Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02215

ALY 6110 Data Management & Big Data

CRN: 72083

Project name: U.S. Traffic Accident Analysis

Instructor: Valeriy Shevchenko

Team:

Anju Yeh: yeh.an@northeastern.edu

Binyu Zhai: zhai.b@northeastern.edu

Liuzhao Tang: tang.liu@northeastern.edu

Jincheng Wang: wang.jinc@northeastern.edu

Zhongying Lan: lan.z@northeastern.edu

Summary

In this project, we use the U.S. Traffic Accidents dataset from Kaggle. There are 3.5 million records and 49 columns. This is a countrywide car accident dataset, which covers 49 states of the USA. It was collected from February 2016 to June 2020. The columns contain the Location of accidents, the Severity of accidents and the Weather condition of accidents, etc.

We will use Databricks as our cloud platform for analyzing. Besides, we choose Python language to do the whole process. Include cleaning up the raw data, performing the exploratory data analysis (EDA) and drawing a series of traffic accidents maps to figure out what is the trend of the U.S. traffic accidents. And then, we build a model to do the feature selection. At last, we would provide some relevant insights and suggestions.

This project aims to reveal the factors that might lead the accident and even its severity changes. We would like to find out as well as to show the concepts such as what time scale is the peak, which city takes the lead by the highlights and visualizations to the audience. Also, we will make modeling and predictions to indicate what factors are significant to the accidents' severities and try to predict the result. In general, we expect the result of the project can give an overall view to the audiences. As well as we hope that people can avoid tragic accident by paying attention to the key factors that are pointed out by the project.

Content

Step 1. Data clean

We conduct a preliminary clean-up of the dataset in order to perform further analysis. At very first, we quickly peek through the data shows we have all sorts of data types, like string, datetime, float, boolean, and integers. Then, we check the missing values, the outcome is showing as follow:

	Total	Percent
End_Lat	2478818	70.548896
End_Lng	2478818	70.548896
Number	2262864	64.402694
Precipitation(in)	2025874	57.657793
Wind_Chill(F)	1868249	53.171675
TMC	1034799	29.451104
Wind_Speed(mph)	454609	12.938490
Weather_Condition	76138	2.166941
Visibility(mi)	75856	2.158915
Humidity(%)	69687	1.983341
Temperature(F)	65732	1.870779
Wind_Direction	58874	1.675595
Pressure(in)	55882	1.590441
Weather_Timestamp	43323	1.233003
Airport_Code	6758	0.192337
Timezone	3880	0.110428
Zipcode	1069	0.030424
Nautical_Twilight	115	0.003273
Astronomical_Twilight	115	0.003273
Civil_Twilight	115	0.003273
Sunrise_Sunset	115	0.003273
City	112	0.003188
Description	1	0.000028

Then, fill 'Humidity (%)', 'Precipitation(in)', 'Wind_Chill(F)', 'Wind_Speed(mph)', 'Visibility(mi)' with 0 because it is possible to have no records for these columns. For example, it's possible to have zero rain if rain didn't fall that day. Besides, we fill the 'Temperature(F)' and 'Pressure(in)' with the average value. Then, we Fill the 'City' column. Since we have a 'State' column. So, we fill the city column with the most occurring city of the state it belongs to. At last, we drop useless columns and empty rows. Following are the reasons for deleting these columns:

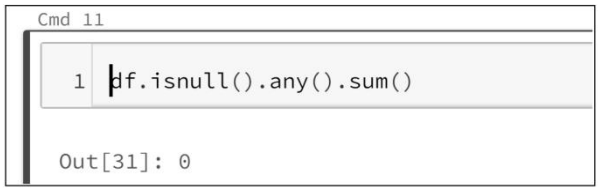
1.It's hard to impute End-lat and End-lng, because it can only be collected when the accident affected a huge area of road. And it's 77% the same as start-Lat and start-lng. So, dropping is the logical choice.

2. TMC stands for Traffic Message Channel. Its purpose is to deliver information about traffic distortions or warnings to mobile receivers such as navigation devices. Which we don't need.

