USA Car Accidents Severity Prediction

IBM Data Science Capstone Project

Nikita Kohli

Introduction

- Motivation:
  The economic and societal impact of traffic accidents cost U.S. citizens hundreds of billions of dollars every year. And a large part of losses is caused by a small number of serious accidents. Reducing traffic accidents, especially serious accidents, is nevertheless always an important challenge. The proactive approach, one of the two main approaches for dealing with traffic safety problems, focuses on preventing potential unsafe road conditions from occurring in the first place. For the effective implementation of this approach, accident prediction and severity prediction are critical. If we can identify the patterns of how these serious accidents happen and the key factors, we might be able to implement well-informed actions and better allocate financial and human resources.
- Objective:
  The first objective of this project is to recognize key factors affecting the accident severity. The second one is to develop a model that can accurately predict accident severity. To be specific, for a given accident, without any detailed information about itself, like driver attributes or vehicle type, this model is supposed to be able to predict the likelihood of this accident being a severe one. The accident could be the one that just happened and still lack of detailed information, or a potential one predicted by other models. Therefore, with the sophisticated real-time traffic accident prediction solution developed by the creators of the same dataset used in this project, this model might be able to further predict severe accidents in real-time.

Dataset Overview

US-Accident dataset is a countrywide car accident dataset, which covers 49 states of the United States. It contains 3 million cases of traffic accidents that took place from February 2016 to June 2020, using two 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. In this project, only the data of accidents that happened after February 2016 and were reported by MapQuest was finally used in exploration analysis and modeling so that irrelevant factors can be eliminated to the greatest extent. Details about features in the dataset:

- Traffic Attributes (12):
- ID: This is a unique identifier of the accident record.
- Source: Indicates source of the accident report (i.e. the API which reported the accident.).
- TMC: A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
- Severity: Shows the severity of the accident, a 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).
- Start_Time: Shows start time of the accident in local time zone.
- End_Time: Shows end time of the accident in local time zone.
- Start_Lat: Shows latitude in GPS coordinate of the start point.
- Start_Lng: Shows longitude in GPS coordinate of the start point.
- End_Lat: Shows latitude in GPS coordinate of the end point.
- End_Lng: Shows longitude in GPS coordinate of the end point.
- Distance(mi): The length of the road extent affected by the accident.
- Description: Shows natural language description of the accident.
- Address Attributes (9):
- Number: Shows the street number in address field.

- Street: Shows the street name in address field.
- Side: Shows the relative side of the street (Right/Left) in address field.
- City: Shows the city in address field.
- County: Shows the county in address field.
- State: Shows the state in address field.
- Zipcode: Shows the zipcode in address field.
- Country: Shows the country in address field.
- Timezone: Shows timezone based on the location of the accident (eastern, central, etc.).

- Weather Attributes (11):
- Airport_Code: Denotes an airport-based weather station which is the closest one to location of the accident.
- Weather_Timestamp: Shows the time-stamp of weather observation record (in local time).
- Temperature(F): Shows the temperature (in Fahrenheit).
- Wind_Chill(F): Shows the wind chill (in Fahrenheit).
- Humidity(%): Shows the humidity (in percentage).
- Pressure(in): Shows the air pressure (in inches).
- Visibility(mi): Shows visibility (in miles).
- Wind_Direction: Shows wind direction.
- Wind_Speed(mph): Shows wind speed (in miles per hour).
- Precipitation(in): Shows precipitation amount in inches, if there is any.
- Weather_Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.).

- POI Attributes (13):
- Amenity: A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location.
- Bump: A POI annotation which indicates presence of speed bump or hump in a nearby location.
- Crossing: A POI annotation which indicates presence of crossing in a nearby location.

