Traffic Accident Analysis and Predictive Indicators in Denver, Colorado

By: D. Nguyen

Dated: September 10, 2020

Introduction

This is a report on predicting the probability of the severity of a traffic accident based on various factors particularly in the City of Denver, Colorado. The predictive analysis can be used by the government and its citizens to mitigate or avoid such accidents, and to improve circumstances where possible. Lowering the impact or avoiding traffic accidents will decrease the amount of property damage and human injury, and can also reduce traffic jams within the community.

Data

Data from collisions that occurred within the City and County of Denver are tabulated and provided by the Denver Police Department. The data includes periods from January 1, 2013 through September 1, 2020 with total observations of 181,837 collisions reported by the police. The severity of a collision is indicated by the number of individuals reported as 'Seriously Injured' or as a 'Fatality'. Details about each collision includes location information, collision descriptions, types of transportation involved, environmental conditions such as weather and road conditions, and any human negligence such as speeding or driving under the influence. This dataset is used for exploratory analysis and to determine correlations with the severity of the traffic accidents.

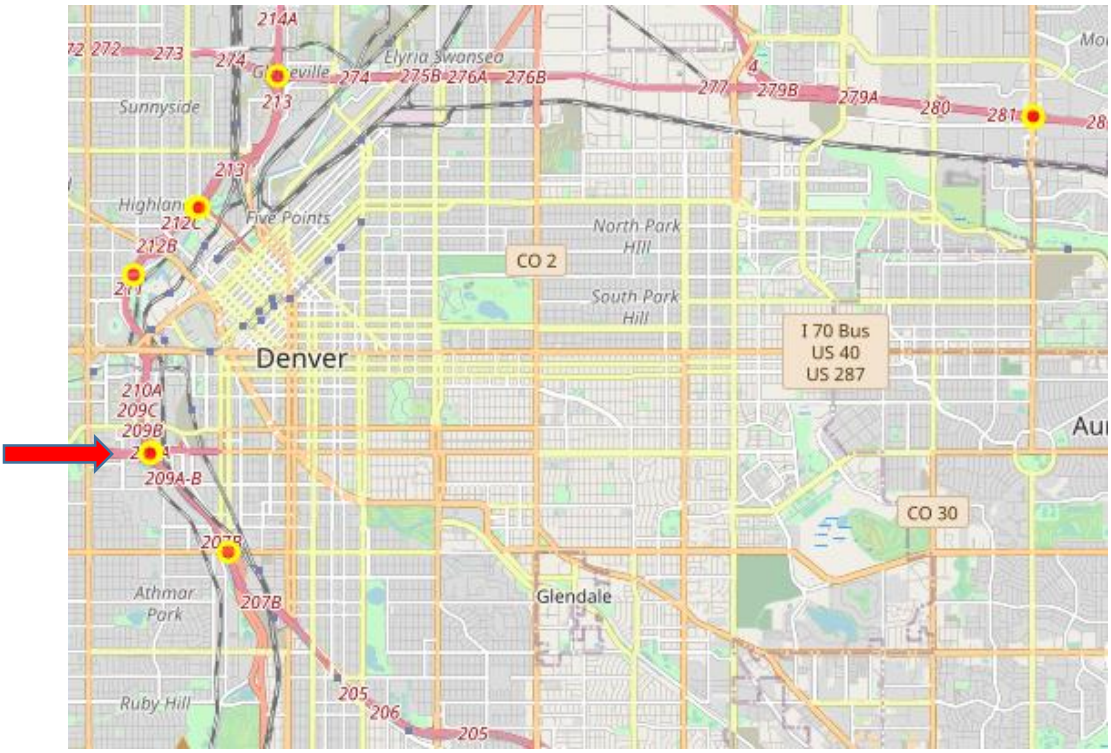
Note that the Denver Police Department completes a State of Colorado Accident Report if there is $1,000 or greater in damage, an injury or fatality, or drug/alcohol involvement. Motor vehicle crashes reported directly to the State of Colorado, even if they occurred within the City and County of Denver, are not included in this dataset.