3. Weather_Timestamp: Shows the time-stamp of the weather observation record (in local time).

4. Drop some useless columns like ID, Source, TMC, etc.

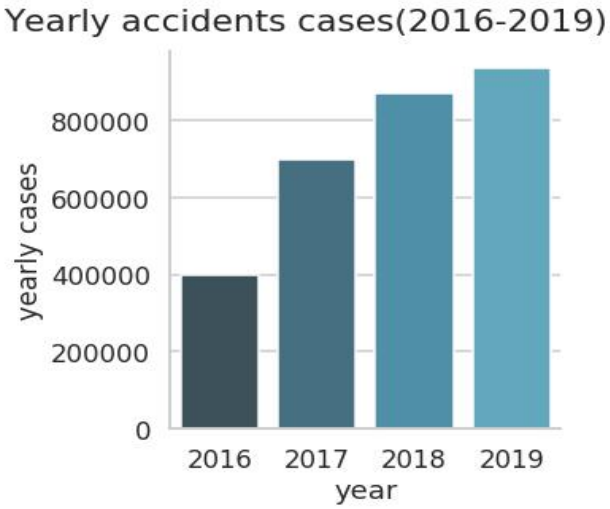
There is zero missing value in our dataset now.



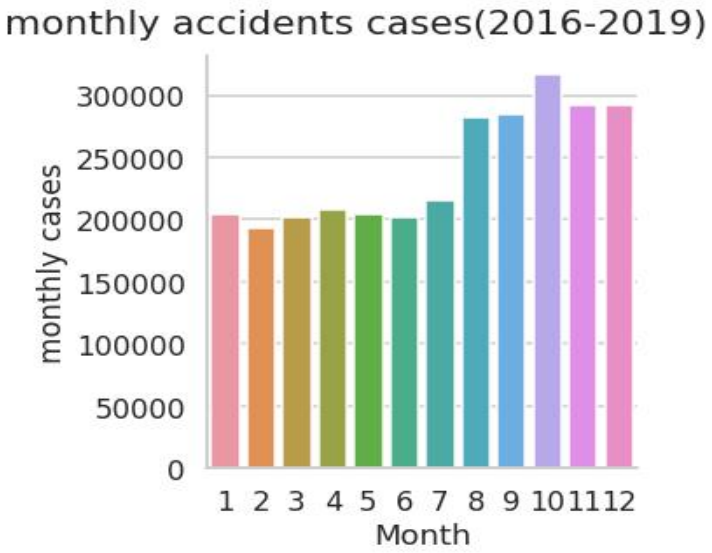
Step 2. Exploratory Data Analysis

2.1 Analyze number of accidents by Time

In this section, first we have made statistics on traffic accidents that occurred in the four years from 2016 to 2019. From the figure, we can find that in these four years, traffic accidents have been increasing year by year. In 2019, there were approximately 900,000 traffic accidents in the United States.

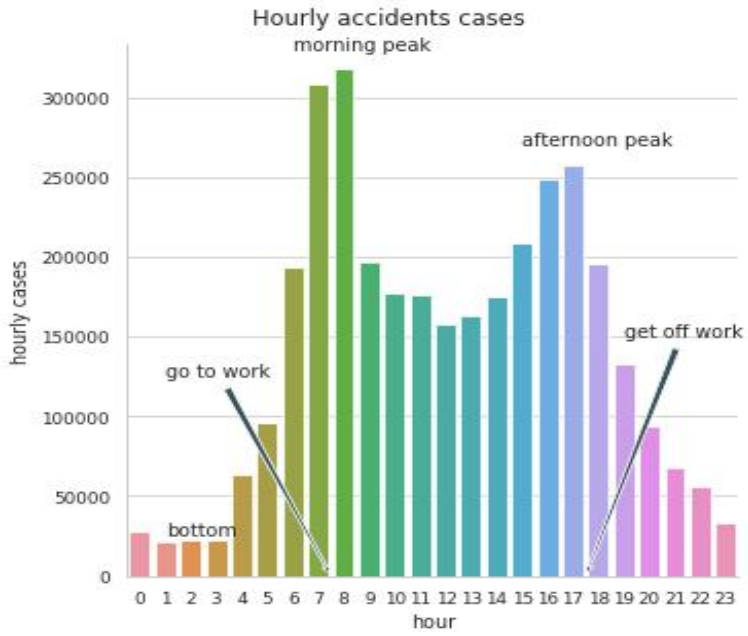


Next, we counted the average traffic accidents that occurred every month for the period from 2016 to 2019. Through analysis, we can know that the number of traffic accidents in the five months from August to December is significantly higher than other months. We guess there are worse weather conditions in the summer and winter. Weather factors may be the cause of traffic accidents