  - Give_Way: A POI annotation which indicates presence of give_way sign in a nearby location.
  - Junction: A POI annotation which indicates presence of junction in a nearby location.
  - No_Exit: A POI annotation which indicates presence of no_exit sign in a nearby location.
  - Railway: A POI annotation which indicates presence of railway in a nearby location.
  - Roundabout: A POI annotation which indicates presence of roundabout in a nearby location.
  - Station: A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location.
  - Stop: A POI annotation which indicates presence of stop sign in a nearby location.
  - Traffic_Calming: A POI annotation which indicates presence of traffic_calming means in a nearby location.
  - Traffic_Signal: A POI annotation which indicates presence of traffic_signal in a nearby location.
  - Turning_Loop: A POI annotation which indicates presence of turning_loop in a nearby location.
- Period-of-Day (4):
  - Sunrise_Sunset: Shows the period of day (i.e. day or night) based on sunrise/sunset.
  - Civil_Twilight: Shows the period of day (i.e. day or night) based on civil twilight.
  - Nautical_Twilight: Shows the period of day (i.e. day or night) based on nautical twilight.
  - Astronomical_Twilight: Shows the period of day (i.e. day or night) based on astronomical twilight.

Due to the limit of computer capacity, I am focusing on Montgomery County in the State of Pennsylvania. I will only select a few features I believe are more relevant to severity. Categorical data will be treated

with Pandas get_dummies method. Rows with missing values will be dropped.
Data source: https://www.kaggle.com/sobhanmoosavi/us-accidents

Methodology

- <u>Severity Level</u>

  As a nation gets on wheels, driving around is part of the life. Imagine you just dressed yourself up and on the way to your first date, what's the last thing you want to happen? Or you just took off a family road trip to a long-dreamed national park, what's absolutely the thing you want to avoid?
  The good thing is that, although it varies state by state, most accidents are light accidents with severity level 2, with the most severe one (level 4) the least (Figure 1). The time impact is not the mere thing; the emotional impact can last much longer, or even ruin your whole day or entire vacation.

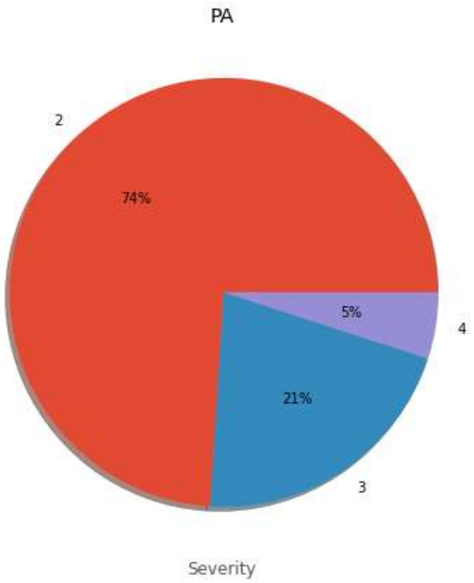

Fig.1. Accident severity for the states of PA

- <u>Exploratory Data Analysis</u>

  To avoid car accidents, it is meaningful to see when, where and under what weather conditions did most accidents happen.

- WHEN

  According to the results, most accidents happened during the daytime (Figure 2). There are more accidents on weekdays than weekends (Figure 3). On weekdays, rush hours are most dangerous times while on weekends; early afternoon is more dangerous than other time (Figure 4). Based on this information, you may plan your travel better if possible, or pay extra attention while driving during highly risky time.

