Predicting Accidents in US using Time Series Analysis

Introduction:

As per US Department of Transportation, there are over 5,891,000 vehicle crashes, on average, each year. Approximately 21% of these crashes - nearly 1,235,000 - are weather-related. On average, nearly 5,000 people are killed and over 418,000 people are injured in weather-related crashes each year. (Source: Ten-year averages from 2007 to 2016 analyzed by Booz Allen Hamilton, based on NHTSA data).

Weather acts through visibility impairments, precipitation, high winds, and temperature extremes to affect driver capabilities, vehicle performance (i.e., traction, stability, and maneuverability), pavement friction, roadway infrastructure, crash risk, traffic flow, and agency productivity.

Climate change can influence road environment and driving behavior, which in turn affect the risk of road traffic accidents. The climatic factors affecting the traffic accident include temperature, precipitation, and wind. Temperature has an influence on the risk level of road traffic accidents, especially the extreme low temperature conditions and the hot weather. In stormy weather, when the gust speed is higher than a certain threshold, the probability of traffic accidents increases accordingly. Strong winds increase the frequency of rollovers, sideslip, and spin, especially rollovers.

There have been numerous studies to understand accidents due to unsafe driving conditions, road conditions or drunk driving patterns. However not many studies are dedicated to day-to-day weather change impacts on accidents. This study tries to predict the number of future accidents using time series data of historical accidents and weather reports.

Data:

The study uses accident data* from 2016 to 2021 sourced from Kaggle. It contains 2845342 rows and 47 columns with information on the accidents across the country, date and timestamp of accidents, their geospatial mapping, weather data from the nearest airport station at the time of accident and traffic/road condition (like pedestrian crossing, stop junction, railway crossing etc.)

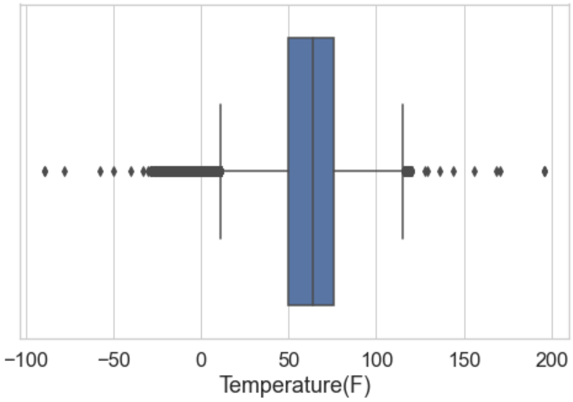
While this study focuses primarily on weather, the research can be expanded in future to examine the following.

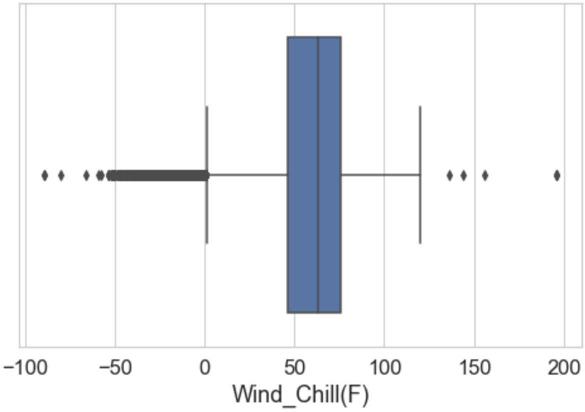
- Role of road conditions combined with weather conditions on the number of accidents.
- Weather impact on the Severity of accident using a Time series Classifier.
- Forecasting accidents using geospatial mapping of accidents.
- Effect from Covid and work from home jobs on the number and severity of accidents.

Data Wrangling:

We begin by analyzing the missing values. We notice that a few columns may not aid in the current study but may potentially be useful for future study. Weather conditions are represented by the continuous variables of Temperature, Wind Chill, Visibility, Pressure, Humidity, Wind Speed and Precipitation. We notice almost 20% of the data in Precipitation is missing.

Temperature and Windchill have some extreme outliers.