Discussion

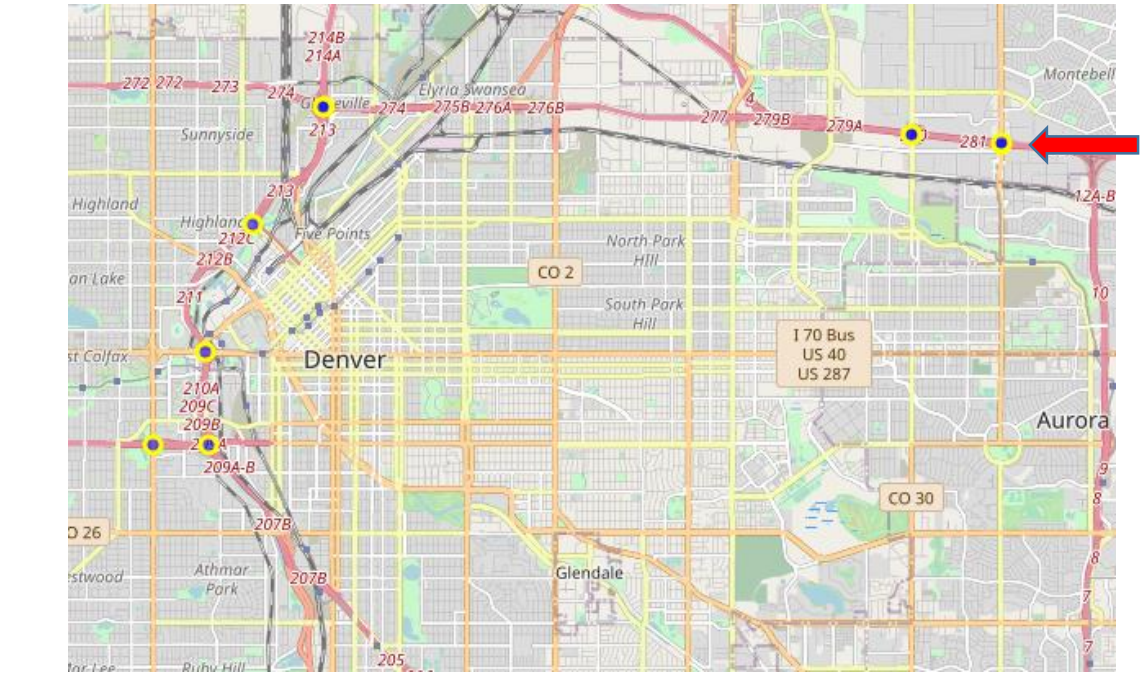
Exploratory analysis is performed on the following segments: By location, time series, environmental condition, and the human factor.

Location

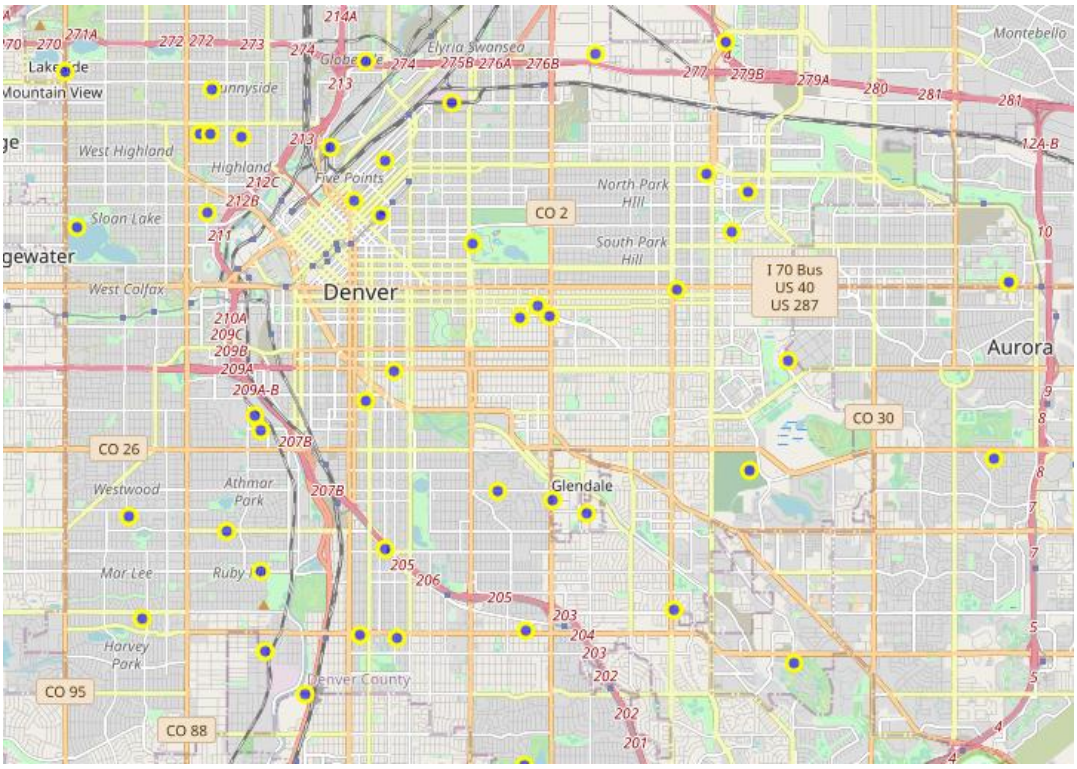
Mapped below are the top locations with the greatest number of traffic accidents which shows that incidents most frequently occur on highways. Of the plots presented, the intersection of HWY Interstate 25 and 6th Avenue has the highest count of accidents during the last five years.



Top locations with serious injuries including fatalities are also along major highways with Interstate 70 and Peoria Street being the most hazardous. Based on the dataset, 2.26% of total traffic accidents had resulted in serious bodily injury or fatality.