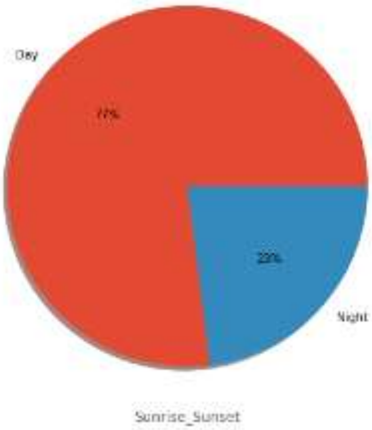

Fig.2. Percentage of accidents during daytime versus nighttime

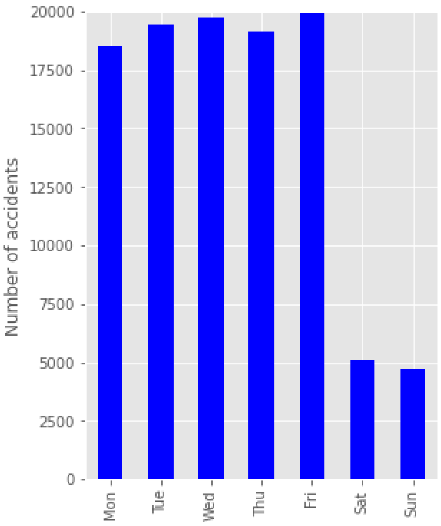

Fig.3. Average number of accidents per day for each weekday/weekend

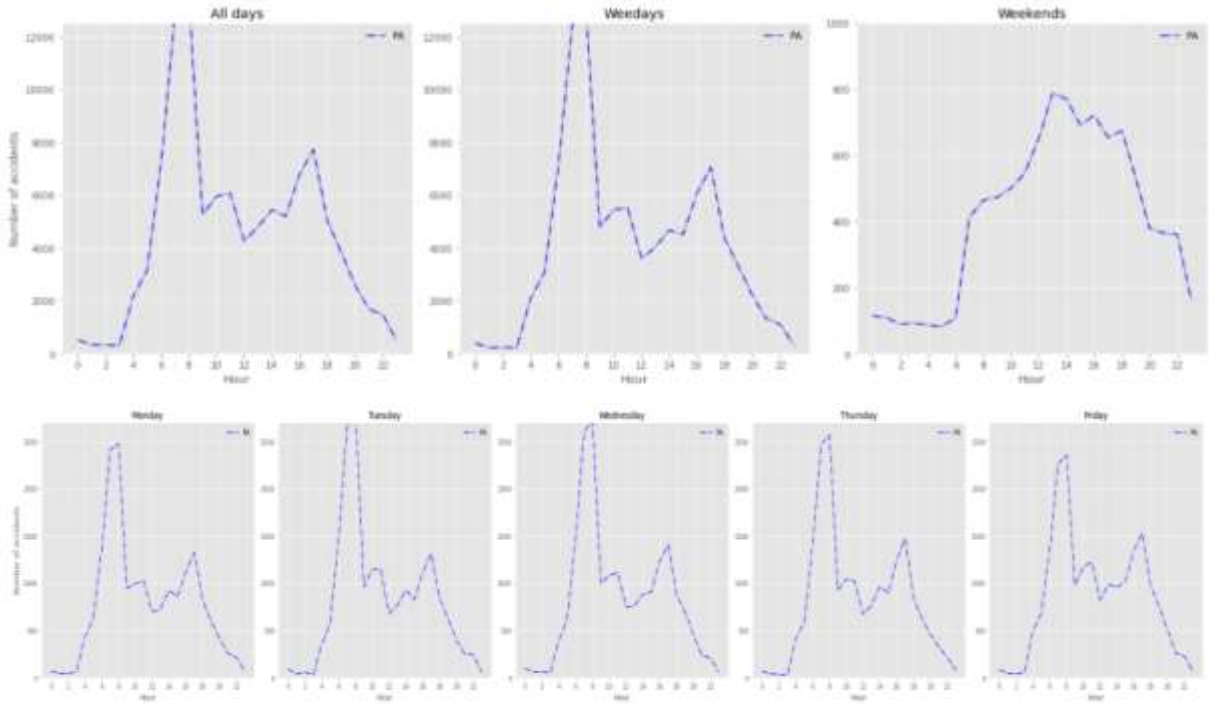

Fig.4. Average number of accidents per severity level per day for each hour on weekday/weekend

- WHERE
  While most accidents happened on the right side of the street, still quite a significant portion (~ 30%) of accidents occurred on the left side, especially in the states of PA (Figure 5). While Montgomery County in PA has the most accidents (Figure 7), Philadelphia, Lancaster, and Pittsburg are the most dangerous cities in PA (Figure 8).