Finally, we counted the average number of traffic accidents that occurred per hour in a day. From the figure, we can know that most accidents occur during the day, and peaks occur at 8 am and 5 pm. Eight in the morning is when people go to work and five in the afternoon is when people get off work to get home. These two time periods are

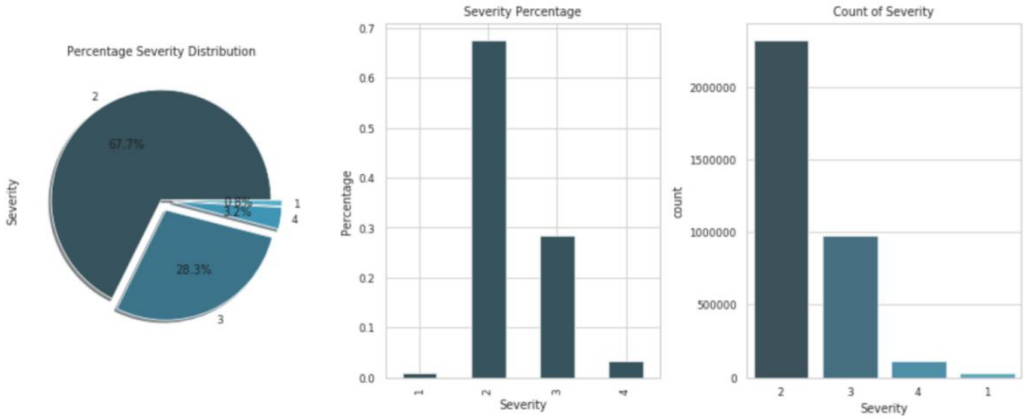
peak times for people to travel, so the possibility of traffic accidents is greater. Compared with these two times, at other times, people tend to rest at home instead of traveling, so the number of traffic accidents is much lower.



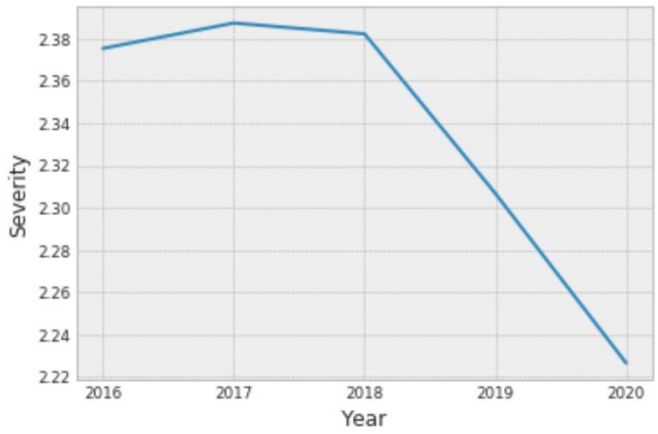
2.2 Analyze number of accidents by Severity

Column ‘Severity’ shows the severity of the accident, the number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).

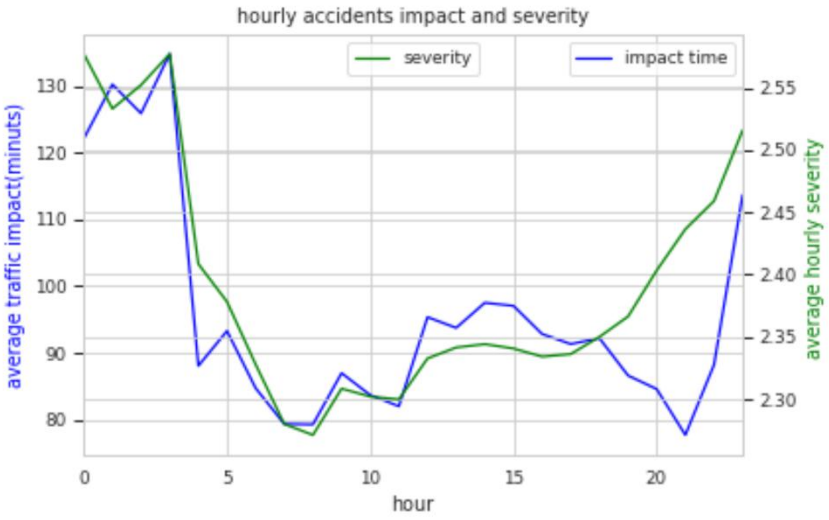
First, we take a look at this feature. Obviously, 67.7 % Accidents fall in the Severity class 2 followed by Severity class 3,4 and 1.



