A comprehensive study of US traffic accident evaluation and prediction

Team #145

Introduction

According to the World Health Organization, traffic accidents result in 1.35 million deaths and 20 to 50 million sustained non-fatal injuries annually. Solely in the US, traffic accidents incur a cost of $871 billion annually, just over 4% of the total GDP, and is one of the leading causes of death for those between 1-54 years of age.[19] Therefore, traffic accident evaluation, prediction, and prevention are crucial topics to be explored. *(HQ4)* By providing a more accurate method of risk prediction and granting the accessibility of an interactive risk prediction platform to the public, many groups, including drivers, first responders, and transportation departments can take on precautionary measures and actions to reduce the risk. The success of our algorithm will be measured by prediction accuracy and precision, while the interactive platform will be measured by user studies, and it is our goal to observe traffic accident reduction in our serviced areas in the long run.

Problem Definition

The objectives of this project are to evaluate US car accidents based on various environmental and geographical conditions and develop an interactive platform of car accident risk and severity prediction available to the public. Our specific tasks include data manipulation and evaluation for data training, development of a highly accurate risk and severity prediction algorithm using machine learning techniques, and development of an interactive web-accessible platform for public access.

Literature Survey

Traditionally, accident prediction models have been designed based on a causal relationship between traffic accidents and a variety of human, road geometry, and environmental factors. When each subset was examined individually, it has been found that the frequently studied attribute of weather conditions and road geometry is correlated with accident occurrences.[1] More specifically, the effect of precipitation has been found to be consistent in its ability to increase accident frequency[2], especially in urban areas[3]. However, these traditional statistical methods are built on static assumptions and are limited by their tendency to follow patterns.[4] Therefore, when tasked to predict car accidents, basic causal statical predictions are merely adequate.

Since traffic accident data is of great heterogeneity, traffic accident studies are usually limited by the scale of their data. In most cases, analysis has only been able to be performed on a localized city and town level.[1,3] In this project, we built our model based on a US country-wide traffic accident dataset[5,6] with 47 available attributes.

More recently, researchers have begun to shift towards using machine learning to tackle traffic accident risk evaluation and prediction and improve model accuracy[7] , such as Decision trees, and random forests. Specifically, the K-means clustering has the ability to overcome the heterogeneous nature of the traffic accident prediction thus improving the performance.[8,9,10] When paired with random forest, the resulting model was able to achieve a predictive accuracy of 99.86% for road accidents in Dehradun, India.[11] More complex deep learning and neural networks algorithms tend to have higher accuracy than machine learning algorithms, however, they are limited by their inability to capture intervariable dependencies – and for such a heterogeneous nature of traffic accident data, this limitation poses a major drawback.[12,13,14,15,16]

Proposed Method

Our approach to traffic accident occurrence and severity prediction will employ the US country-wide traffic accident dataset developed by Sobhan et. al., the largest and most comprehensive to date[5], and be modeled using several machine learning algorithms such as classification, decision tree, random forest, and regression. We propose that the predictive power of our model will surpass those of previous studies based on the superior quality of our training data. Furthermore, despite having several developed algorithms, these methods currently only exist on paper. Our team aims to develop a publicly accessible interactive web-accessible platform using python and JavaScript to allow users to input various attributes and conditions to forecast traffic accident risk and severity of their future travels. Aside from building an innovative platform and a more powerful algorithm, we also propose to examine the relationship between the US census population and population density data and the US traffic accident data, which has only been researched at the regional level before.[17,18]

Algorithm

The machine learning algorithms used in this project include classification, decision tree, random forest, and regression. To model the probability of traffic accident occurrence, based on the nature of our output, we propose to use a random forest model trained on merged weather data and the comprehensive US traffic data. Doing so, given the extent of the dataset, we expect our model to be valid with high accuracy and predictive power. To model the severity of occurred accidents, we seek to employ a combination of classification methods and regressions to find the model with high accuracy and high predictive power. It is expected that the algorithm will be built using python Flask.