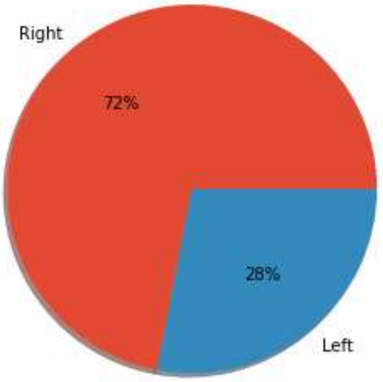

Fig.5. Street sides for accidents

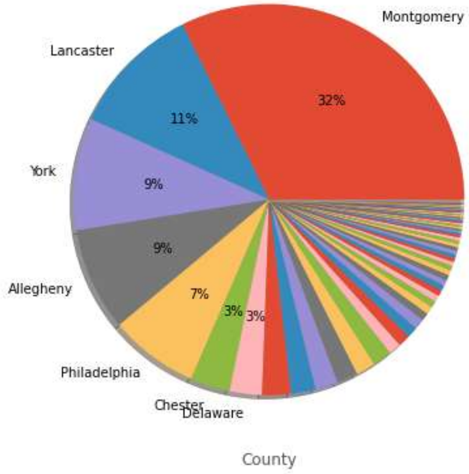

Fig.6. Most dangerous counties

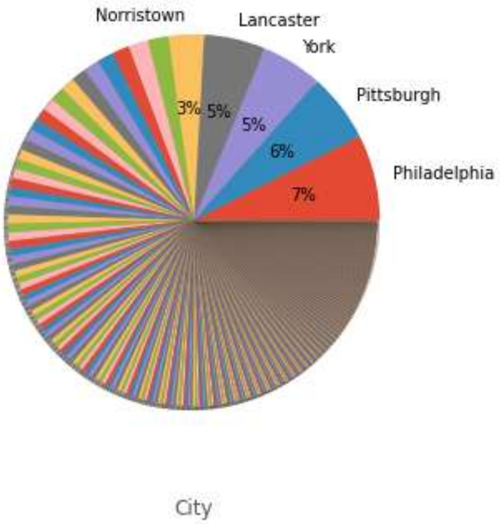

Fig.7. Most dangerous cities

Take a closer look at each accident; most accidents happened at traffic signals, followed by Junctions and Crossings (Figure 8). In each of these locations, most accidents are still light accidents with severity level 2 (Figure 9).