Then, we try to figure out the ‘severity’ changes in years from 2016 to 2020. It's easy to find that it has declined year by year from 2017 to 2020.

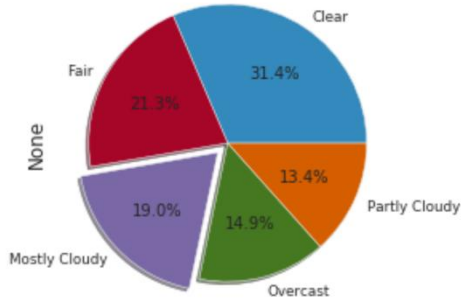


At last, we create the following chart to discuss the severity changes hourly. Serious car accidents always happen between 5pm and 3am.



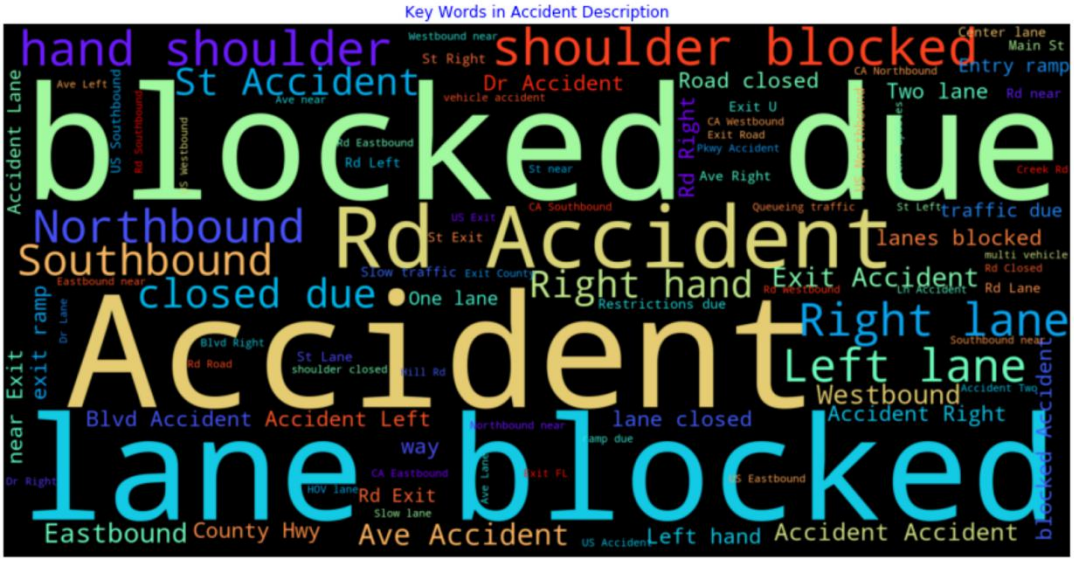
2.3 Analyze number of accidents by Weather condition

In this section, our goal is trying to identify what kind of weather conditions cause the most car accidents. There are a total five weather conditions that can be found from the following pie chart such as partly cloudy, overcast, mostly cloudy, fair and clear. As we have excluded the minimal cause data such as light rain, rain, light snow, thunderstorm and scattered clouds etc. in order to extract the most significant factors that caused the car accidents. Surprisingly, most car accidents happened during the clear weather with 31.4%. Therefore, we could infer that drivers are usually more cautious driving during poor weather conditions.