User Interface

Our team proposes to build a web-based user interface using python Flask and JavaScript, specifically the D3 library. The ideal interface would allow users to input data into specific attributes used to build the algorithm and return the daily traffic accident occurrence prediction at the city level. Furthermore, the proposed interface would return the predicted severity for the associated predicted traffic accident. Additional visualization such as heatmap and choropleths are also subject for implementation.

Experiments and Evaluation

All experimentation and evaluation of this project is done using python and JavaScript. Our experiments are designed to answer the following questions:

1. What is the probability of traffic accident occurrence given specific user inputs?
2. What is the severity level of the traffic accident should it occur?
3. Is there a relationship between population density and traffic accident occurrence?

Data Cleaning

The dataset was first cleaned by accessing each attribute for missing values. The missing percentage of each attribute is summarized in Figure 1 below.

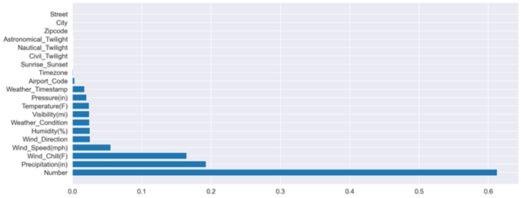


Figure 1. Attributes ranked by percent missing.

After all rows containing missing values were removed, we evaluated the numerical attributes in the dataset for any outliers. Figure 2 below shows the box-and-whisker plot for all numerical attributes in the dataset. All rows with outliers were removed. After examining the statistical outliers, we evaluated the remaining dataset for any intuitive outliers. For example, we removed the pressure data that are not within the normal range for atmospheric pressure of around 25 to 32 inHg. Similar processing was conducted for wind speed, precipitation for wind speed greater than 50 mph, precipitation greater than 2.5 in, and distance greater than 20 miles. After all pre-processing were completed, the remaining data accounted for 75% of the original dataset.

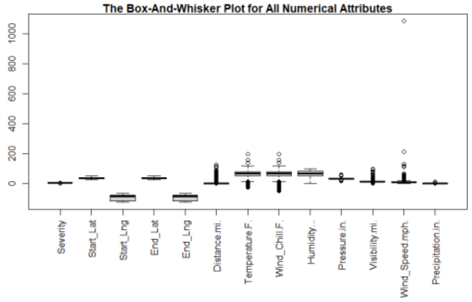


Figure 2. Box-and-Whisker plot of all numerical attributes after removing rows with missing data.

Exploratory Data Analysis (Data Visualization)

Prior to building a predictive model, we explored the cleaned dataset for any visible trends or characteristics. Figure 3 below is a heatmap of all the accidents. From the map, we can see that most accidents are congregated along the west and east coast where population density is higher, with seemingly sparse accident in the central rural United States.

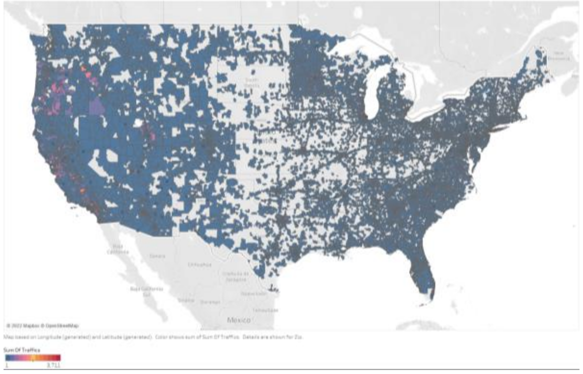

Mexico
© 2022 Mapbox © OpenStreetMap
Map based on Longitude (generated) and Latitude (generated). Color shows sum of Sum Of Traffics. Details are shown for Zip.
Sum Of Traffics
1
3,711

Figure 3. United States heatmap of traffic accidents from 2016 to 2020.