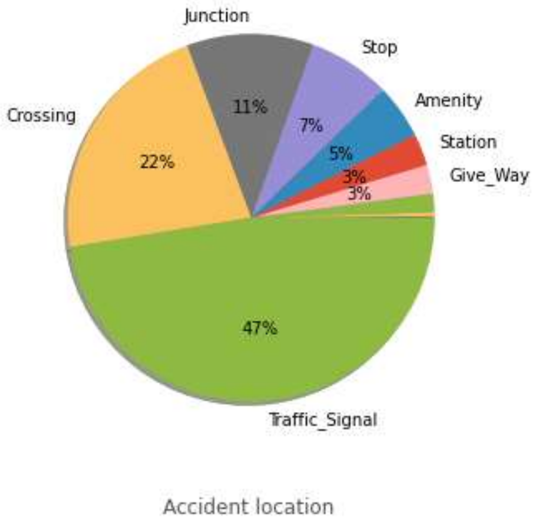

Fig.8. where did most accidents occur

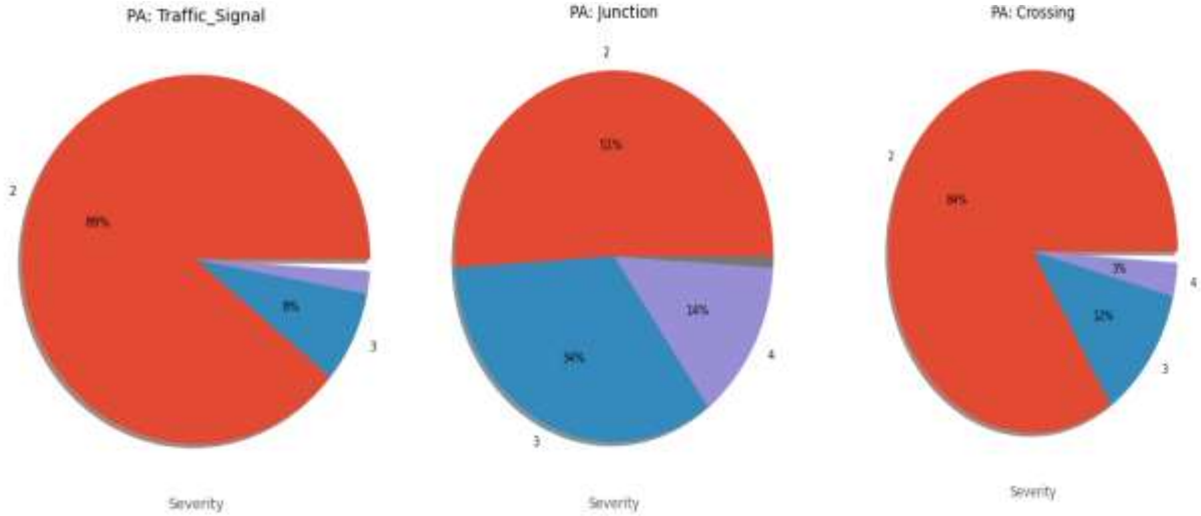

Fig.9. Accident severity levels for each location

Based on this information, you know where you need to avoid if possible, otherwise, stop texting or messaging but rather focusing on the road!

- WEATHER
  Based on this data set, most accidents occurred under clear weather condition. It may be due to the fact that percentage of clear days being the highest, or drivers tend to pay less attention to the road under good weather conditions (Figure

10). While under bad weather conditions, people may be more cautious while driving, thus less accidents. While for each weather condition, most accidents have severity level 2 (Figure 11), for each severity category, most accidents occurred under clear weather condition (Figure 12).