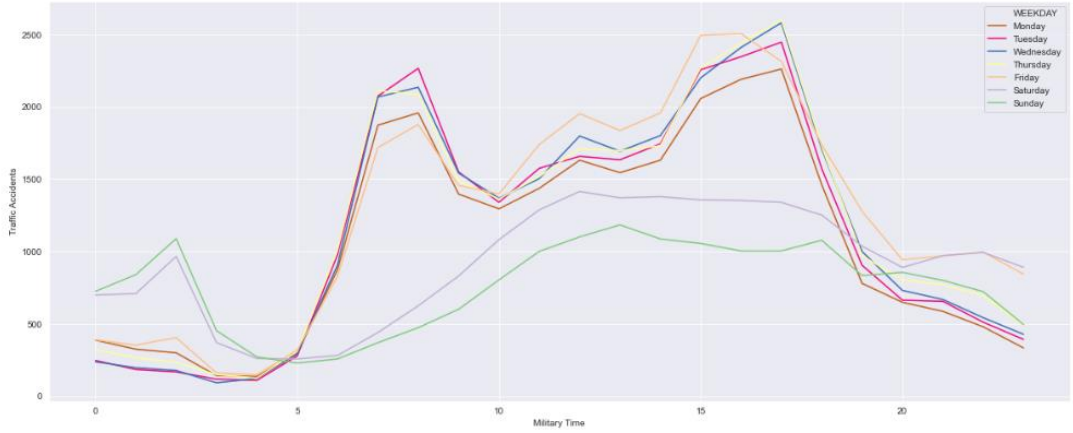
A random sample of locations with traffic accidents that do not have serious injury or fatalities are shown as located outside of highways. This indicates that location should be included as a predictive feature in severity of a traffic accident.



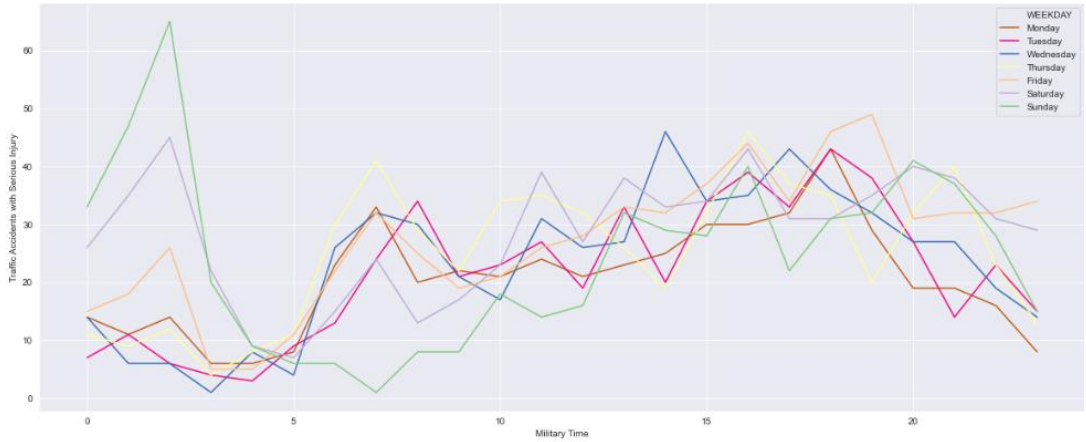
Time Series

The next exploratory analysis is performed on time series- the time of day, day of week, and months when traffic accidents occurred.

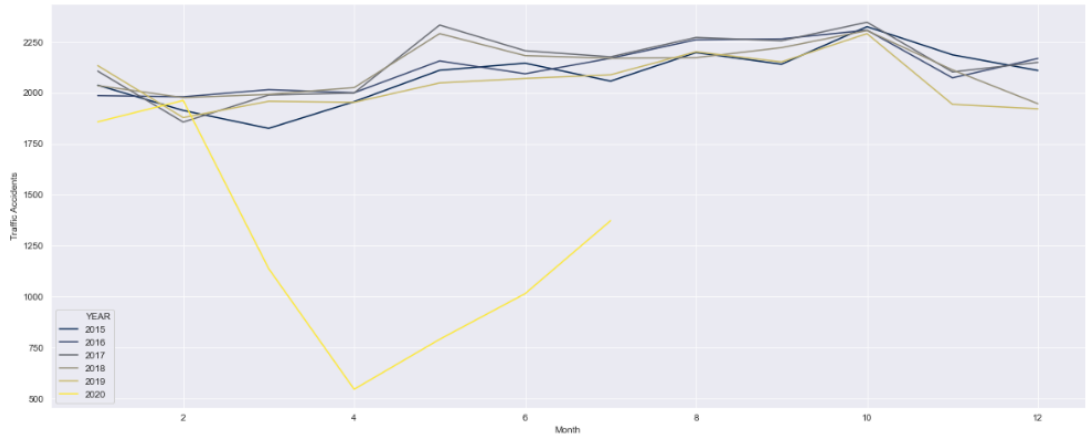
Below is a line chart depicting the number of total traffic accidents by day of week, with the time of day on the x-axis. The plot indicates that traffic accidents happen most during rush hour, peaking at around 8 am and 5 pm on weekdays from Monday through Friday. For weekends, Saturday and Sunday, there is a relative spike of traffic accidents at around 2 am, followedby a steep drop until 5 am, where it then gradually increases.