Further assessment into the states and cities with most accidents were conducted.

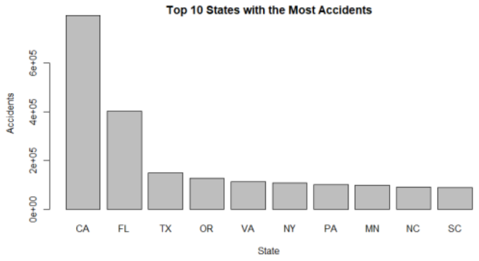


Figure 4a. Top 10 states with the most accidents.

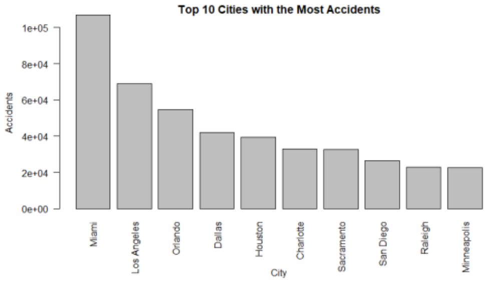


Figure 4b. Top 10 cities with most accidents.

Figure 4a and 4b ranks the top 10 states and cities with the greatest number of traffic accidents. US census data on population density for each US city was also plotted against the number of corresponding traffic accidents. The results are displayed in Figure 5 below.

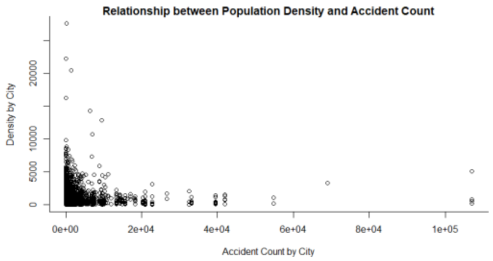


Figure 5. Relationship between population density and traffic accident count based on city.

The resulting graph did not suggest an obvious positive linear relationship between the two, but we would further explore in our models and data visualization web platform.

Potential Impact Analysis of the Pandemic

Since our data contains both pre-pandemic and post-pandemic period, we sampled the data from 2019 to 2021 to evaluate the potential impact of the pandemic period on model accuracy.

Figure 6a calculates the proportion of accidents in 2019-2021 by month. While 2019 and 2021 follow roughly the same pattern with accidents steadily growing from January to December, July and August of 2020 have exceptionally low levels of accident events due to the lockdown period during the pandemic.

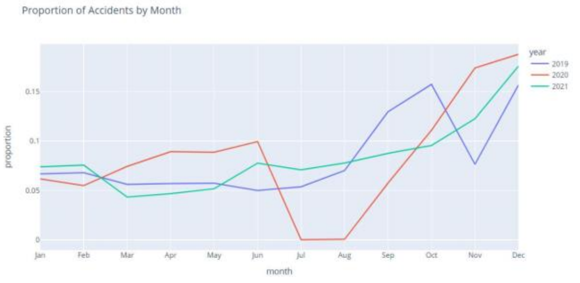


Figure 6a. Proportion of accidents by month.

Furthermore, Figure 6b demonstrates the proportion of accidents changed in 2019 to 2021 by the top 10 cities with the greatest number of traffic accidents. From the chart, we observe that the pandemic period has different impact on the traffic incidents of different cities. Take Miami and Orlando as examples, they didn't make up a large portion of the total accidents in 2019 but have grown largely in 2020 and 2021, which revealed that the pandemic didn't impact the traffic of these cities as much as it did with other cities such as Portland.