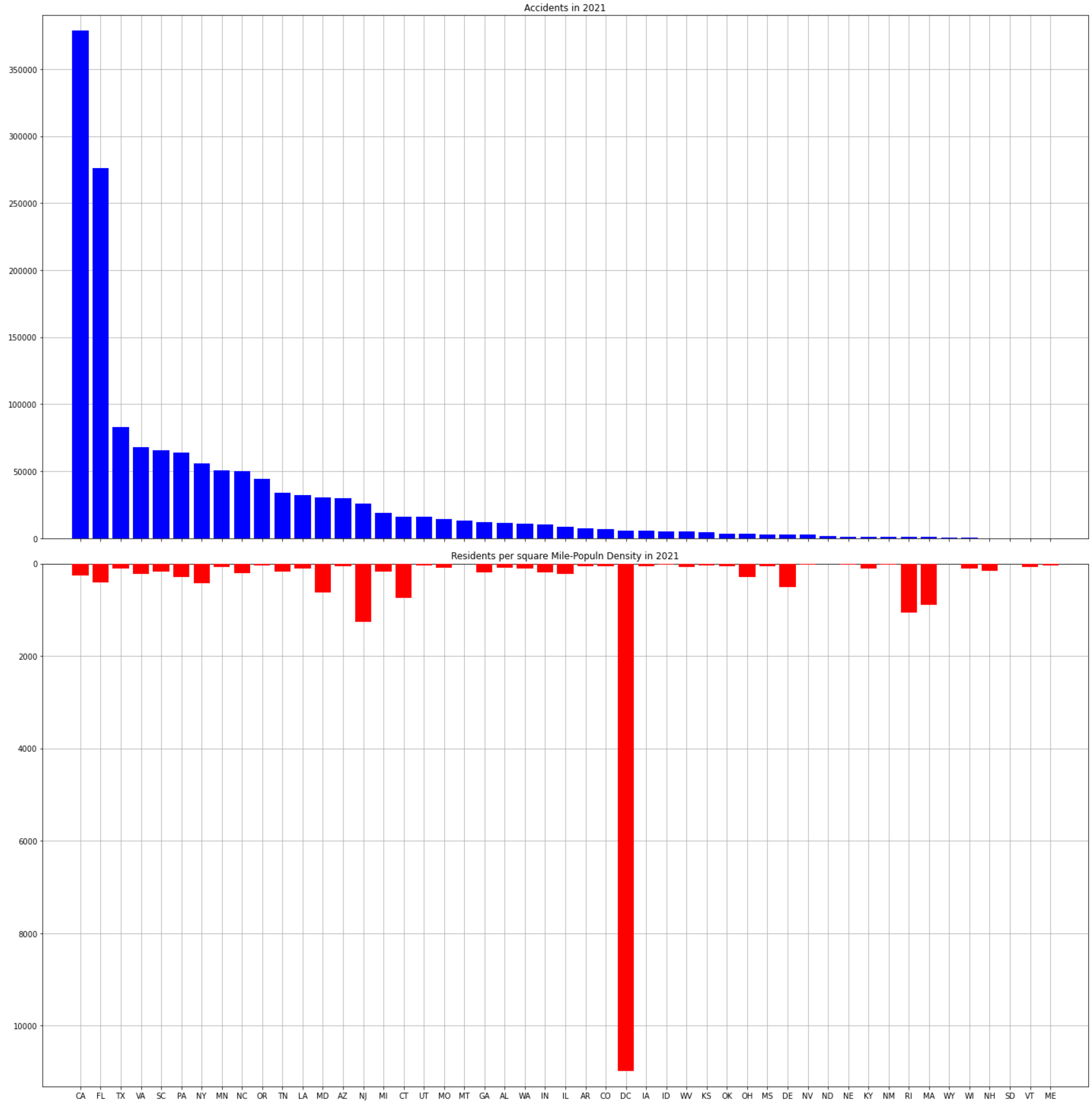




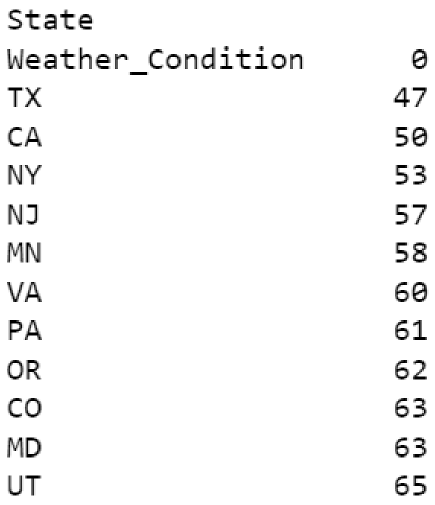
A few temperature values over 150F seemed impossible. To deal with such outliers, we engage in some historical weather research. The highest temperature on record belonged to California's Death Valley which, in 1913, reached a temperature of 134 degrees Fahrenheit. Further on checking weather data for TX, FL, AL, AZ, NY, and IL, none of the locations have ever recorded anything over 120F. TX had recoded a maximum of 120F. We delete these outliers to avoid bias in the study. We do a similar study for the lowest temperatures and delete a few rows. We also check for duplicates and verify the data types of different columns.

Exploratory Data Analysis:

While California recorded the highest number of total accidents, followed by Florida, these two states featured amongst the least densely populated states. Contrasting to popular belief that more traffic density causes more accidents, this data indicated that population density has no effect on the total number of accidents.



Since the scope of this study is to analyze the effect of weather, we reduce the sample set. But how do we reduce it? The criteria used for reduction is to include diverse weather conditions in the representative sample. Weather Conditions like snow, fleeting, blowing wind etc. are described in a column.

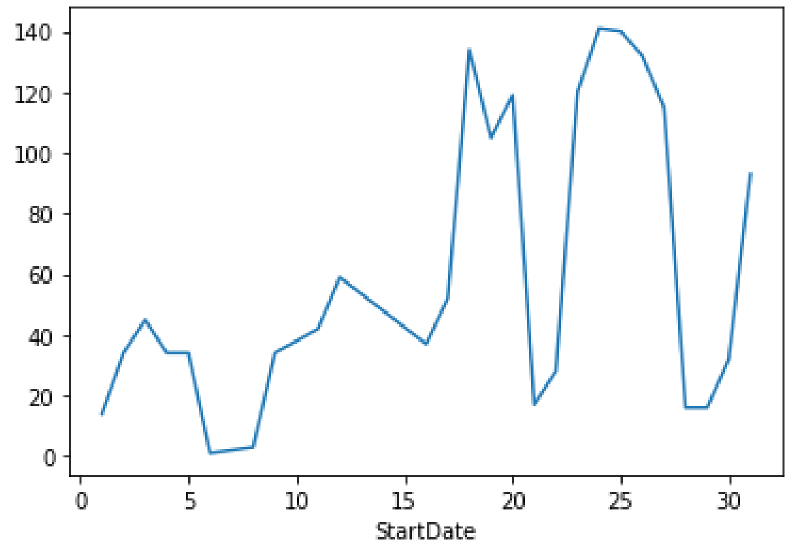


```
State
Weather_Condition      0
TX                    47
CA                    50
NY                    53
NJ                    57
MN                    58
VA                    60
PA                    61
OR                    62
CO                    63
MD                    63
UT                    65
```