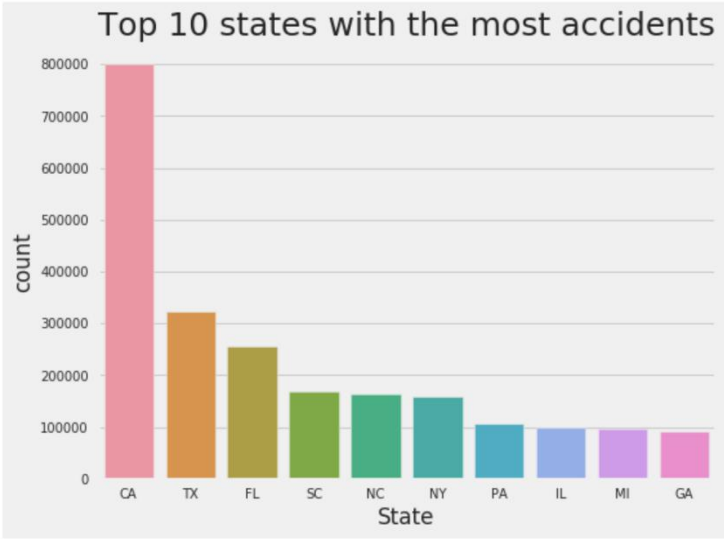
2.4 Analyze number of accidents by Description

Description column indicates the natural language description and details of the accidents. We executed the wordcloud package in order to help us in analyzing the texts and quickly visualize the key words. We can find from the following picture that accident, lane blocked and Rd Accident are the most frequent words that appeared in the description of US traffic accident dataset.

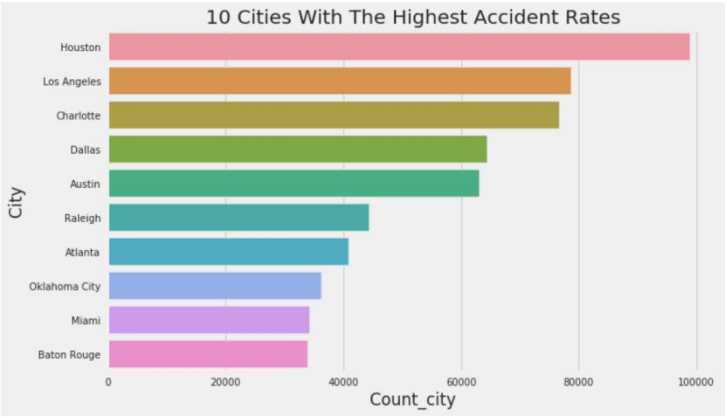


2.5 Analyze number of accidents by State and City

State column records all the states that contain car accidents. Based on the following bar chart, we identify that CA has the most car accidents followed by TX and FL. With the logical assumption, we could infer that this resulted from the fact that CA has the highest population in the US.



Taking a closer look into the city bar chart, we found out that Houston has the most accidents followed by Los Angeles and Charlotte among all US cities. In this case, we came up with a point that Taxes is the 2nd biggest state in the USA, thus this can lead to more drivers speeding to reach their destination faster.



Step 3. Draw maps

Figure X displays maps of traffic accidents from 2016 to 2020. The first finding from the map is that the cases of traffic accidents are increasing. Second, we can find that most of the cases occur along the coastline area and Great Lakes Region. Finally, the traffic accidents in the central region are increasing in number and scope year after year.