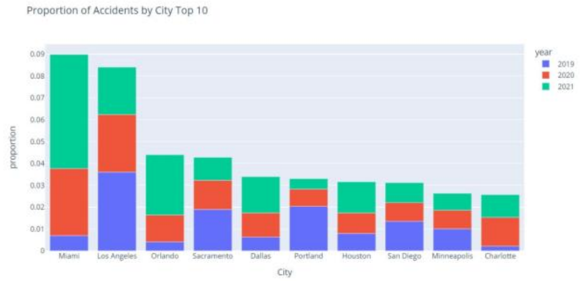


Figure 6b. Proportion of accidents by top 10 cities

To conclude, we found several impacts of the pandemic and lockdown period on the traffic data patterns based on the data from 2019 to 2021, which might cause our model to be slightly biased and predict lower probabilities for July and August periods. However, since the data volume of these two months in 2020 are small, we estimate it to have little impact on the accuracy of our model overall. This could be a future point of development when more data available.

Model – Traffic Accident Occurrence Prediction

The goal of the traffic accident occurrence prediction model is to predict the probability of traffic occurrence given specific date, location, and weather conditions inputs as chosen by the users. The output of the model, probability of accident occurrence, will be able to assist the users in deciding whether to take the trip.

The processed data were split into 20% testing, 16% validation, and 64% training sets. With the target output being binary in nature: accident or no accident, a random forest was used as the baseline model. The initial hyperparameters are set to ten trees with a maximum depth of ten. The features used in the random forest models were *state*, *weather_type*, *weather_severity*, *day_of_week*, *time_in_day,* and *month_in_year*. These features are all categorical and one hot encoding was applied to improve predictions. With the baseline random tree model, we were able to achieve a 91.6% accuracy and a 76.6% AUC, see Figure 7 below, on the testing data.

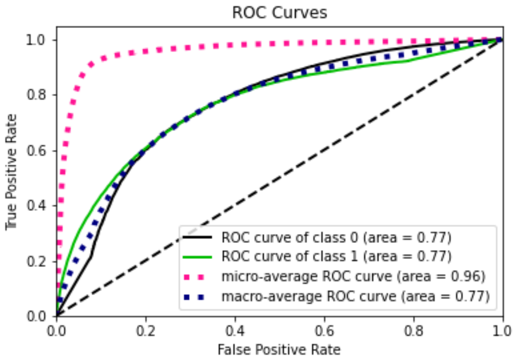


Figure 7. ROC Curve of random tree model.

Model – Traffic Accident Severity Prediction

The second model of interest is the traffic accident severity prediction model. Aside from the pre-existing features in the dataset, additional features such as *hour_of_day*, *day_of_week,* and *month* were also used to build the model. The numerical and categorical variables used for severity prediction are as follows:

$$\begin{aligned} Numerical\ Colum = {} & Severity + Temperature + Humidity \\ & + Pressure + Visibility + Wind\_Speed \\ & + Hour + Day\_of\_Week + Month \end{aligned}$$

$$\begin{aligned} & Categorical\ Column \\ & \quad = City + Weather\_Condition \\ & \quad + Civil\_Twilight \end{aligned}$$

In total, four modeling methods were assessed: Logistic Regression, Decision Tree Classifier, Linear Support Vector Classifier, and Random Forest Classifier. The results of the four models are summarized in Table I below.

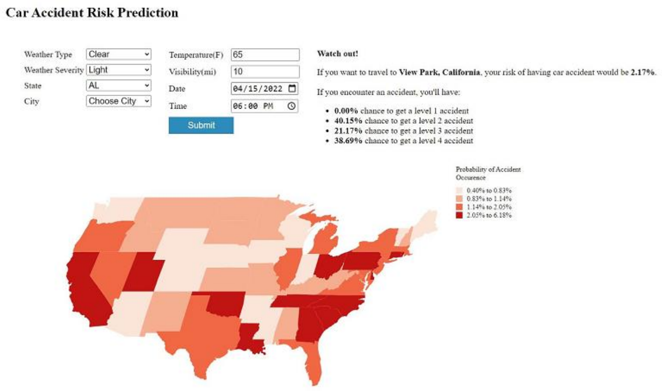
Table I. Model Summary Table

Model	Accuracy	F1-score	Precision	Roc_Auc
Logistic Regression	0.934997	0.903863	0.892018	0.776316
Decision Tree Classifier	0.937548	0.917452	0.911884	0.834185
Linear SVC	0.935100	0.903808	0.906701	-
Random Forest Classifier	0.840936	0.877676	0.934114	0.906396

The model with the best performance is the Decision Tree Classifier model (max depth = 10) with an accuracy of 93.7%, precision of 91.2%, and AUC of 83.4%.