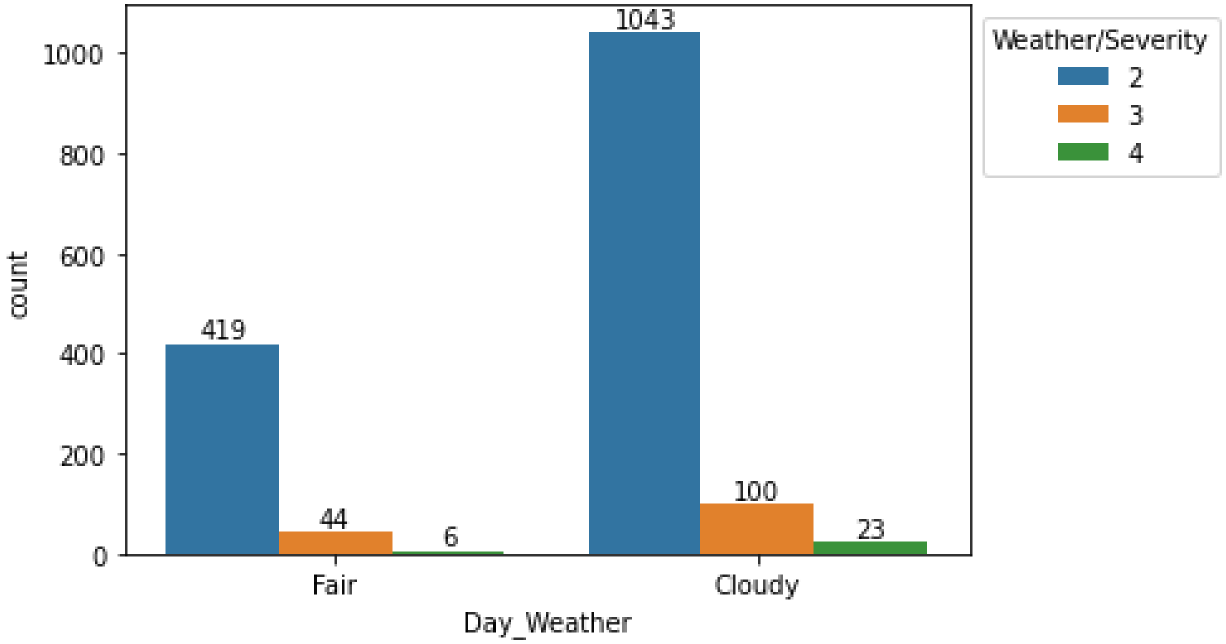
We map each weather condition for all the different states to check the ones with the least null values. This table shows the states which records least null values for weather conditions. These are also the ones with diverse weather conditions. CA not just records high number of accidents; it also records a variety of unique weather conditions. We subset the dataset to just the state of CA.

We narrow down the unique values in Weather Conditions. For instance, fair weather and clear weather may represent normality.

We repeat the earlier procedure of checking for diverse weather conditions and highest number of accidents to further reduce the sample set. Now we select county Los Angeles as it records a high number of accidents and diverse weather conditions. We randomly select a month to check patterns.

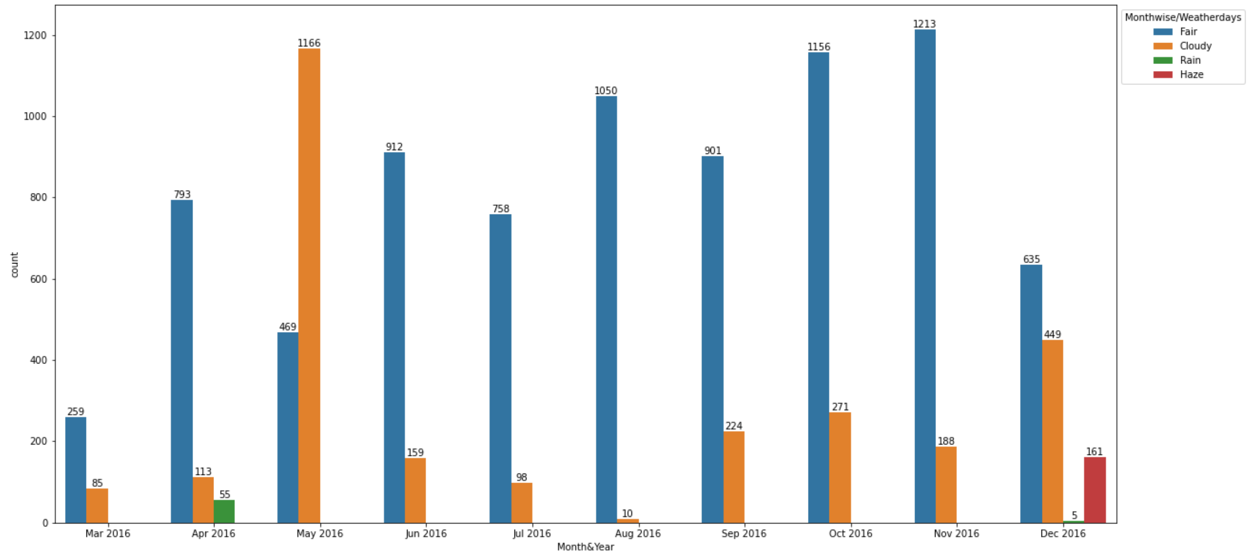


We notice that number of accidents peak around the mid of the month with frequent highs and lows till month end.

