate safety measures and valuable information can be deduced by deep analysis of the data, thereby providing safety suggestions.

We have used many techniques and algorithms to find patterns and relationships among different features taken in the dataset. Intensity of role of weather in accidents, prevention measures of accidents, severity of accidents at different times of the day, locations of frequent accidents, all these are predicted by using appropriate models. Firstly, in pre-processing handling of missing data is done by both mean imputation and median imputation, outlier analysis is done by using box plot. Next comes the handling of categorical data which is done through binary encoding, then to reduce the dimension we have used Principal component analysis and lastly to get the relationship among the features we calculated correlation coefficient for each pair of features we got after PCA.

II. Literature Survey

Farhan[1] uses supervised machine learning techniques to classify the severity of accidents that take place in Bangladesh. The paper also uses different visualizations to interpret the impact of a given feature on the accidents being caused. The authors have collected a dataset from the ARI of BUET that consists of a total 43,089 traffic accidents [2001-2015] in Bangladesh. The dataset has 34 features with 8.7% missing values in total. The paper uses three feature selection algorithms to reduce the number of features: Univariate Feature Selection, Recursive Feature Elimination, and Feature Importance. Mean imputation is used to deal with 1.65% missing values in the selected features. Decision Tree, K-Nearest Neighbours (KNN), Naïve Bayes and AdaBoost are applied to classify the severity of the accidents. The target column is found to be imbalanced, hence Feature Creation is used to balance the target column. Model performance is found to improve slightly when the target column is balanced. Adaboost outperforms the other three algorithms as it is a combination of multiple weak classifiers. Accidents are relatively unforeseen and spontaneous, so direct observation is quite difficult. For that reason, getting 100% accurate data is quite impossible. The paper claims that training time and the risk of overfitting can be reduced with a suitable feature selection algorithm. Missing values also affect the performance of the model, thus an appropriate data imputation method should be used.

The key takeaways from this paper[1] is that data pre-processing techniques like data cleaning and feature selection are very integral before passing the data to a machine learning model. Imbalanced data can lead to a biased/skewed model.

FARS (Fatal Accidents Reporting System)[2] dataset, containing 37,248 records and 55 attributes. Since our paper is about effects of weather on road accidents, this paper is most relevant as it explains all features affecting the accident. Careful analysis of road traffic data is critical to find out the variables that are closely related to fatal accidents. Association rules were discovered by Apriori algorithm, classification model was built by Naïve Bayes classifier, and clusters were formed by simple K-means clustering algorithm with Euclidean distance as dissimilarity measure.

Data cleaning is important so all the tuples with missing values are removed and numerical values are converted into nominal values in accordance with the requirement.

Liling[2] claims that if more data, like non-fatal accident data, weather data, mileage data, and so on, are available, more tests could be performed thus more suggestions could be made from the data. The takeaway from this paper is that selection of appropriate features is crucial as they will help in more accurate analysis of data.

Zhang [3] represents the categorical variables numerically using one-hot encoding. Then the authors use Logistic Regression with and without Recursive Feature Elimination (RFE) to perform feature selection and identify whether a certain set of conditions would increase the likelihood of an injury-inducing accident. The dataset for this paper was a randomly selected sample of 46,077 crash entries for the last five years, retrieved from MassDOT's crash portal, based in Massachusetts.

They[3] assumed that crash data is better interpreted by linear modelling. They removed columns such as co-ordinates which had too many distinct possibilities to group together. The paper claims that prediction accuracy can be increased by using suitable feature selection algorithms. They also suggested using SVM's in the future to better model the non-linearity of certain attributes contributing to the accident. According to the authors features like tough weather and road conditions, senior/teen drivers were the major accident-causing factors. The key takeaways from this paper is that data pre-processing techniques like data cleaning and feature selection are very integral before passing the data to a machine learning model. By encoding the categorical values, the machine learning algorithms can better understand data. The attributes of the accident dataset might not always be linearly related and imbalanced data can lead to a biased/skewed model.

III. METHODOLOGY

The countrywide car accident dataset collected from February 2016 to June 2020, which covers 49 states of the USA. The road and weather conditions during the time of the accident and also the location of the accident are recorded. The severity of the accident is also recorded on a scale of 1 to 4.