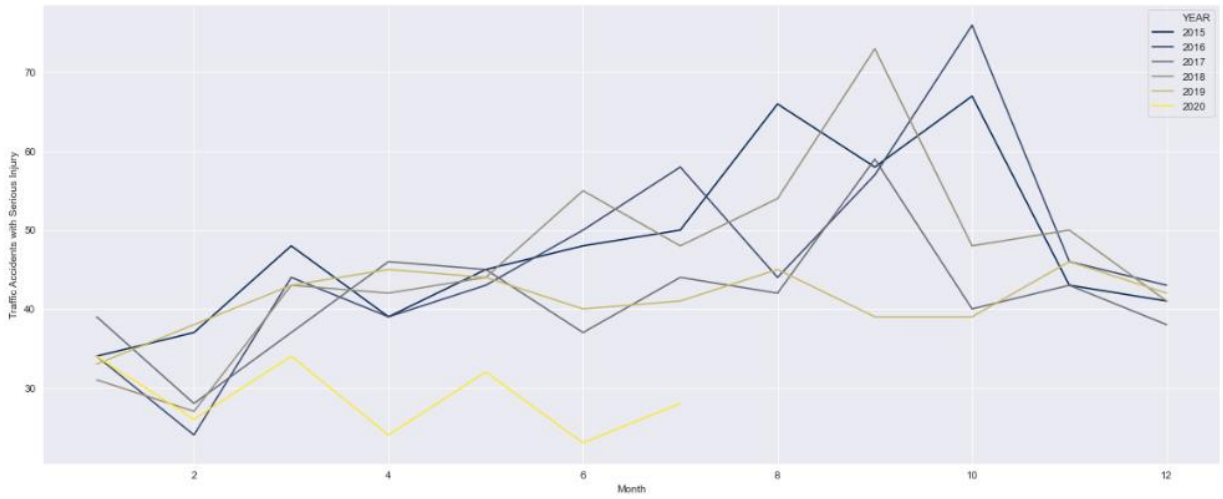
The graph below plots for traffic accidents that resulted in serious injury or fatality. During the weekdays, the number of incidents stayed relatively low from around midnight to 5 am, with the exception of Friday, and starts to increase as the day progresses. For weekends including Fridays, the highest count of traffic accidents resulting in injury happened at around 2:30 am, with Saturday being the greatest.



Traffic accidents by year during the calendar months are shown below. The peak points are in May and October for all preceding years before 2020. Year 2020 is an anomaly due to covid-19, which reflects a steep dip in March and April, with a pivot increase as it progresses to May and beyond. This graph does not include August and September traffic data due to the likelihood that these months have not yet closed and are incomplete.

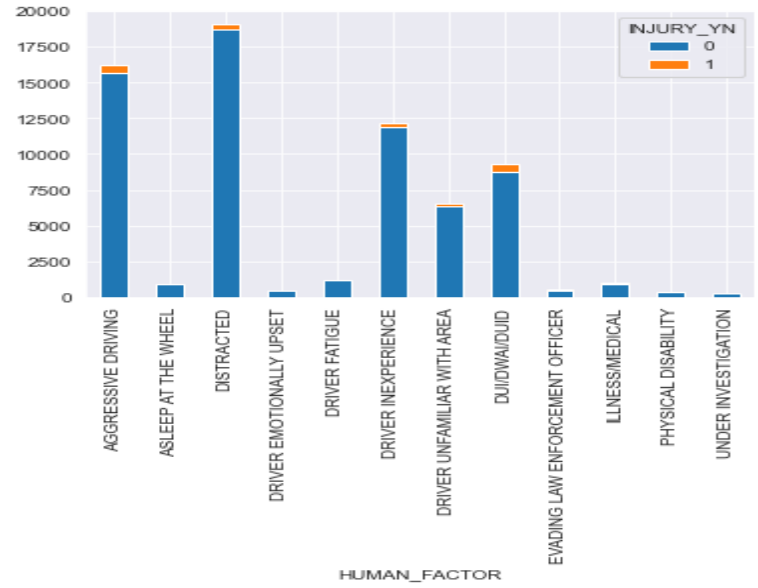


When graphing traffic accidents resulting in serious injury (excluding year 2020), the previous years show peaks in the fall at around September and October, followed by a dip from November through February, indicating that although inclement weather tend to happen during the winter, it appears that weather is not a strong factor in causing or correlating with serious injury from a traffic incident. However, this can also mean there are not as much driving during the holidays.



