AI for Social Good: Spatio-Temporal Forecast Modeling and Fairness of Traffic Fatality

Fiona Victoria Stanley Jothiraj
University of Washington, Bothell

Keywords: Traffic Fatality, Forecasting, Spatio-Temporal, Fairness, Deep learning, Social Good

Abstract
We are more prone to traffic accidents with an increased number of automobiles on the road and increased maze-shaped routes everywhere. Since the behavior of traffic accidents is not all random but caused due to several factors like road and weather conditions, understanding these relationships between factors and road accidents would help us build an accident predictor. The proposed solution consists of implementing various forecasting and deep learning modeling techniques to estimate the hotspots for traffic fatality in the future. The prediction performance in terms of accuracy and error shows promising results.

1. Introduction

The motivation behind this research project was to highlight the importance of applying AI knowledge to areas that are beneficial to human lives. According to World Health Organization statistics, approximately 1.25 million people lose their lives in road traffic accidents worldwide every year, which means that one person is killed every 25 seconds. Another motivation for choosing the traffic fatality domain is because it provides the ML developer to provide causal inference, which is far more important than prediction. Typical machine learning and deep learning methods have proved to be efficient in finding correlations in data but are unskilful in determining causation or interpretability.

2. Related Work

Scientists from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and the Qatar Center for Artificial Intelligence have developed a deep learning model that predicts high-resolution crash risk maps. The model used historical crash data, road maps, satellite imagery, and GPS traces as input. The risk map output describes the expected number of crashes over a period of time in the future to identify high-risk areas and predict future crashes.

3. Data for Traffic Fatality Analysis

A. Data

FARS is a nationwide census providing NHTSA, Congress, and the American public yearly data regarding fatal injuries suffered in motor vehicle traffic crashes. Data is present from the year 1975 to 2020. To keep the sensitive data anonymous, personal identifying information such as names, addresses, or social security numbers are kept private. This keeps the publicly available FARS data fully conformed to the Privacy Act.

There are generally two criteria based on which the data get chosen to be added to the dataset:
Criteria 1: Fatal motor vehicle crashes involving a motor vehicle traveling on a traffic way customarily open to the public.
Criteria 2: Death of a motorist or a non-motorist within 30 days of the crash

B. Properties of Data

i. The data is neatly presented in a tabular fashion within several .CSV files
ii. Approximately 25 .csv files are present each year with an average of 50,000 rows in each CSV
iii. Detailed information on nearly 30,000+ accident cases
iv. There are several types of features contained in the dataset such as

a. Date and time
b. Geospatial data such as latitude and longitude
c. Textual information describing the accident
d. Categorical - Nominal features are usually coded

v. Presence of several redundant features - both categorical (with text) and encoded

C. EDA

For this research project, data were collected from 1995 to 2020. Initial analysis of the data shows that the majority of the fatalities occurred in the states of California, Florida, and Texas from 2015 to 2020. Also, seasonality bar plots indicated that weekends and the Fall season were most likely the times for chances of a traffic fatality. Graphs also revealed that most accidents occurred during a clear day with daylight which raises some questions about how that could be possible (since it is human intuition to think that a cloudy or night times is a more obvious choice). However, since most of the data is collected from the southern states of North America, it is possible that data containing long-summer months with clear skies could be skewing such graphs.

D. Spatio Temporal Analysis

Initially, a study analyzing all the information available about traffic accidents was performed. Fatal accident information present spatially, temporally, and based on factors such as age, gender, weather, and lighting conditions were studied. Figures 1 and 2 show the temporal and spatial analysis results.

