US Traffic Accidents Analysis and Predictions

This dataset has been collected in real-time, using multiple Traffic APIs. Currently, it contains accident data that are collected from February 2016 to Dec 2021 for the Contiguous United States. US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the impact of COVID-19 on traffic behavior and accidents.

1. Data

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2021, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 2.8 million accident records in this dataset.

2. Method

There are three main types of recommenders used in practice today:

1. **Content-based filter:** Recommending future items to the user that have similar innate features with previously "liked" items. Basically, content-based relies on similarities between features of the items & needs good item profiles to function properly.

2. **Collaborative-based filter:** Recommending products based on a similar user that has already rated the product. Collaborative filtering relies on information from similar users, and it is important to have a large explicit user rating base (doesn't work well for new customer bases).

3. **Hybrid Method:** Leverages both content & collaborative based filtering. Typically, when a new user comes into the recommender, the content-based recommendation takes place. Then after interacting with the items a couple of times, the collaborative/ user-based recommendation system will be utilized.

   - The dataset is a structured and organized dataset with 47 metrics. It's collected continuously, including APIs that provide streaming traffic event data.
   - The parties that capture these data are US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.
   - Nearly around 2.9 million accidents were recorded from Feb 2016 to Dec 2021.
   - The dataset is about describing environmental types of metrics and geospatial information of the reported US car accidents. The user-based collaborative filtering system has been chosen for this analysis.

3. Data Cleaning

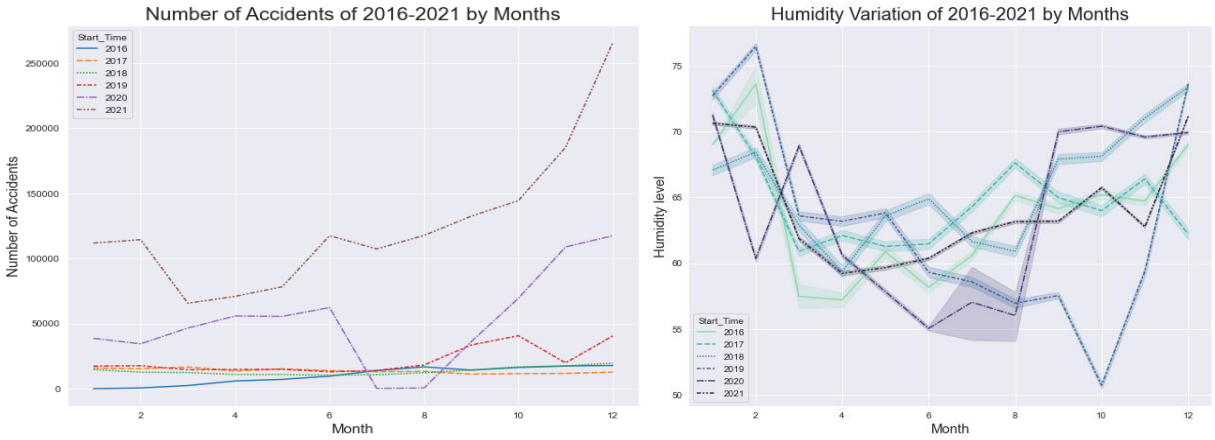
Data Cleaning Report

- The number column has been dropped as it has more than (50%) missing values (62%).
- The street column also has been dropped because it has so many unique features.
- The side columns had one N value that was removed. It also showed that most accidents toke place in the right side (more than 82%)!
- Miami has the largest number of accidents and it ranked number one. Miami also the number three populus city in the US, which is clearly justifies that, and sure we are going to get interesting information to help this in the EDA stage.
- Other cities like LA have the same growth rate population wise but also has lower accidents numbers and in comparison, with Miami, it considered too low. We need to check further behind the population information as it is not the main reason as it appears from the side research.
- Further looking into counties information and the State information is required in EDA.