User Interface

To tie the two models together, a web-based user interface is made on a local host. Figure 8 below is a depiction of the interactive interface.



Car Accident Risk Prediction

Weather Type	Clear	Temperature(F)	65
Weather Severity	Light	Visibility(mi)	10
State	AL	Date	04/15/2022
City	Choose City	Time	06:00 PM

Watch out!

If you want to travel to **View Park, California**, your risk of having car accident would be **2.17%**.

If you encounter an accident, you'll have:

- **0.00%** chance to get a level 1 accident
- **40.15%** chance to get a level 2 accident
- **21.17%** chance to get a level 3 accident
- **38.69%** chance to get a level 4 accident

Figure 8. Web-based interactive interface.

The interface allows for input of attributes such as $weather\_type$, $weather\_severity$, $state$, $city$, $temperature$, $visibility$, $date$, and $time$. These are all attributes used to train the algorithm. For those users who are unsure of input for a specified field, they may wish to leave the field empty or with the standard default value. The standard default value is the most prevalent data point of all for that specific category. A US choropleth map using GeoJSON of traffic accident occurrence prediction is displayed for users to visually examine the traffic accident occurrence in relation to other US states. When hovering over the map of different states, a tooltip would show the information of the state and the corresponding predicted risk of the state.

In addition to the accident occurrence prediction and visualization, an accident severity prediction is also provided with the fields submitted by users. The probabilities of level 1 to level 4 accident severities will be calculated and presented on the screen, with level 1 indicating least impact on traffic and level 4 indicating significant impact on traffic such as long delays.

With the fields suggested in Figure 8, the risk of having traffic accident in View Park, California is 2.17% and the most probable accident severity is level 2 at 40.15% which is a minor accident event.

List of Innovations

Different from historical analyses performed on traffic accident data, we used a comprehensive dataset that consists of US country-wide traffic incidents including accidents, constructions, and congestions. In addition to a more comprehensive data, we have introduced the US population density into traffic accident analysis for the first time based on our research. Due to the geographic features of US, we had more complex traffic condition and weather dynamics in our dataset, which would make our analysis more representative and comprehensive than those of previously studies.

Conclusion and Discussion

This project explored the correlations between geographic features, weather conditions and the occurrence and severity of traffic accidents. From the current models we have, we could reach the prediction accuracy of above 90% on both traffic accidents occurrence and severity. Although previous studies have reached model accuracies as high as >99%, those models are not as representative of the dynamic nature of traffic accident prediction as our models are. Our models are based off sets of comprehensive traffic accident data and weather data of the United States such that our models are trained using more dynamic attributes than those of previous studies, thereby resulting in reduced saccuracies of our model. However, although our model predictive accuracies are lower, we suspect that our models have higher explanatory powers than those of other studies.

Our current study also sought to correlate traffic accident data with population density. We proposed that the number of traffic accident should be positively correlated with population density such that as population density increases, so will the number of traffic accidents. However, from our exploratory data analysis, we found that our hypothesis was not supported. Instead, we found that there does not seem to be a correlation between the two. A possible explanation for this behavior may be that the areas with high population density are mostly metropolitan areas with many alternative modes of transportation besides personal vehicles. Therefore, with congestion and high vehicle associated prices such as parking, toll fees, and gas, people would very much opt to own a car in the first place, and possibly travel by other means of transportation such as subways or electric scooters. Hence, since a smaller proportion of the population is driving, the number of traffic accidents in relation to the population density would be much smaller.