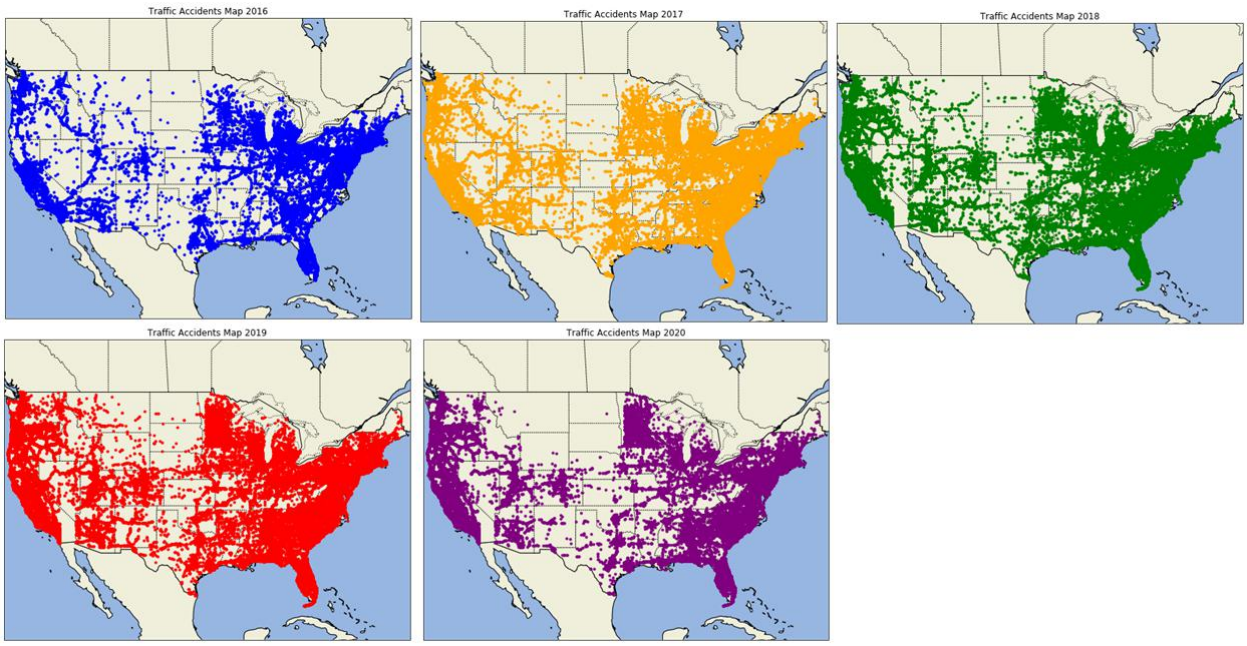


Figure X: Maps of traffic accidents from 2016 to 2020

Step 4. Feature Selection and Modeling Prediction

In this step we mainly perform two processes to make some determinations, which are feature selections by random forest and forecasting severity of accident by multi-logistic regression modeling. By the feature selection, we would like to introduce the key factors that lead the severity, as well as predict the severity of accidents by model.

We believe that this is an important and interesting step because data can usually give the audience some more surprising information than the imagination. Here, we can logically assume some key factors such as the humidity, wind level, and weather might affect the severity of accidents.

First, we need to prepare more on reshaping the data set before doing the work. If we directly use “get_dummies” to transform the categorical information, we will get more than 10000 columns, which makes the work to be harsh. So, we drop some not essential columns like “Country”, “City”, “Descriptions”, etc. and then the data size has been cut down to the 203 columns as shown below.

Feature Selection:

We use randomforest classifier and import feature selection function from sklearn package to get the result as the figure shown below.

```
Out[19]: array([ True,  True,  True,  True,  True,  True,  True,  True,  True,
        True, False, False,  True, False,  True, False, False, False,
       False, False, False,  True, False, False,  True,  True, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False,  True, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
```

Command took 3.10 minutes -- by wang.jinc@northeastern.edu at 12/4/2020, 6:04:53 PM on 6110

(In the figure, True represents that the columns are selected as the feature. False represent that the columns are not as significant as the feature)

In general, among the more than 200 factors, only 15 of them are constantly significant to the severity of accidents. They are ‘Start_Lat’, ‘Start_Lng’, “'Distance(mi)', ‘Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)’,’ Crossing’, ‘Junction’, ‘Traffic_Signal’, ‘Side_L’, and ‘Side_R’. The result is a little amazing that the state and weather are not the factors that lead to the severity of the accident.