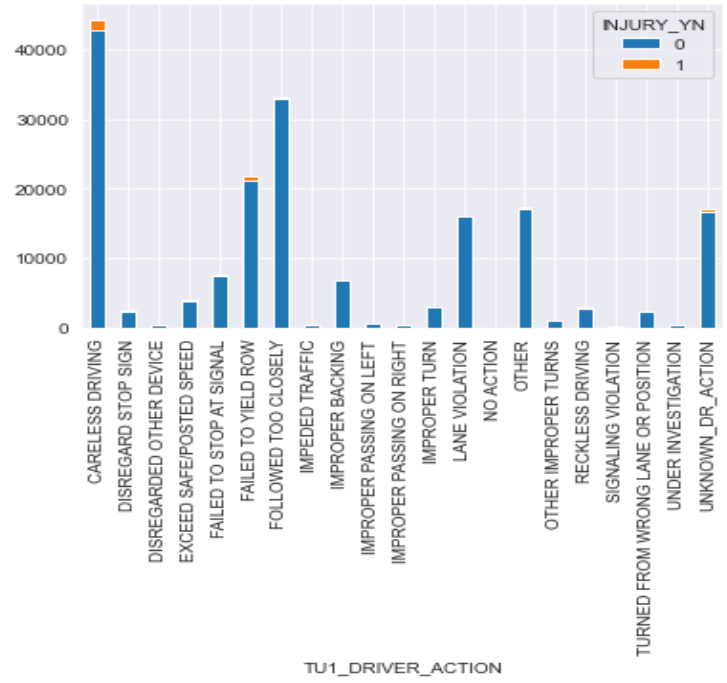
Human Contributing Factors

Below are descriptions of human behaviors contributing to traffic accidents. The top accident causing behaviors are from Distraction, Aggressive Driving, Driver Inexperience, DUI's, and Driver Being Unfamiliar with the Area, in their respective order. Top behaviors that resulted in serious injuries are from DUI's, Aggressive Driving, Distraction and Driver Inexperience, respectively. The orange '1' indicates incidents with serious injury (including fatalities), and blue '0' indicates no serious injuries (or fatalities).



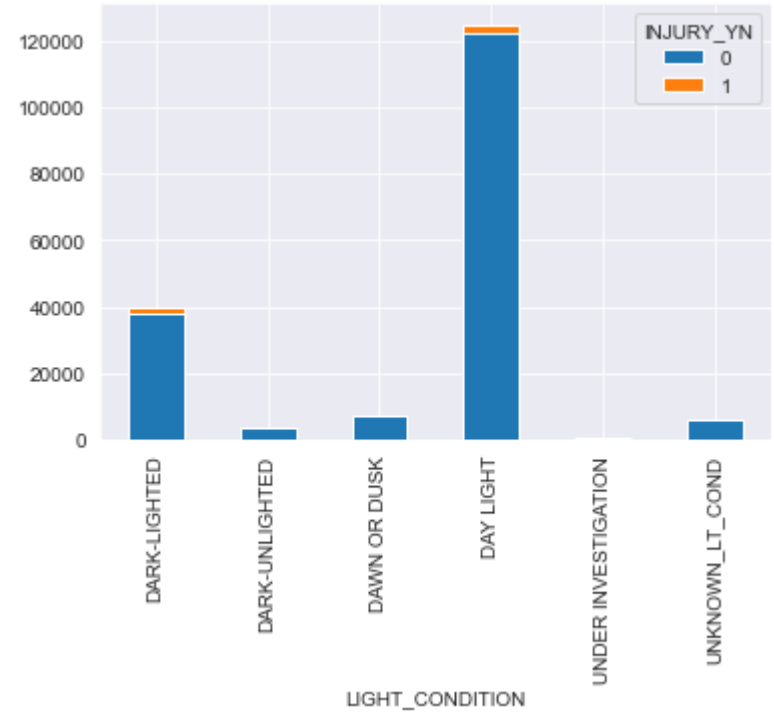
Driver Action

Below are descriptions of driver actions that are attributable to traffic accidents. The top known accident causing driver actions are Careless Driving, Following Too Closely, and Failing to Yield, in their respective order. Top actions that resulted in serious injuries or fatalities are from Careless Driving and Failing to Yield.