With more available traffic accident data in the coming future, we will be able to further polish our models to improve their accuracies and predictive power. Furthermore, we aim to build a mobile application of our current web-based interface to allow for ease of use on the go. Future work on this related topic could involve using population density and traffic accident data to identify emerging metropolitan areas.

"All team members have contributed similar amount of effort on this progress report."

References

1. Roland J, Way P, Sartipi M. Studying the effects of weather and roadway geometrics on daily accident occurrence using a multilayer perceptron model. InProceedings of the Fourth Workshop on International Science of Smart City Operations and Platforms Engineering 2019 Apr 15 (pp. 49-53).
2. Theofilatos A, Yannis G. A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention. 2014 Nov 1;72:244-256.
3. Jaroszweski D, McNamara T. The influence of rainfall on road accidents in urban areas: A weather radar approach. Travel behaviour and society. 2014 Jan 1;1(1):15-21.
4. Park RC, Hong EJ. Urban traffic accident risk prediction for knowledge-based mobile multimedia service. Personal and Ubiquitous Computing. 2020 Aug 19:1-11.
5. Moosavi S, Samavatian MH, Parthasarathy S, Ramnath R. A countrywide traffic accident dataset. arXiv preprint arXiv:1906.05409. 2019 Jun 12:1-6.
6. Moosavi S, Samavatian MH, Parthasarathy S, Teodorescu R, Ramnath R. Accident risk prediction based on heterogeneous sparse data: New dataset and insights. InProceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems 2019 Nov 5 (pp. 33-42).
7. Brühwiler L, Fu C, Huang H, Longhi L, Weibel R. Predicting individuals' car accident risk by trajectory, driving events, and geographical context. Computers, Environment and Urban Systems. 2022 Apr 1;93:101760.
8. Kumar S, Toshniwal D. A data mining framework to analyze road accident data. Journal of Big Data. 2015 Dec;2(1):1-18.
9. Yassin SS. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Applied Sciences. 2020 Sep;2(9):1-13.
10. Razali NA, Shamsaimon N, Ishak KK, Ramli S, Amran MF, Sukardi S. Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning. Journal of Big Data. 2021 Dec;8(1):1-25.
11. Kumar S, Toshniwal D. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). Journal of Big Data. 2016 Dec;3(1):1-11.
12. Yin X, Wu G, Wei J, Shen Y, Qi H, Yin B. Deep learning on traffic prediction: Methods, analysis and future directions. IEEE Transactions on Intelligent Transportation Systems. 2021 Feb 10:1-16.
13. Yu L, Du B, Hu X, Sun L, Han L, Lv W. Deep spatio-temporal graph convolutional network for traffic accident prediction. Neurocomputing. 2021 Jan 29;423:135-147.
14. Lee K, Eo M, Jung E, Yoon Y, Rhee W. Short-term traffic prediction with deep neural networks: A survey. IEEE Access. 2021 Apr 5;9:54739-56.
15. Yuan Z, Zhou X, Yang T. Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. InProceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018 Jul 19 (pp. 984-992).
16. Yasin Çodur M, Tortum A. An artificial neural network model for highway accident prediction: A case study of Erzurum, Turkey. PROMET-Traffic&Transportation. 2015 Jun 26;27(3):217-225.
17. Guerra E, Dong X, Kondo M. Do denser neighborhoods have safer streets? population density and traffic safety in the Philadelphia Region. Journal of Planning Education and Research. 2019:0739456X19845043.
18. Fischer K, Sternfeld I, Melnick DS. Impact of population density on collision rates in a rapidly developing rural, exurban area of Los Angeles County. Injury prevention. 2013 Apr 1;19(2):85-91.
19. He S, Sadeghi MA, Chawla S, Alizadeh M, Balakrishnan H, Madden S. Inferring high-

resolution traffic accident risk maps based on satellite imagery and GPS trajectories. InProceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 11977-11985).