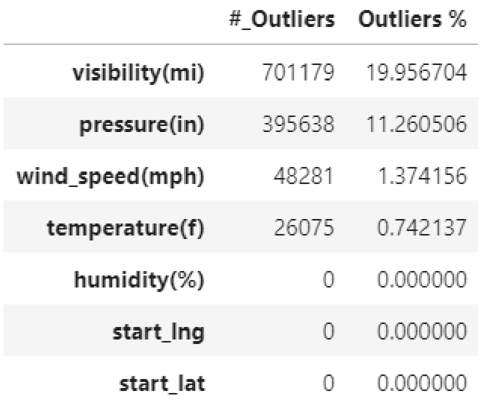
- *Shape* = 3513617 x 49
- Total unique records(*rows*) = 3513616
- Total Attributes(*columns*) = 49

Target column is *Severity*, which is of *ordinal data* type.

23 columns out of 49 columns have missing values in them, as shown below (which is a part of the code). The importance of the columns with missing values are gauged with respect to the goals we have put forth.

Mean Imputation is done for temperature(f)*,* pressure(in) and humidity (%) because of low standard deviation and a smaller number of outliers. *Median imputation* is done for *visibility(mi)* due to a large number of outliers.

There are only 7 numerical type columns after dropping the irrelevant columns. Visually determining the outliers using a boxplot is impossible due to the sheer volume of data. Hence, we have used *(Q3 + 1.5IQR)* and *(Q1 - 1.5IQR)* as upper and lower bound to filter out the values that lie out of that range.



	#_Outliers	Outliers %
visibility(mi)	701179	19.956704
pressure(in)	395638	11.260506
wind_speed(mph)	48281	1.374156
temperature(f)	26075	0.742137
humidity(%)	0	0.000000
start_lng	0	0.000000
start_lat	0	0.000000

It is observed that the number is very high and hence no immediate action is taken to eliminate the outlier.

Our initial insights are shown in following figures

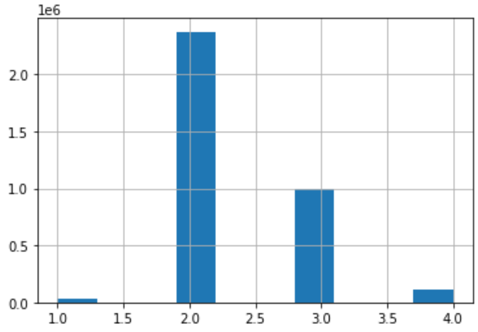


Fig: 1 [severity]

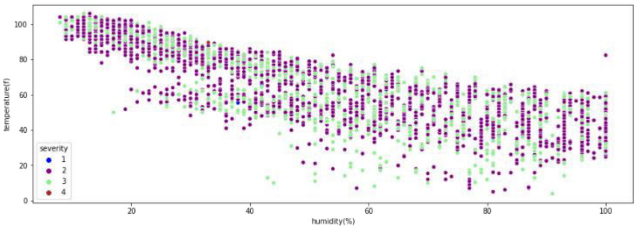


Fig: 2 [temp – humidity]

Fig: 1 shows that there are more data points with a severity rating of 2 when compared to other ratings. Thus this class imbalance has to be dealt with in the future while building the model by resampling the dataset. Fig: 2 shows the scatter plot of *temperature* vs *humidity* with severity index. The two are negatively correlated to each other and show no formation of clusters as such. Further sections such as *Correlation* and *PCA* provide further insights on correlation type and cluster absence.

