Applied Data Science Capstone

Capstone Project - Car accident severity (Week 2)

1. Introduction

1.1 Background

Road crash fatalities and disabilities have become very common these days and are recognised as a major public health issue. Approximately 1.35 million people die and 30-50 million suffer non-fatal injuries in every year from road accidents globally. There are different factors that cause road accidents, including speeding, weather, time, road conditions etc. Data analytics has emerged as a great technique that allows data scientists to extract meaningful information from a large set of data. Accident modelling on the different contributing factors of road accidents could provide an insight into the leading conditions for road accidents.

1.2 Problem

In this project, road accident data for Washington state has been analysed. Machine learning techniques have been used on road accident data for King county to build a model which can be used for real time accident prediction for the same county.

1.3 Interest

People who live in Washington state, especially the drivers would be very interested in real time accident prediction so that they can drive more carefully or change their route. People from other states would also be interested in this analysis since they can also avoid similar travel situations to prevent accidents.

2. Data acquisition and cleaning

2.1 Data sources

The dataset used for this project is obtained from Kaggle website and can be downloaded from here. The dataset consists of traffic collision events from 2016 to 2020 for the 49 states of United States of America (USA) and it contains 3513617 rows and 49 columns.

2.2 Data cleaning

The dataset available from Kaggle website was generated by combining the traffic event data from several providers. The columns containing Start_Time and End_time character data was converted to date/time data type. The amount of time was extracted in minutes for each accident and after that time duration values were checked for negative numbers if any. The outliers were set to not a number (NAN) and the negative time duration rows were dropped.

2.3 Feature selection

The original dataset contains 49 features. After the preliminary data analysis, only the relevant features were selected based on their impact on accidents. The following 34 features were selected for the improved accuracy of the prediction.

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

3. Exploratory data analysis

In this project, as the first step, the road accident data for Washington (WA) state was analysed and the possible factors that could lead to accidents were identified. Among the various counties in WA, it was found that most of the road accidents are happening in King county. Therefore, machine learning algorithms were applied on various traffic collision events from King to build models that could provide a reference guide to people in King to drive safely.

3.1 Monthly crash rates

The traffic collision events from June 2016 to May 2020 were analysed for each month for WA and is shown in Figure 1. Looking at the Figure, there in no regular trend for the number of accidents by month and the highest number of accidents happened in the month of October 2017 for the given time period.

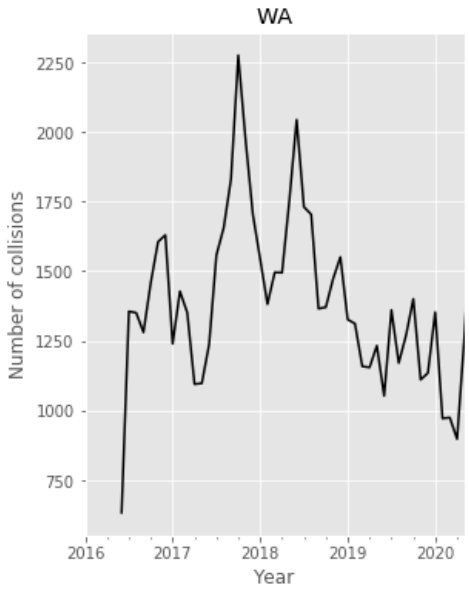

Fig.1 Number of collisions per month for WA.

3.2 Severity of an accident

Figure 2 shows the percentage distribution of accident severity for WA. From the Figure, most of the accidents are in severity level 2 (61%), followed by level 3 (34%). Only 4% of the accidents fall into the most severe case (level 4).

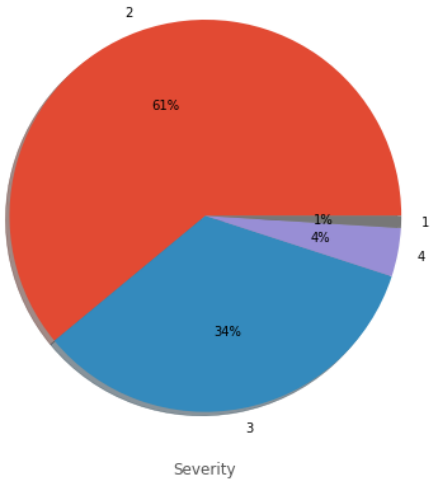

Fig. 2 Percentage distribution of road accidents severity in WA.