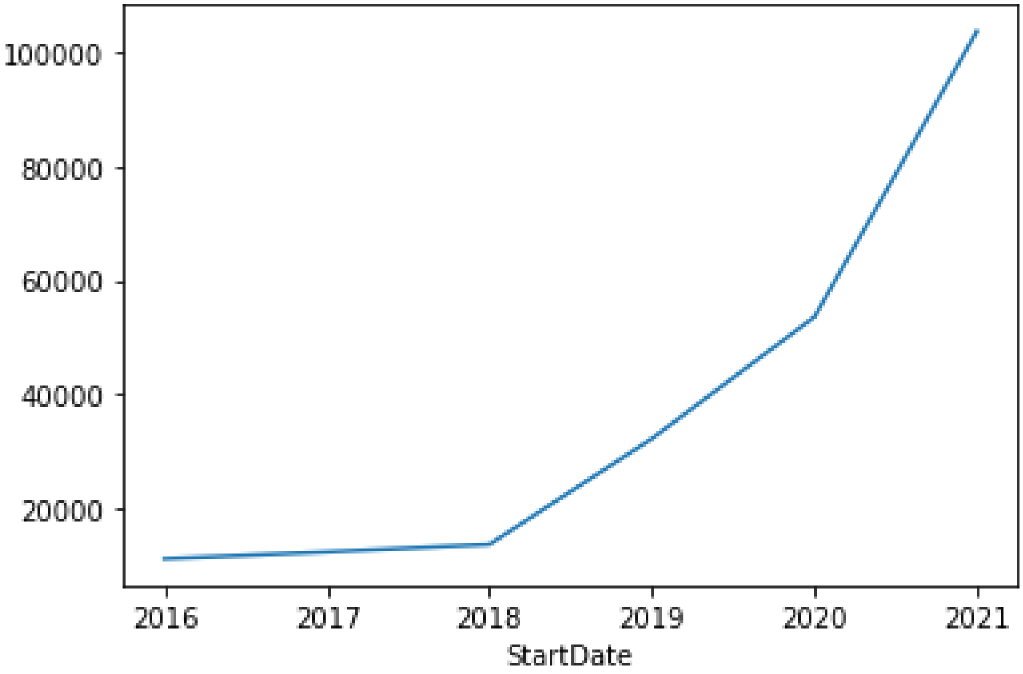


A cloudy day in LA county records more accidents than a fair-weather day, clearly indicating that weather impacts accidents.

A deeper look into all the accidents in year 2016 in LA indicates a peak only in the month of May. Except for May 2016, every other month has higher number of Fair-weather days. However earlier we did see that May 2016 had recorded a peak in accidents, which again indicates to a relation between weather and accidents.



Comparing the total number of accidents for all years, we see an upward trend in the number of accidents in 2018-2021.



Preprocessing and Training:

We prepare the data for Time series split by setting Start date as index and sorting it. We find the total number of accidents each day for LA county and average out the continuous variables for weather conditions. But we are still talking of

16758 Zip codes and 16 Airport codes. We apply our earlier reduction technique to zero in on one Airport code.

Our target variable is to predict the number of accidents for a particular day using past data- Total Accidents per day. Here we will use the number of accidents that have occurred to guide us. This will be achieved by creating a lag of the target variable - "Total Accidents". We lag it by 30 days. For instance, to predict for May 2018, we could use data of April 2018. Hence, we lag the column total accidents by 30 days. Similarly, when we have weather conditions, we will have to lead them by one day as we will not know the weather for next day in advance. Hence, we will use the weather conditions of the previous day and the lagged features of the target variable to forecast the accidents of a given day.