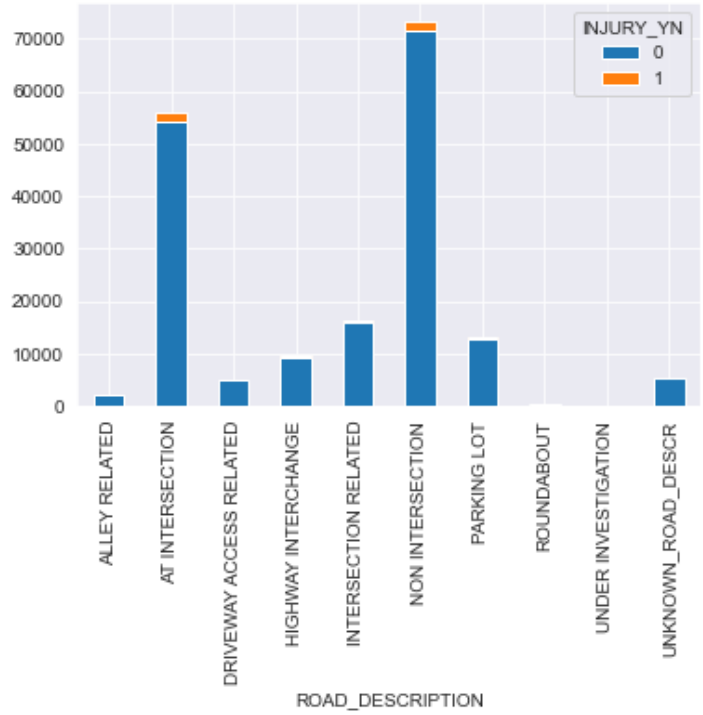
Light Conditions

Below are lighting conditions that may be attributable to traffic accidents. The majority of traffic accidents occurred at day light, and dark-lighted areas coming at a distant second. Serious injuries (and fatalities) happened both during day light and dark-lighted conditions.



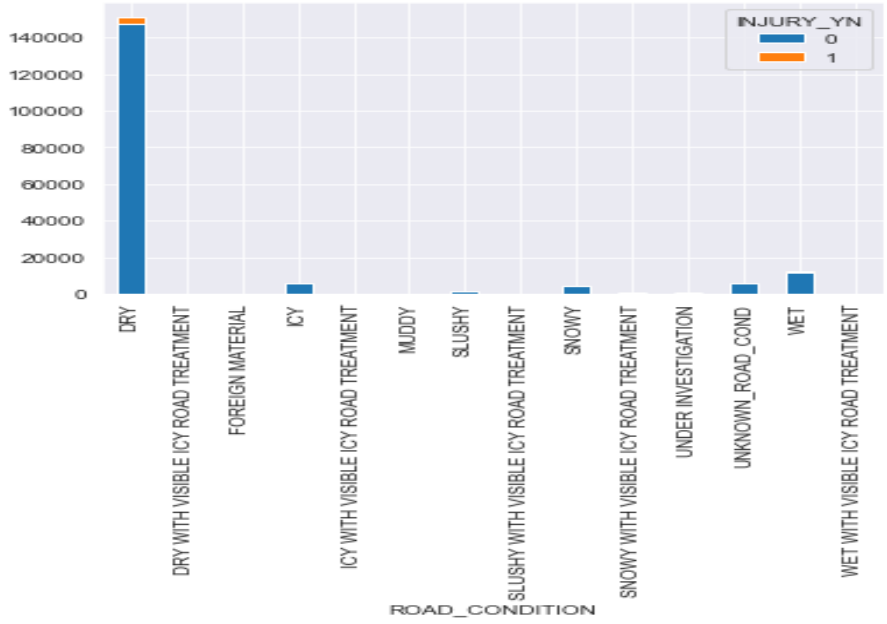
Road Description

Traffic accidents with and without serious injury occur most often at non-intersections and intersections.



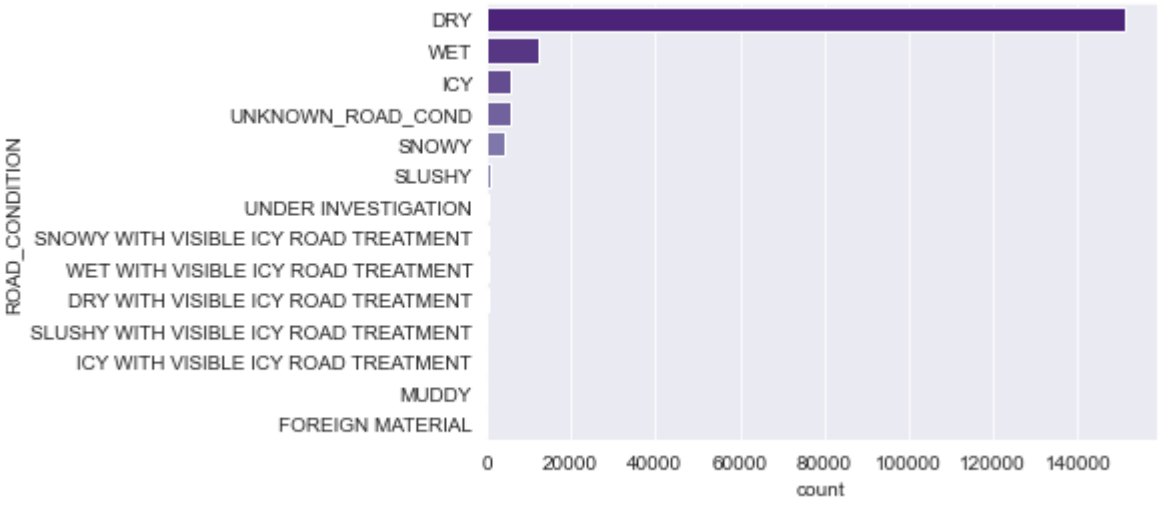
Road Conditions

Below are road conditions that may be attributable to traffic accidents. With the high majority of traffic accidents coming from dry roads; the road conditions seem to have less of an affect in causing traffic accidents or serious injuries.