4. EDA

EDA Report

- The map shows where car accidents occur more frequently. Indeed, if we locate to New York City, we don't have any data to show. If we locate to Miami and Los angeles, the number of car accidents is high.
- East and west coast have the highest number of car accidents, whereas the middle of the states are relatively more 'safe'.
- This map may not be able to tell much things about the accidents on the COVID-19 period.
- For future work, we might need more information about number of cases and COVID-19 regions in the form of heatmap.



5. Algorithms & Machine Learning

ML Notebook

Statistical Conclusion

- Distance: 25%/50% percentile are all 0; 75% percentile is 0.279 which is `quite` abnormal(too small).
- Precipitation: 25%/50%/75% percentile are all 0 which is also quite abnormal.
- Street: 1-5N seems to be a common street name that an accidents would happen. The name of the street could mean something or contain some special characters.
- Side: It seems that accidents usually happen on the right side of the road.
- City/County: Los Angeles is the place that traffic accidents are likely to happen.
- State: CA(California) is the state that traffic accidents are likely to happen.
- Wind_Direction: CALM is the most frequent wind direction.

- Weather_Condition: Fair is the most frequent weather condition.

Location Analysis Conclusion

- *Top 10 states: CA, FL, TX, NC, SC, NY, OR, VA, PA, IL*
- *Top 10 cities: Houston, Charlotte, Los Angeles, Miami, Dallas, Austin, Raleigh, Orlando, Sacamento, Baton Rouge.*
- *Most of the accidents happened at the right side of the road which is quite an interesting finding.*

Time Analysis Conclusion

- *The number of the accidents are increasing every year(2016~2020).*
- *It seems that there is an increase in the number of the accidents from July(7) to December(12). Fall and winter is reasonable to be more dangerous. However, the number of the accidents in January and February are so much lower than December, which is quite interesting.*
- *There is a drop in the number of accidents during the weekend.*
- *It seems that 7:00/8:00(Go to work) and 16:00/17:00(go back home) are the time where accidents happen during a day.*
- *The decrease of the accidents number at weekends mainly because the decrease at 7:00/8:00(Go to work) and 16:00/17:00(go back home). The peak of the we*

Environmental Analysis Conclusion

- *Freezing rain with windy, light blowing Snow and patches of fog with windy are the top 3 dangeous weather condition.*
- *It doesn't show that the wind direction have a significant influence on severity.*
- *Temperature(F): lower temperature -> higher Severity*
- *Humidity(%): higher humidity -> higher Severity*
- *Pressure(in): lower pressure -> higher Severity*
- *Visibility(mi): lower visibility -> higher Severity*
- *Wind_Speed(mph): higher wind speed -> higher Severity*
- *In summary, all the results fit with cold, chill, freezing weather condition. Eg. Freezing rain with windy, snow and so on.*
- *Accidents mostly happened at daytime.*