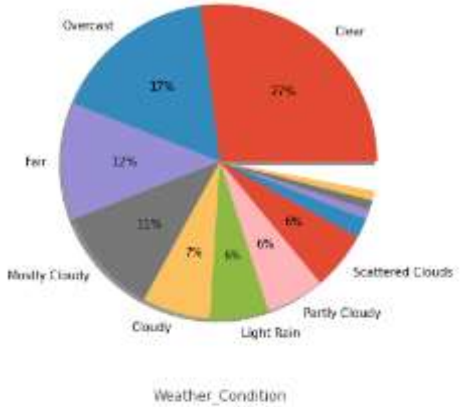

Fig.10. Percentage of accidents under various weather conditions

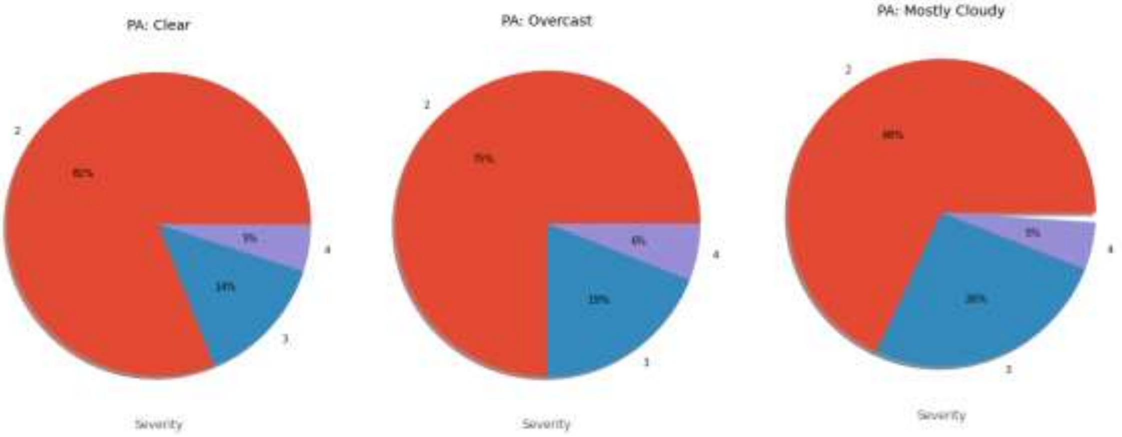

Fig.11. Accident severity under top three weather conditions

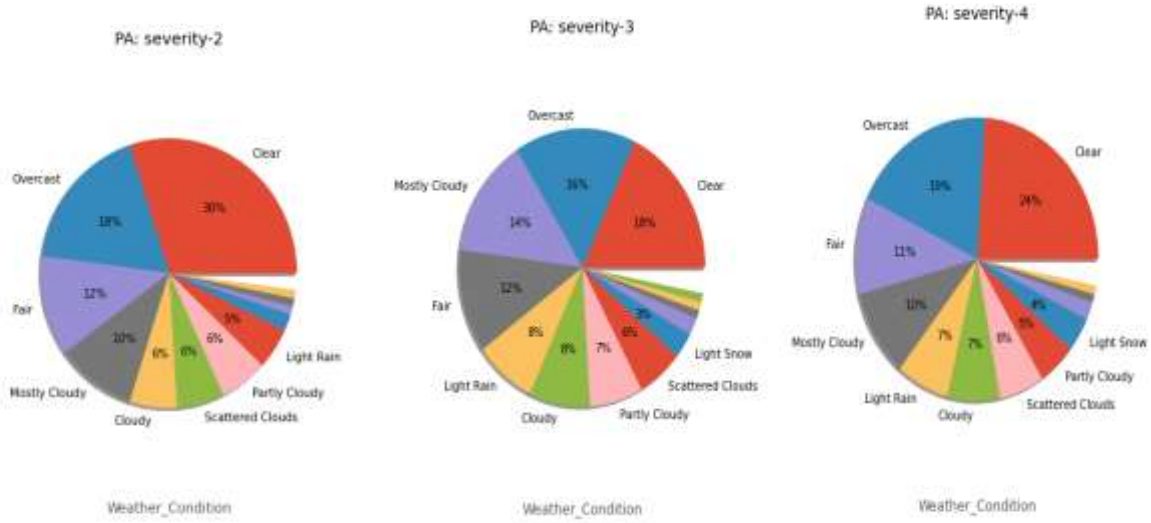

Fig.12. Weather conditions for accidents with each severity

Now you know you should not relax and diverge your attention to other unrelated affairs while driving under the sun.

Accidents are everywhere, we should do everything we can to avoid it. It is just for your own, but also for your families and loved ones. Especially for those deadly accidents, there is no regret.

Nevertheless, if you already did your best but still cannot avoid it. Can we try to reduce the severity, thus impact, of the accident? Can we predict the severity with machine learning algorithms?

- <u>Severity Prediction</u>

  Applying Machine Learning Algorithms to Predict the Severity of Car Accidents

  - DATA PREPROCESSING

    Besides no missing value is allowed, most machine learning algorithms work only with numerical data. For that, records with missing values were dropped from the calculation. Some outliers, especially with extremely short or long time to clear the accident, were processed and replaced with median values. As it doesn't make sense to take 0 minute or many years to clear an accident.

- PREDICTING ACCIDENT SEVERITY WITH VARIOUS MACHINE LEARNING ALGORITHMS

  In this study, four classification machine learning algorithms were evaluated. These are Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees and Random Forest. As you can see in Figure 13, Random Forest algorithm is always the winner and KNN is the last on the list in terms of accuracy. The accuracy is quite high for Montgomery county in PA.