3.3 Distribution of accident by day and time

Most of the accidents are happening during daytime compared to night (Figure 3(a)). As expected, more accidents are happening during weekdays than weekends. There is a slight increase in the percentage of accidents (1%) when you go through Monday to Friday (Figure 3(b)).

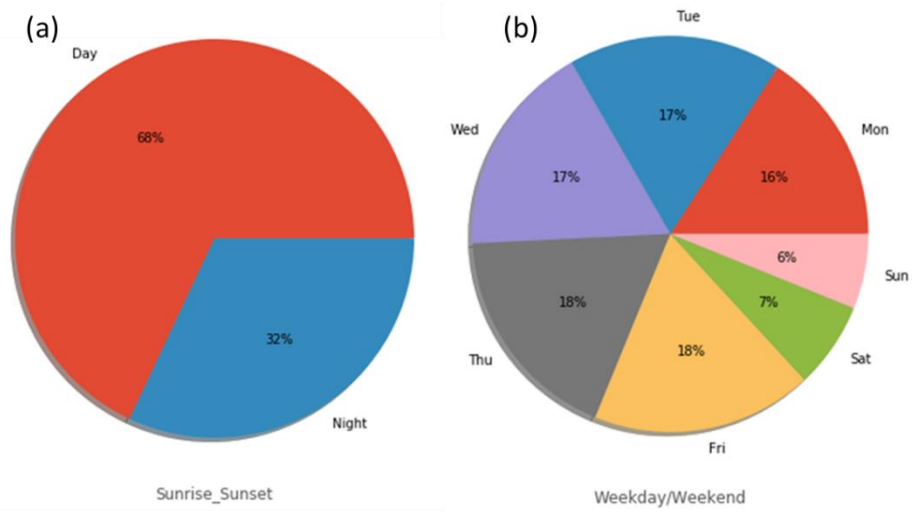

Fig. 3 Percentage of accidents for (a) day vs. night and (b) days of a week.

As shown in Figure 4, it is not recommended to travel 7-8 am in the morning and 4-5 pm in the evening as most of the accidents happened during these hours. During the weekends, most of the accidents are happening right after noon.