We create lag features before splitting so that the model is trained to forecast on previous data, and it will then be able to do the same for test data.

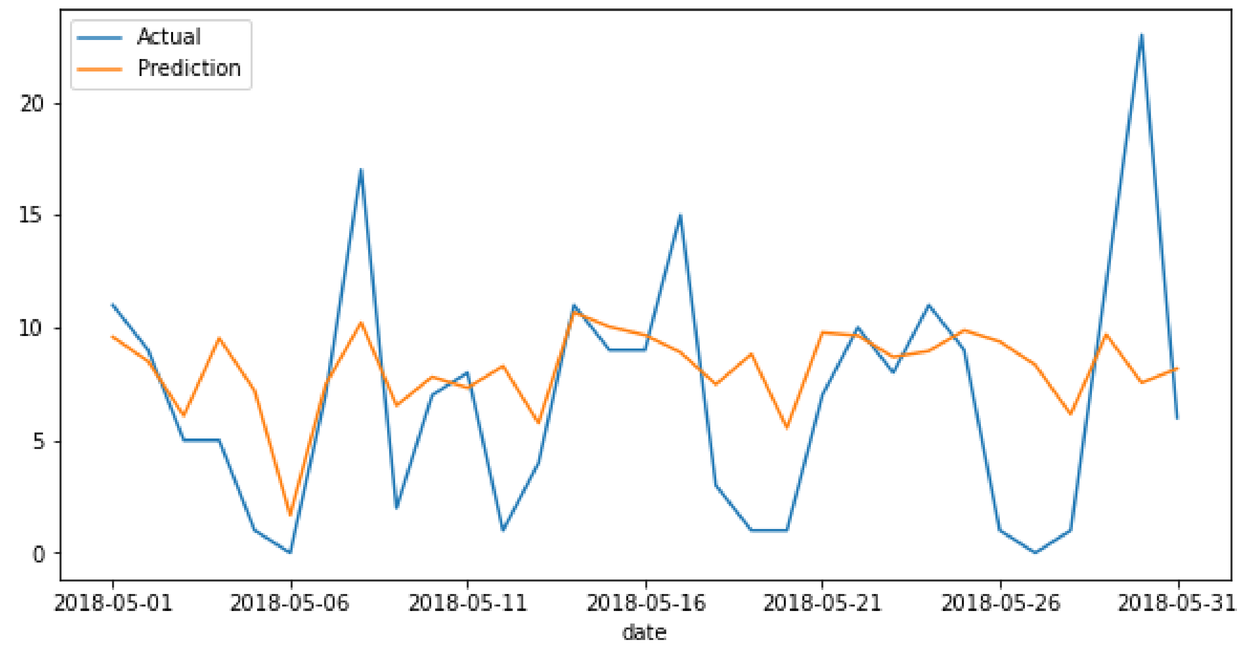
We split the dataset into train which contains May 2016 to April 2018 data. This will be two years of data. The test set will be for the month of May 2018. After splitting, missing values of Wind Speed in train and test are imputed using Next Observation Carried Backward, where the next non-missing values are copied and replaced with the previous missing values.

We now have a multivariate time series forecasting dataset ready for modeling.

Modeling and Conclusion:

We apply three ensemble models – Random Forest Regressor, XGBoost and LightGBM with GridSearch cross validation.

The Random Forest model scores an impressive RMSE of 4.9 and MAE of 3.5.



RMSE has the benefit of penalizing large errors more so can be more appropriate in some cases, for example, if being off by 10 is more than twice as bad as being off by 5. But if being off by 10 is just twice as bad as being off by 5, then MAE is more appropriate.

Since we are predicting accidents, a significantly large error could be more disastrous. Imagine a day when the actual accidents are to be 50 but the model predicts just 30. A large error of this kind could result in the county being unprepared to handle emergencies. Since RMSE would penalize such large errors, in this case, we could opt for a model with lower RMSE. We now have a model that could predict/forecast accidents into the future using past data and other variables.