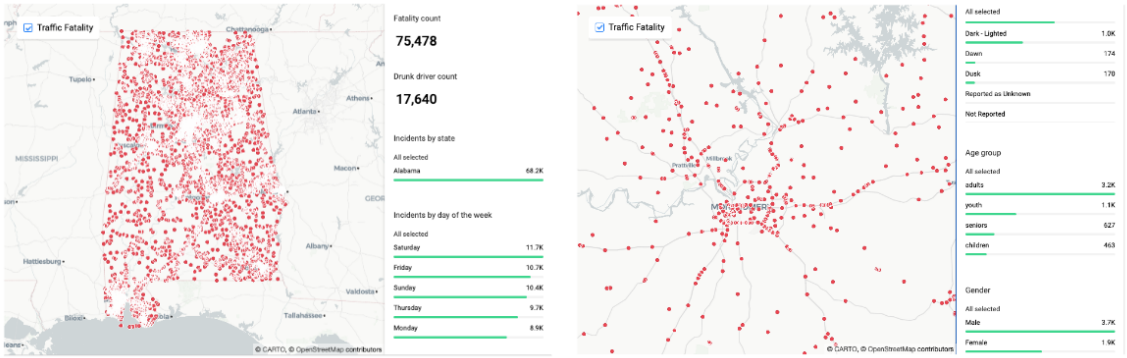

Fig 1. Spatio-Temporal analysis with multiple factors

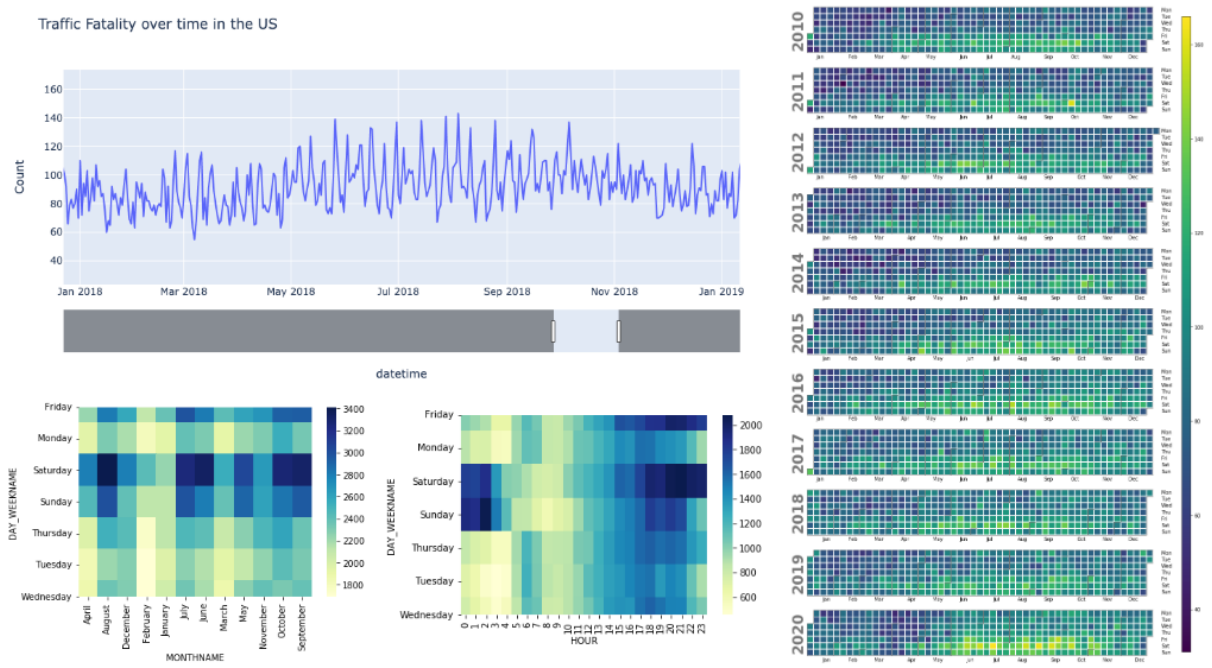

Fig 2. Temporal analysis

4. Methods

A. Factors causing multiple fatalities

The project's first step involved using a classic machine learning model to visualize the critical causal features that caused multiple traffic fatalities in the US. The ML model used for this purpose is a *Random Forest Classifier* which achieved an accuracy score of 0.918. The results of the classifier were visualized using a SHAP plot that provides the interpretability of factors causing multiple fatalities, as shown in Figure 3. The plot shows that factors like a motor vehicle, day, month, hour, and drunk drivers all contribute to multiple traffic fatalities.

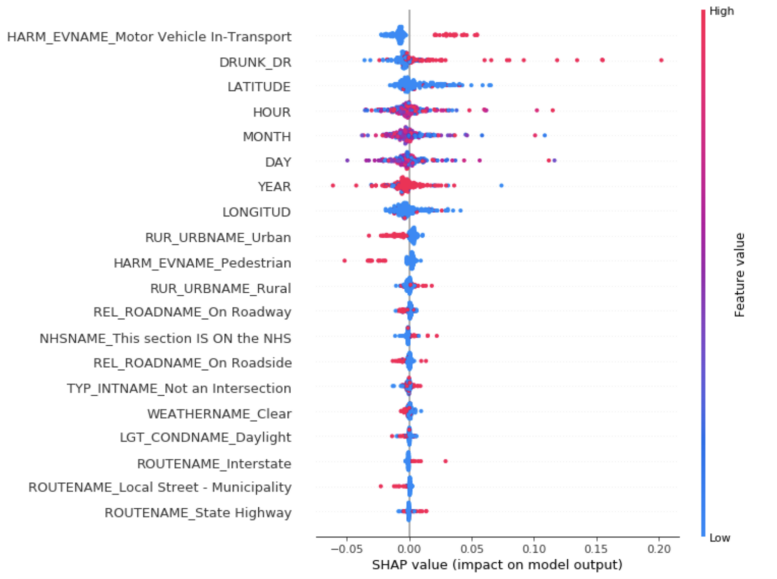

Fig 3. Variable Importance Plot - Global Interpretability (SHAP plot)

B. Traffic accident hotspot prediction

Once the critical causal features were identified, deep-learning networks were used to predict when and where the accidents happened. To achieve this, an unsupervised *ConvLSTM* model (Shi et al., 2015) was employed to predict the subsequent frames of accident hotspots. ConsLSTM is a convolution neural network combined with the LSTM Network that determines the future state of a certain cell in the grid by the inputs and past states of its local neighbors. This model is chosen because it captures the underlying local spatial-temporal correlations (Zhang., 2021). The input to the model is a collection of heatmaps (for each state in the US) indicating the exact location on the map where an accident occurred for every DateTime entry. Figure 4 shows the structure of a convlstm cell and Figure 5 shows the prediction results.