Infrastructure Analysis Conclusion

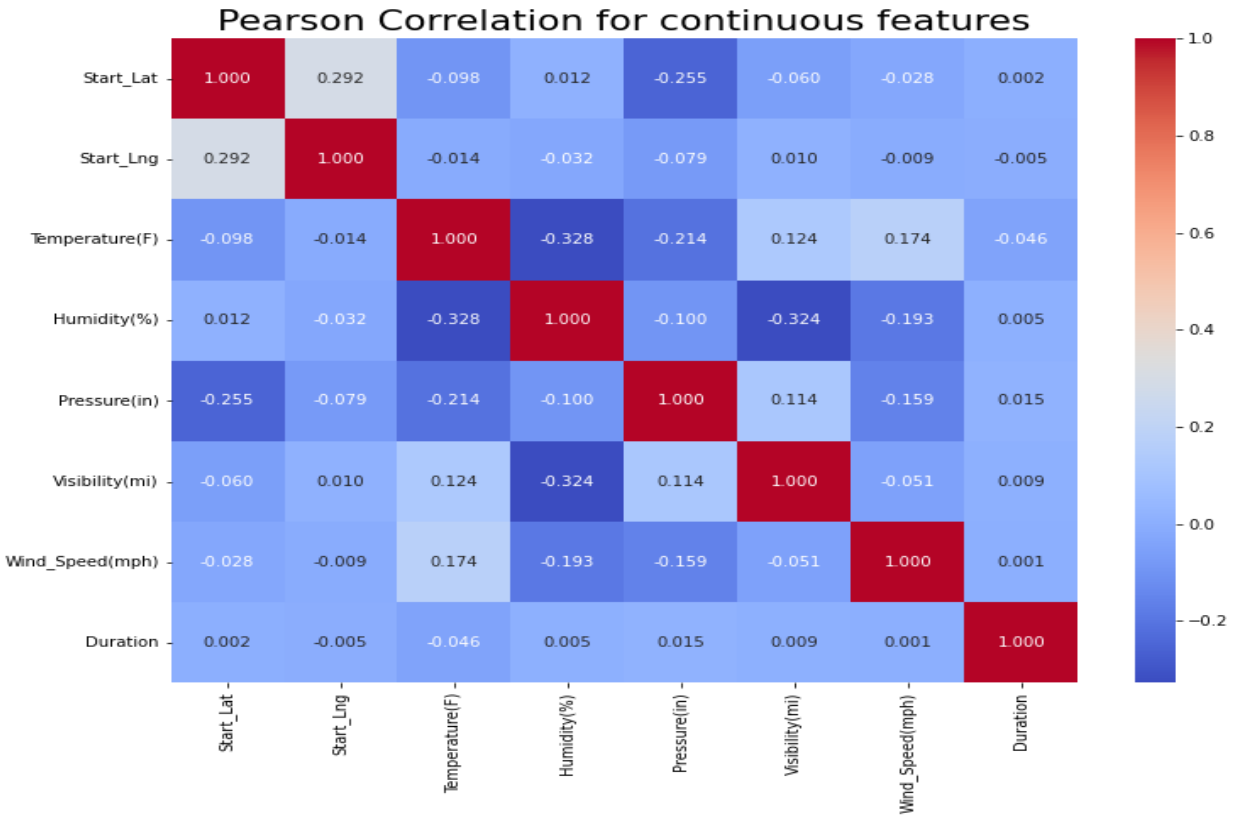
- *The accident usually happend at the palce where there are less traffic facilities.*

Correlation Analysis Conclusion

- *There is weak relationship between: Pressure and Temperature, Wind_Speed and Humidity, and Visibility and Humidity.*
- *There are Moderate relationship between: Bump and Traffic_Calming and Weather_thunder and Weather_Thunderstorm.*
- *There are Strong relationship between: Crossing and Traffic_Signal*

Modeling Conclusion

- *Best model to accidents End-time prediction: Gradient Boost Tree; {'gradientboostingregressorlearning_rate': 0.1, 'gradientboostingregressormax_depth': 2*



We found: the following from the expanded heatmap:

- Moderate relationship: Bump and Traffic_Calming and Weather_thunder and Weather_Thunderstorm.
- Strong relationship between Crossing and Traffic_Signal.
- At this point we can move to the modeling phase as we are ready for it.

6. Future Improvements

- In the future, I would love to spend more time creating a filtering system, wherein a climber could filter out the type, difficulty of climb, & country before receiving their top ten recommendation

- This recommendation system could also be improved by connecting to the 8a.nu website so that the user could input their actual online ID instead of just their user_id number

- Due to RAM constraints on google colab, I had to train a 65% sample of the original 6x dataset. Without resource limitations, I would love to train on the full dataset. Preliminary tests showed that the bigger the training size, the lower the RMSE. One test showed an increase in sample size could increase the RMSE by .03 (in contrast to the .005 improvement I received when increasing the coldstart threshold)

7. Credits

Thanks to Tony Peak and Juan Antonio- My mentors at Springboard