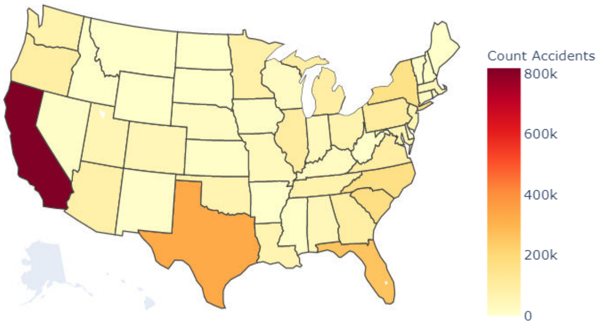


Fig: 3

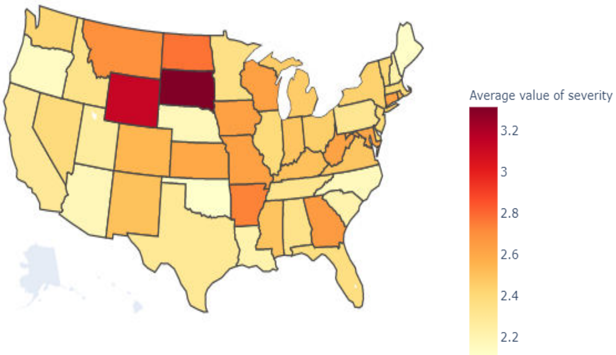
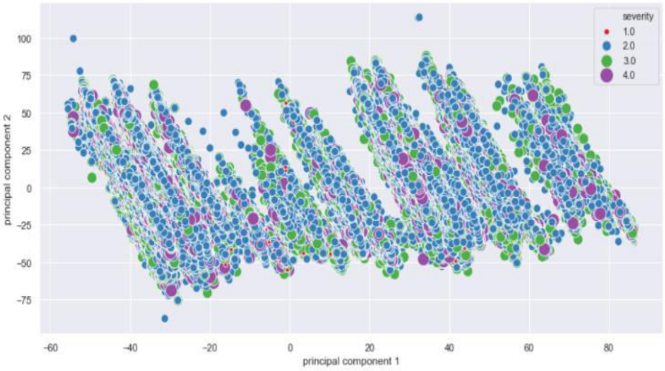


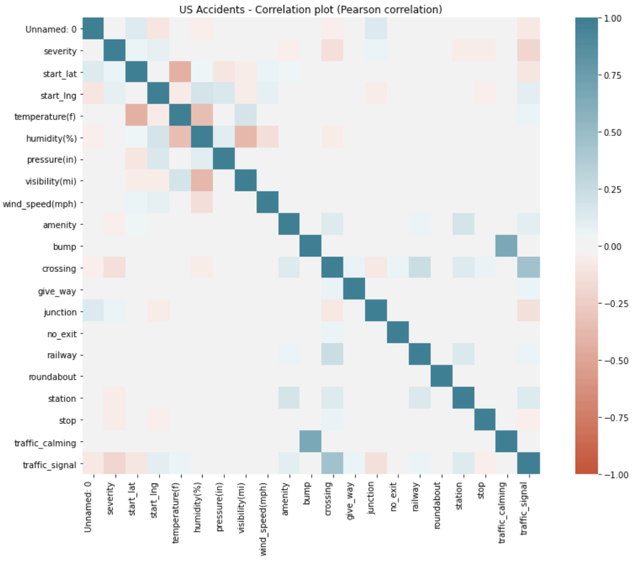
Fig: 4

From the above two plots it is evident that although the number of accidents is more in the extreme west (California) and extreme south (Texas), the severity of accidents in the northern region (Wyoming, Dakota, Montana) is higher.

We have applied Principal Component Analysis using the sklearn's inbuilt PCA() function. we noticed that only the first 6 principal components are enough to explain 99.9% variance of the features shown above. For visualization purposes we have selected the first 2 principal components which explain 84.7% variance. When these 2 components are plotted using a 2-D scatter plot, there are no differentiating clusters for each severity level.



A Pearson correlation has been performed to understand the linear relationship among different features. From the correlation it can be observed that *traffic_signal* and *crossing* have a positive correlation with each other. *Traffic_calming* and *bump* too have a significantly higher positive correlation with each other. *Temperature* and *Humidity* show a negative correlation between them. However, though *visibility* shows negative correlation with *humidity*, due to the volume of data, scatter plot failed to reveal a significant negative slope. The correlations possibly indicate important features that influence the occurrence of an accident, in general.



IV. REFERENCES

[1] Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine. *Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh.* 7th International Conference on Smart Computing & Communications (ICSCC), 2019

[2] Liling Li, Sharad Shrestha, Gongzhu Hu. *IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 363-370. IEEE, 2017.*

[3] Zhang, Aaron, Evan W. Patton, Justin M. Swaney, and Tingying Helen Zeng. *A Statistical Analysis of Recent Traffic Crashes in Massachusetts*