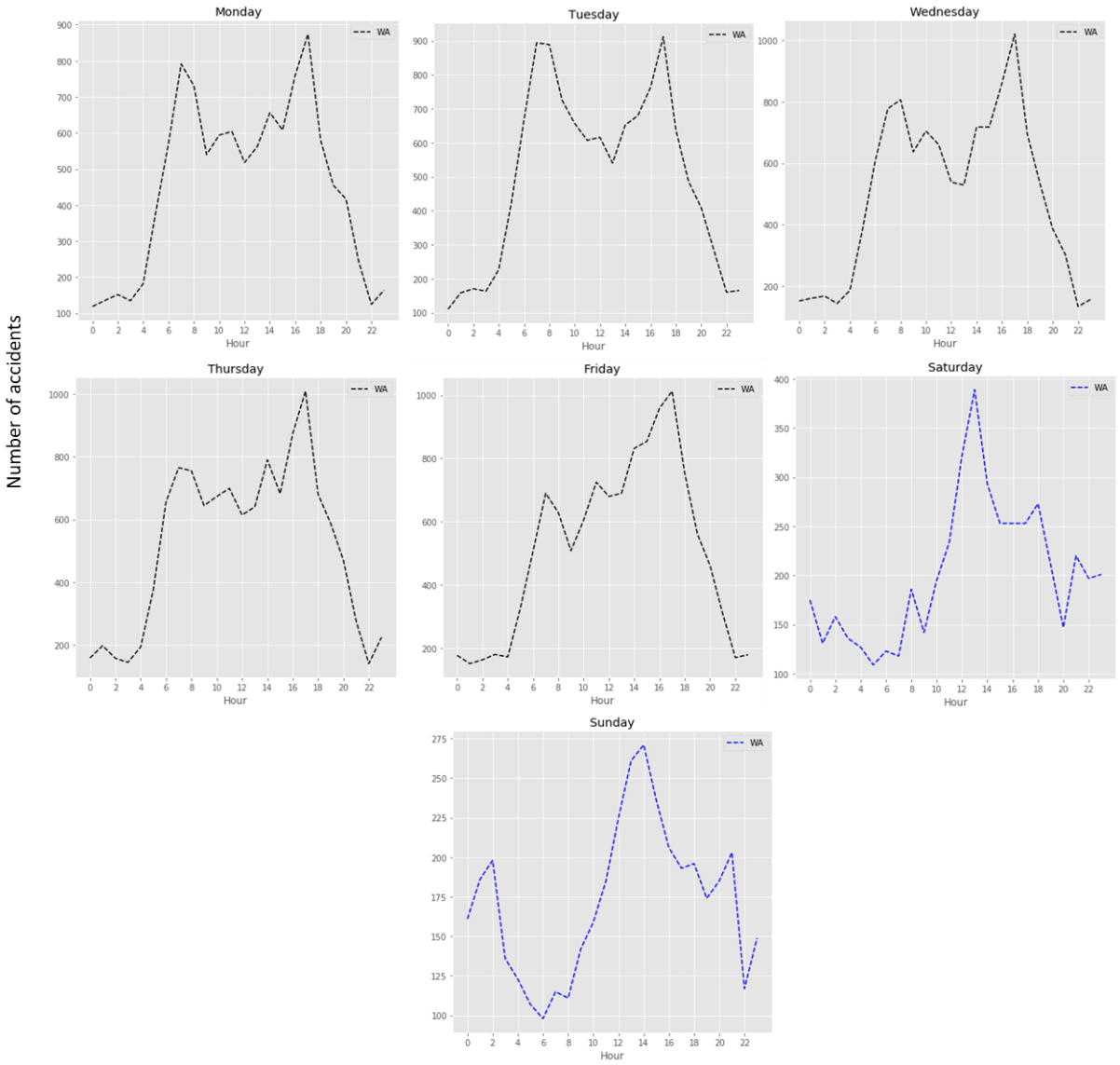

Fig. 4 Hourly distribution of accidents on each day for WA.

3.4 Distribution of accidents by County and City

In WA state, the highest number of accidents are reported from King county (Figure 5(a)). Above 50% traffic collisions are happening in King compared to other counties. The largest city in King, Seattle, reports the maximum number of traffic collision (38%) when compared to other cities (Figure 5). All other cities have a collision rate less than 10%.

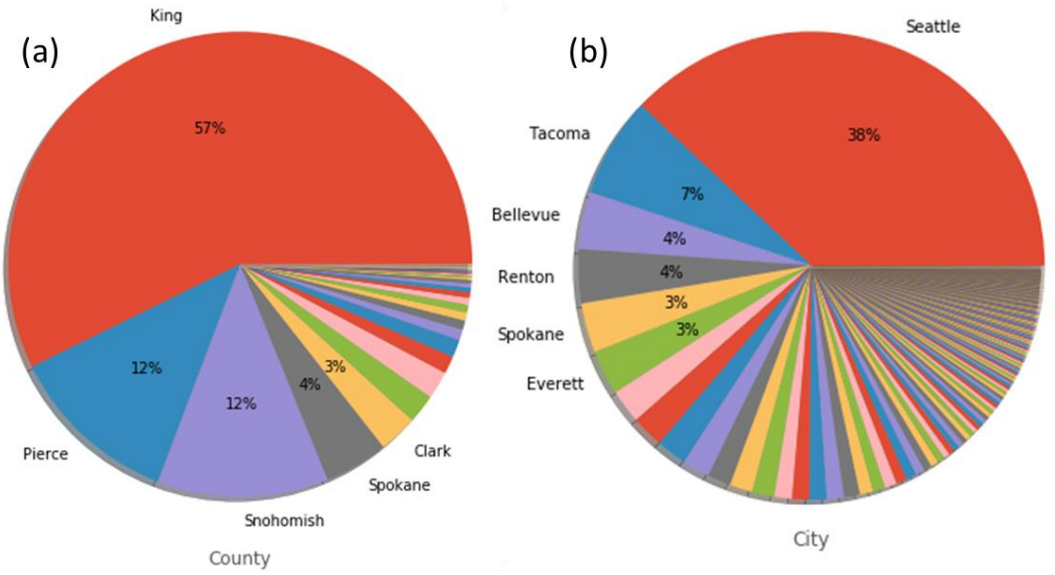

Fig. 5 Percentage distribution of accidents by (a) county and (b) city.