Prediction by Multi-Logistic Regression:

Because the Severity is from 1 to 4 and each number represents its accident severity situation, we prefer to deal with categorical data, which need to use logistic regression as the model. (The result of the Prediction V.S. Reality is briefly shown in the figure below)

Out[21]:

ID	Actually Severity	Predicted Severity	Start_Lat	Start_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
A-927574	2	2	26.608829	-80.065338	0.000	82.0	82.0	69.0	29.95	10.0	13.0	0.00
A-823441	3	2	37.751369	-79.414055	0.000	42.0	33.0	86.0	25.82	10.0	23.0	0.00
A-625908	2	2	40.126068	-75.331192	0.000	55.0	55.0	40.0	29.30	10.0	7.0	0.00
A-1537866	3	2	37.203964	-76.572983	0.000	34.9	0.0	86.0	30.20	10.0	0.0	0.00
A-63005	2	2	33.187656	-117.279549	0.010	60.1	0.0	75.0	30.02	10.0	9.2	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...
A-2571089	3	2	41.943270	-87.716210	0.364	79.0	0.0	77.0	29.87	10.0	9.2	0.02
A-2841848	2	2	46.061200	-95.830190	0.000	58.0	58.0	99.0	28.49	10.0	8.0	0.00
A-2395069	3	2	38.216862	-122.137840	0.000	80.1	0.0	18.0	29.83	10.0	15.0	0.00
A-202006	3	2	41.819302	-71.389595	0.010	44.1	37.4	68.0	29.29	10.0	13.8	0.00
A-1367723	3	3	28.175329	-82.391678	1.610	73.4	0.0	83.0	30.16	10.0	0.0	0.00

687496 rows × 204 columns

Command took 0.86 seconds -- by wang.jinc@northeastern.edu at 12/4/2020, 7:03:33 PM on 6110

The accuracy of the predictive model is around 67% to 70%, which means that this model is appropriate for forecasting, but it is better to do more improvement.

```
acc = metrics.accuracy_score(test_y, a)
print(acc)
```

```
0.6776126697464422
```

Command took 2.23 seconds -- by wang.jinc@northeastern.edu at 12/4/2020, 7:03:36 PM on 6110

Comments

The first difficulty that we encountered was when we were importing the data into Databricks. The error specified that it failed to find a data source which mainly because that the data was loaded in pyspark format which required us to convert the data frame from Spark to Pandas. In addition to this, Databricks sometimes shut down without any signal when we were running the multiple commands. We believe that this issue can be improved by having Databricks create a checklist to detect the performance or pop up any alerts which allows us to maximize the work efficiency. What's more, the challenge of data visualization is to present a clear and concise accident map. At first, we put all data into one map. However, the map is very hard to understand due to the overlap area, even though we use different colors and legend to classify these points as figure Y shows. Then, we try to adjust the transparency of the map by setting a lower alpha value. But it

does not work because the overlap part will become new colors and the reader may be confused about what these colors mean. Finally, we create five plots to demonstrate the trend of traffic accidents from 2016 to 2020. Audiences can easily understand the meaning of each map and find the trend of accident cases.