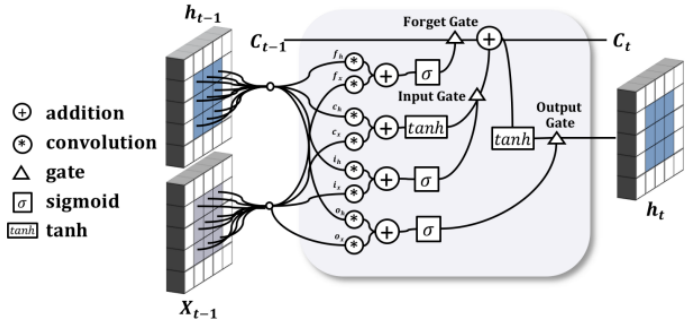

Fig 4. Structure of ConvLSTM cell

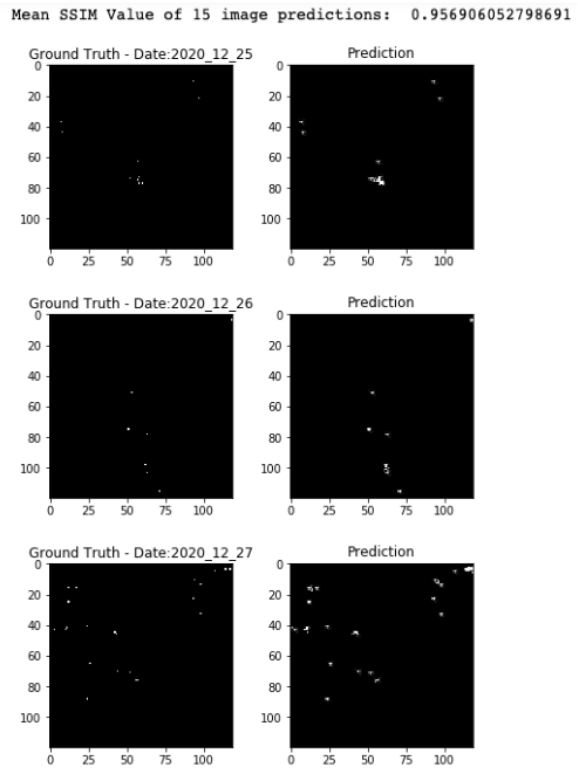

Fig 5. Prediction output

C. Fairness in rich and poor counties of the US

With machine learning systems being used everywhere to make critical life-saving decisions, another task was performed to ensure that traffic fatality occurring in different parts of the US was fair irrespective of the living conditions of people. The fairness of two groups of counties ie. rich or poor based on the SVI index was measured using forecasting model results. For this, the *Prophet model from Facebook* was used as a forecasting model. The prophet model works best with time series with strong seasonal effects and several seasons of historical data. It is also robust to missing data and shifts in the trend and typically handles outliers well.

The experiments were conducted with the same number of counties in both groups (rich and poor). The results from the forecasting model show that the predictive results are very different for both the rich and the poor groups indicating that the fairness metric named *demographic parity* is not satisfied. The forecasting model outputs are dependent on a given sensitive attribute which implies a degree of unfairness between both groups The results are shown in Table 1.

	Number of counties		**Mean absolute error**	
Experiment	**Rich**	**Poor**	**Rich**	**Poor**
1	64	63	0.607	0.761
2	315	413	0.617	0.781
3	472	785	0.639	0.803

Table 1. Performance metrics from Prophet Forecasting model

6. Discussion

We can conclude from the observed results that ConvLSTM models are beneficial for forecasting the hotspots of traffic fatalities in small areas. However, they cannot account for other factors like age, gender, type of car, road condition, and weather during forecasting. The future scope for this work could be converting the essential features that cause multiple fatalities into 4D tensors and feeding them as inputs to a ConvLSTM model.

Another neural network model named ST-Resnet (Zhang et al., 2017) aims to solve this challenge by aggregating external factors, such as weather and day of the week, to predict fatality traffic in every region of the country. This socially beneficial project is a great way to understand or develop future roads keeping in mind the factors involved in crashes.

7. References

Shi, X., Chen, Z., Wang, H., & Yeung, D. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. ArXiv. https://arxiv.org/pdf/1506.04214.pdf

Zhang, J., Zheng J., Qi, D. (2017) Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction, Microsoft Research. ArXiv. https://arxiv.org/pdf/1610.00081.pdf

NHTSA Manuals & Documentation - Fatality Analysis Reporting System (FARS) https://crashstats.nhtsa.dot.gov/#!/PublicationList/106. Accessed on: 05/25/2022.

Zhang C. (2021) Spatio Temporal ConvLSTM for Crash Prediction https://towardsdatascience.com/spatial-temporal-convlstm-for-crash-prediction-411909ed2cfa Accessed on: 05/25/2022.

CDC/ATSDR Social Vulnerability Index https://www.atsdr.cdc.gov/placeandhealth/svi/index.html Accessed on: 05/25/2022.