3.5 Distribution of accidents by street side and location

As shown in Figure 6(a), 79% of accidents occurred at the right side of the street in WA. Figure 6(b) represents the percentage of accidents based on locations where most of the accidents are happening. 30% of accidents are occurring at traffic signal, followed by crossing (25%).

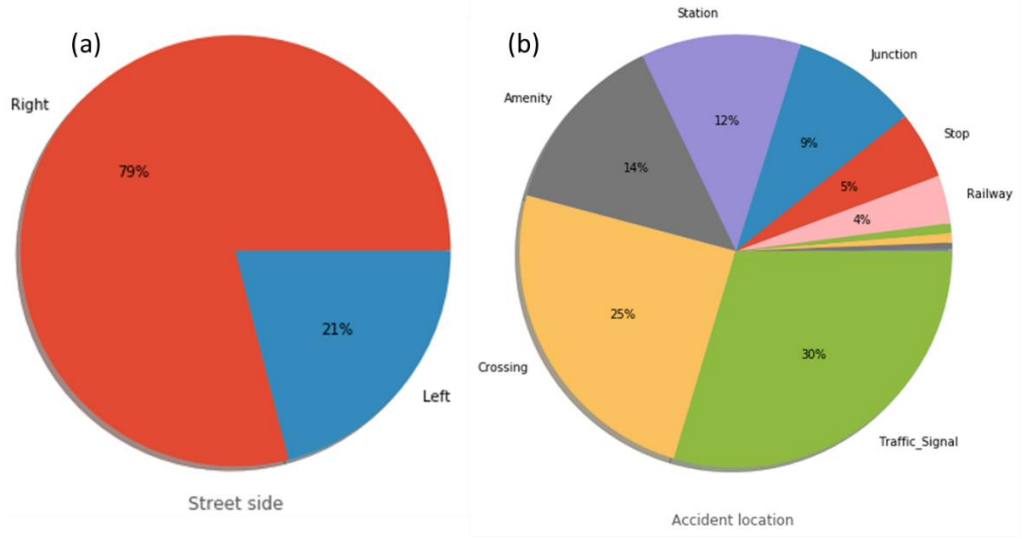

Fig. 6 Percentage distribution of accidents by (a) street side (b) location.

3.6 Distribution of accidents under various weather conditions

The various weather conditions were analysed to understand the effect of these features on accident prediction (Figure 7). Clear, overcast, mostly cloudy, and light rain were the most common weather conditions at which accidents happened at higher rate. Since most of the time the weather is clear, we cannot conclude that clear weather is a contributing factor for accidents. Nevertheless, overcast, mostly cloudy and light rain are realistic factors for accidents.

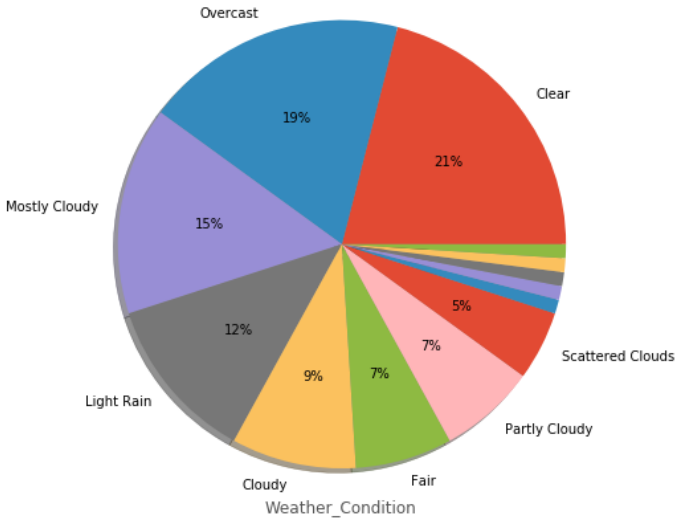

Fig. 7 Percentage distribution of accidents for various weather conditions.