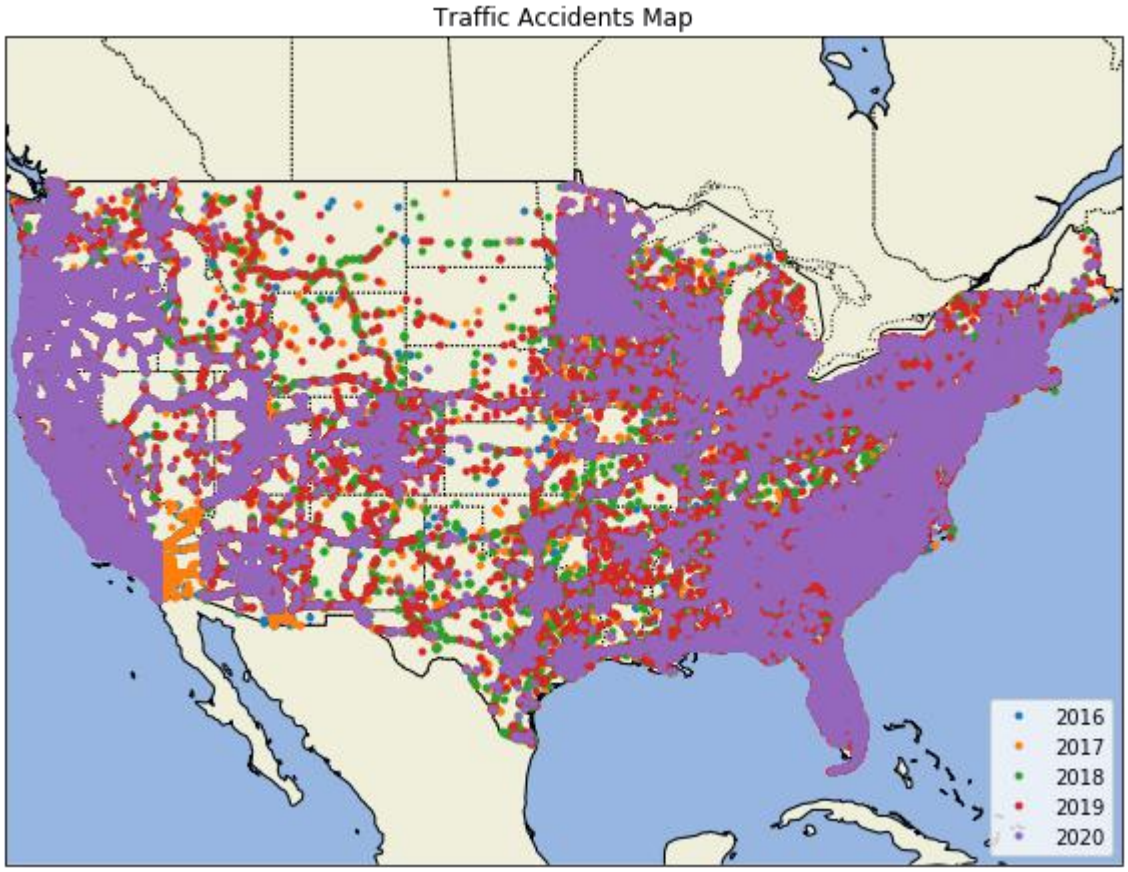


Figure Y: the traffic accidents map (old version)

For the model part, we fell hard to reshape the dataset. Because the dataset contains both numerical and categorical information, we must use 'get_dummies' to transform the categorical data. This will cause the volume of the data set to be too large to do any analysis. Here, the task asks us to determine which information should be included in the model.

Conclusions summary

We have a generally conclusion for the total four parts as the following:

- There were more cases during 8-12 compared to other months, excluding the data from 2020, guess there are worse weather conditions in the winter.

- Most accidents happened during the daytime, and there are two peaks on 7-8 and 16-17 when people are on commute between workplace and home.
- During 23 to 3 o'clock before dawn. Cases numbers are relatively at the bottom level as most people are in sleep
- Accidents are categorized into four groups (1-4) based on their Severity.1 being least and 4 being more severe. 67.7 % Accidents fall in the Severity class 2 followed by Severity class 3,4 and 1.
- Our assumptions generally are that bad weather could lead to more accidents. But here we can see that more accidents occur when the weather is clear. This may be because people drive more carefully when the weather is bad.
- Blocked, Accident, Rd Accident, due, Lane Blocked are the most frequent words in the description of traffic accident.
- CA has the most accidents happened. Since CA has highest population in the US states with logical assumption. Houston has the most accidents among all US cities. Being that Taxes is the 2nd biggest state in USA, this can lead to more drivers Speeding to reach their destination faster.
- For the map section, in the time scale, traffic accident cases are increasing by the dot's density. In terms of scope, these cases are spreading from the coastline to the central region.
- Among all the factors (over than 200 counts), only 15 of them are constant significant in each run of the model. They are 'Start_Lat', 'Start_Lng', '"Distance(mi)', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity (%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)',' Crossing', 'Junction', 'Traffic_Signal', 'Side_L', and 'Side_R'.
- The forecasting of severity of accidents is 67.76%, which means the model is appropriate. We think that in the reshaping process, we probably need to make some improvements to reaching the higher accuracy.

References

Nikam, Swapnil Kisan, "ANALYSIS OF US ACCIDENTS AND SOLUTIONS" (2020). Electronic Theses,Projects, and Dissertations. 979.

https://scholarworks.lib.csusb.edu/etd/979

Binu. (November, 2020). Road Accidents in US. Retrieved December 1, 2020 from https://www.kaggle.com/biphili/road-accidents-in-us.

William Roosevelt. (July, 2020). US Accidents data cleaning + eda. Retrieved December 1, 2020 from https://www.kaggle.com/williamrroosevelt/us-accidents-data-cleaning-eda.