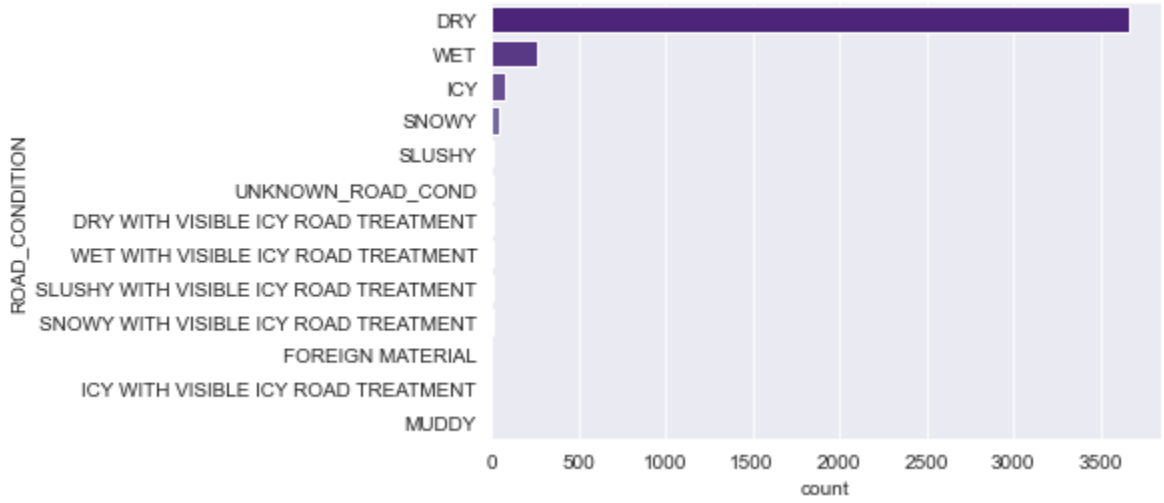


At a closer look, although wet and icy roads may cause an increase in traffic accidents, they do not present a notable impact on increasing the probability of serious injury or fatality.

Bar chart below: Road conditions for all traffic accidents



Bar chart below: Road conditions for traffic accidents with serious injuries



Recommendations

Based on the above exploratory analysis, there are various factors that can impact the probability of serious injury or fatality from a traffic accident. Each area is seen to also have varying degree of affect. Human behavior and driver actions are most correlated with incidents with serious injuries, and road conditions show the least impact.

The following are recommended to decrease a person's chance of being serious injured from a traffic accident:

- Stay off the road from around 2:00 am – 3:00 am on Fridays, Saturdays, and Sundays.
- Avoid drinking and driving.
- Avoid road rage (i.e. aggressive driving).
- Driving on the highway increases the chances of serious injury.

There are other factors that can play into the severity of a traffic accident, but the aforementioned are the most prominent causes or correlations in a traffic accident with serious injury or fatality.

Features Section

Features selected for predictive modeling includes all of the above attributes in the exploratory analysis with exception of Road Condition. Road conditions do not show a significant impact and thus should be excluded from the training dataset to reduce computational cost. The training dataset for predictive modeling is also balanced to ensure the number of traffic accidents with and without serious injuries are not disproportionate.

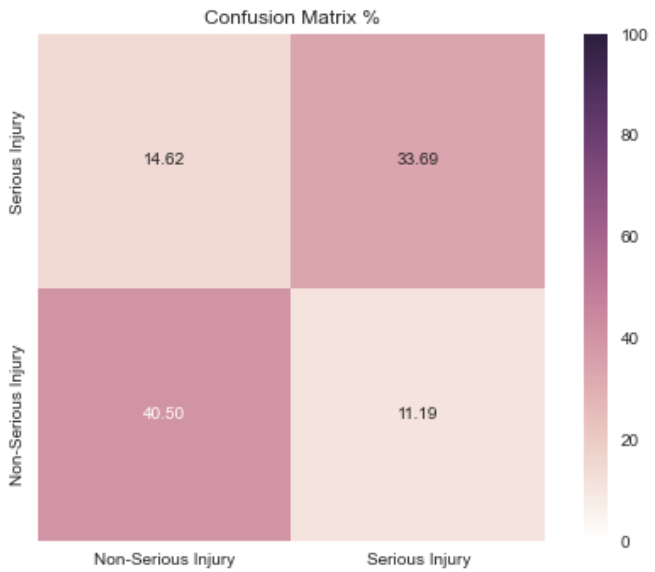
Methodology and Results

The methodologies used for predictive modeling in supervised machine learning on whether or not a traffic accident will result in a serious injury- are logistic regression and K-nearest-neighbors.