Figure 8 represents the percentage of accidents under different weather conditions for each accident severity. Clear and overcast conditions dominate for accident severity level 2,3 and 4.

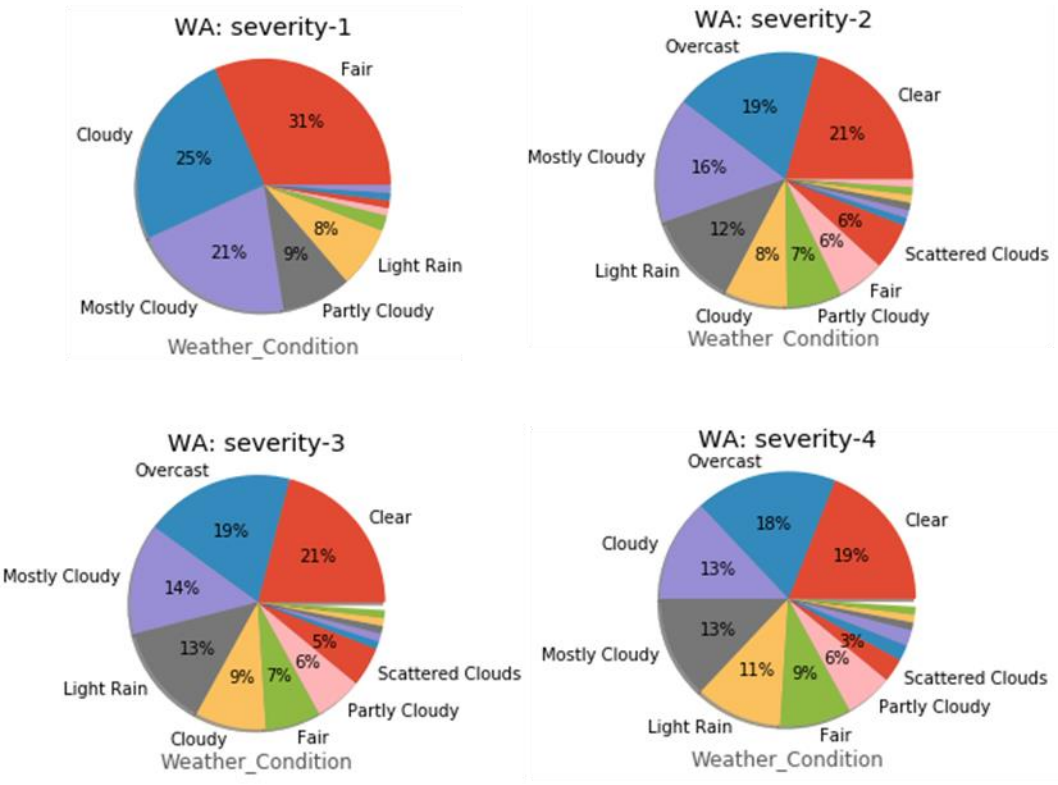

Fig. 8 Percentage weather conditions for each accident severity.

4. Machine learning algorithms and feature engineering

Various supervised machine learning algorithms (Logistic regression, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest classification) are used to predict the accident severity at King county, where most of the accidents are happening in WA. As explained in section 2.4, 34 features were selected for the improved accuracy of the prediction. Figure 9 shows the accuracy of different machine learning algorithms to predict the accident severity for King county. Among the various models, Random Forest classification shows the highest accuracy with a score of 0.92 and KNN shows the least accuracy (0.68). Figure 10 shows the top ten features for prediction of accident severity for King county using Random Forest classification model.