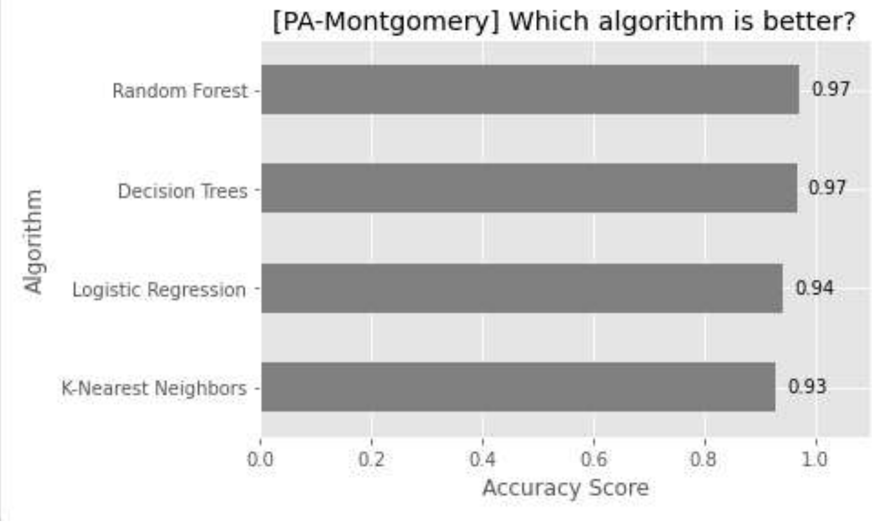

Fig.13. Accuracy of predicting accident severity with various machine learning algorithms for Montgomery county in PA

- MOST IMPORTANT FEATURES

  Shown below are the top 10 features affecting the accuracy of predicting accident severity for the data from Montgomery county in PA. Prediction accuracy can be further improved by removing less important or irrelevant features.

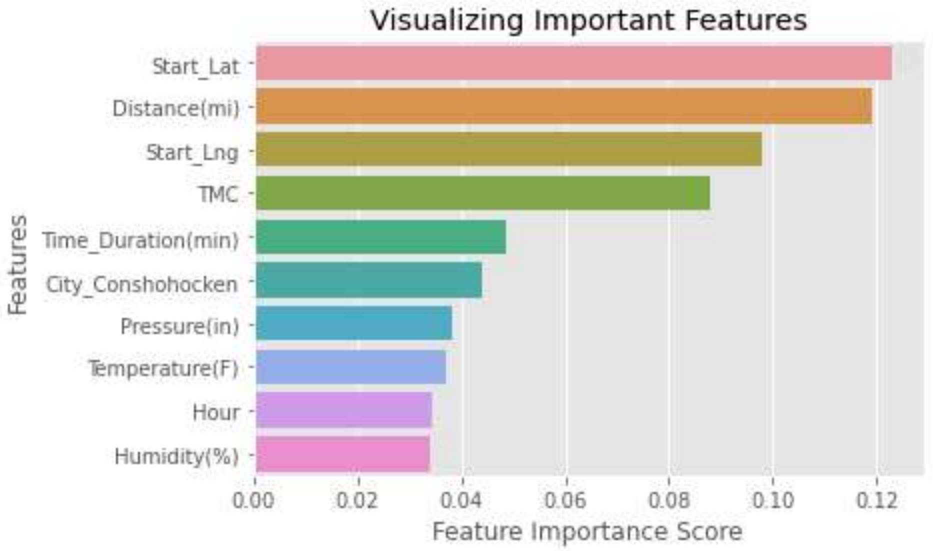

Fig.14. Top 10 features affecting the prediction accuracy for Random Forest algorithm for Montgomery county in PA

Result

This study indicates that there are patterns of when, where and under what weather conditions did most accidents occurred. The severity of each accident can be predicted quite accurately with various classification machine learning algorithms.

Future Work

1. Incorporate this model in a real-time accident risk prediction model or develop a new real-time severe accident risk prediction on grid cells.
2. Detailed relations between some key factors and accident severity can be further studied.
3. Policy implications of this project can be explored.

Conclusion

- Country-wide accident severity can be accurately predicted with limited data attributes (location, time, and weather).
- Spatial patterns are the most useful features. For small areas like **street**, severe accidents are more likely to happen at places

having more accidents while for larger areas like **city**, at places having less accident.

- An accident is much less likely to be severe if it happens near **traffic signal** while more likely if near **junction**.
- Weather features like **pressure**, **temperature**, **humidity**, and **wind speed** are also very important.