Logistic Regression

Logistic regression is a classification model that predicts an outcome using conditional probabilities. This method is suitable as the features from the traffic accidents have varying degrees of impact, thus affecting the probability of a serious injury. The evaluation metrics from the logistic regression performed are below:

Accuracy	0.7419
F1 Score	0.7413
Jaccard Score	0.5661
Log Loss	0.5232

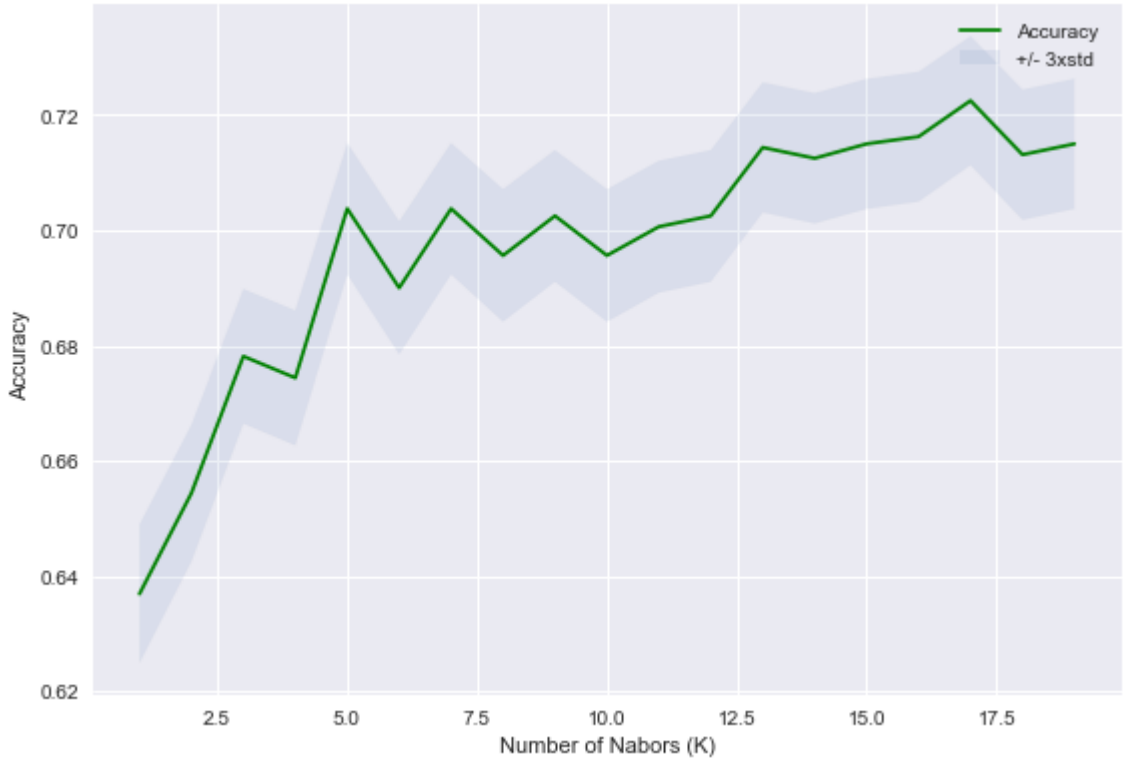


An accuracy rate and F1 score of 0.74 generally reflects 74% accuracy in predicting a serious injury, although, these rates can further be improved. F1 score takes into consideration the false positives and false negatives, accounting for precision and recalls. The confusion matrix above reflects the percentages of true and false positives and negatives, in which includes the underlying data in calculating the F1 score.

Jaccard Score is 0.57; this score is used for gauging the similarity and diversity of sample sets. The Log Loss value is 0.52. Log loss is a measure of uncertainty (i.e. entropy) in the underlying probabilities of the model, it is better for a log loss to be closer to zero.

K Nearest Neighbor

K Nearest Neighbor (KNN) is a classification algorithm that uses a labeled input dataset to predict the output of the data points. It is most useful for target values that have feature similairty. KNN checks how similar a data point is to its neghbor as classifies that data point into the class it is most similar to. Instead of using weighted probabilities such as in a logistic regression, KNN uses distance between plots for classification. It is important to determine the best K value to calculate for KNN. The Elbow Method is used to determine this optimal value of K; the K-Means clustering technique is performed to calculate possible K values. Based on the below visualization, the optimal number of clusters should be at around K = 17, the highest point in the graph.



KNN evaluation metrics at K = 17 are below:

Accuracy	0.7219
F1 Score	0.7205
Jaccard Score	0.5326

The evaluation metrics for Logistic Regression are slightly better than KNN. However, the predictive models can further be refined to improve their accuracy either through revising or updating the features set, choosing different parameters for the algorithms, and/or obtaining additional quality data. To consider selecting a different machine learning algorithm is also an option.

Conclusion

There can be infinite factors and nuances that affect the possibility of a serious injury in a traffic accident, therefore, predictive models cannot realistically predict their outcome at 100%. However, being able to provide insights on the significant causes and conditions can only help the community mitigate such injuries. Raising awareness to better inform people will help them make better decisions when being on the road. This will result in a safer environment with less infrastructure damage, less public and personal property damage, lower medical costs, and most importantly, save more lives.