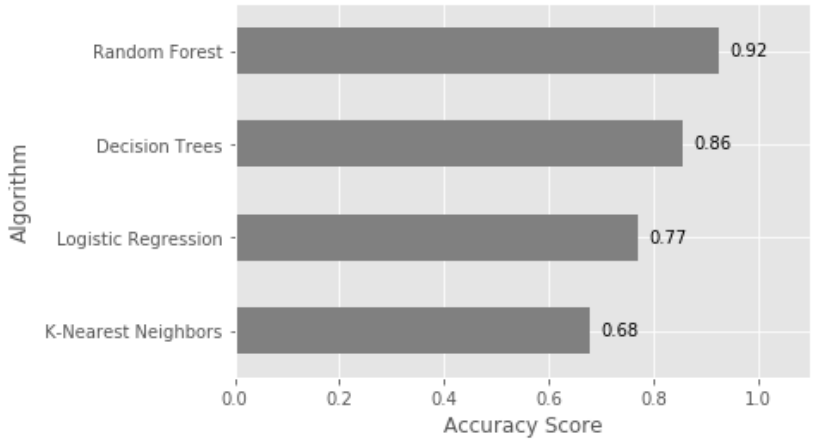

Fig.9 Accuracy of different machine learning algorithms to predict the accident severity.

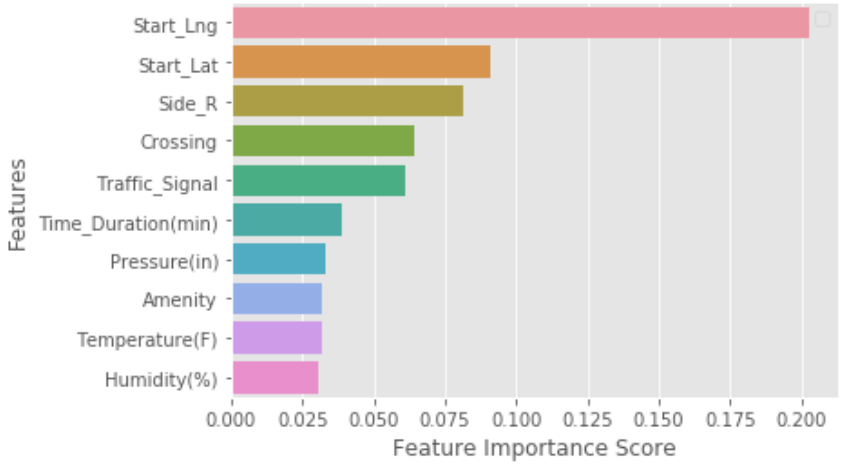

Fig. 10 Top ten features for prediction of accident severity for King county using Random Forest classification model.

5. Discussion

Exploratory data analysis has been done to figure out the various factors leading to road accidents. Monthly collision rate analysis did not give any insight for predicting the possibility of an accident. The severity of accidents was analysed in a scale of 1-4 (level 4-most severe). It was found that most of the accidents are in the level 2. Most of the accidents are happening during daytime and it could be due to the heavy traffic compared to night time. It is not recommended to travel around 7-8 am and 4-5pm on weekdays as most of the accidents are happening during these rush hours. Similarly, early afternoon is not recommended for weekends. Above 50% traffic collisions are reported from King county in WA and most of those accidents are happening in Seattle city. There

is huge possibility for road accidents on right side of the street in WA and in terms of locations, traffic signals and crossings are the most dangerous places. 79% of accidents occurred at the right side of the street in WA. 30% of accidents are occurring at traffic signal, followed by crossing (25%). Clear, overcast, mostly cloudy, light rain, and cloudy are the topmost 5 weather conditions for accidents. Among them we could eliminate clear condition because most of the days are clear. The accident severity for King county was predicted by applying machine learning algorithms and Random Forest classification showed the highest accuracy. Top ten features for prediction of accident severity for King county was extracted using Random Forest classification model.

6. Conclusion

In this project, traffic collision data for WA state has been analysed and machine learning algorithms were applied to predict the accident severity for King county. Various features were analysed and the most important contributing factors which could lead to road crash fatalities and disabilities were identified and listed. The built model to predict the real time accident severity will be very useful for the people in King.

7. Acknowledgement

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

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.