Predicting Severity of Accidents

Shaheer Airaj Ahmed

1. Introduction

Traffic accidents are in no scenario an ideal condition for the average human. The damage to their property, injury of self or others costs not just money but also can slightly disturb a person mentally depending on the severity of the accident.

Getting an idea of cause of accidents and predicting the severity can be useful to car, life and health insurance companies as it effects the number of customers who come to them to buy their insurance plans. Depending on different areas and most common types of weather and other conditions, the insurance plans can be priced and packaged differently to include different services.

So the goal of this capstone becomes to first find general trends, try to look at what is causing accidents, which features effect severity of the accident and finally create a machine learning model to predict the severity of accidents.

2. Data

The data was provided by IBM themselves. It contains 38 features and 194,673 rows. The dataset consists of many columns containing both categorical and numeric data and gives detailed descriptions of keys provided for each type of incident.

The features I am going to be working with are SEVERITYCODE, ADDRTYPE, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, VEHCOUNT, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, INATTENTIONIND, UNDERINFL.

These are the features I've decided to use and uncover the insights they have to provide. First I will try to get an overview of what these features are generally trying to tell us in terms of the number of road accidents for each situation or category and then see how the features relate to the severity codes and what methods will be useful in predicting the severity of accidents.

- SEVERITYCODE: Contains two values, 1 and 2. 1 relates to prop damage and 2 relates to injury.

- ADDRTYPE: This contains information on the location of the accident in relation to either intersection, block or alley.

- COLLISIONTYPE: Tells us the type of collision that took place such as Angles, Sideswipe, Parked car, cycles, rear ended, head on, left turn, pedestrian, right turn, and other.

- PERSONCOUNT: The number of people in the accident
- PEDCOUNT: Number of pedestrians involved in the accident
- VEHCOUNT: Number of vehicles involved in the accident
- WEATHER: Type of weather
- ROADCOND: The road condition at the time of accident
- LIGHTCOND: The light condition at the time of the accident
- SPEEDING: Whether the person was speeding
- INATTENTIONIND: Whether the accident was caused to inattention of the driver
- UNDERINFL: Whether the driver was under influence
- INCDATE: Giving the date of the incident

3. Data Exploration and Cleaning

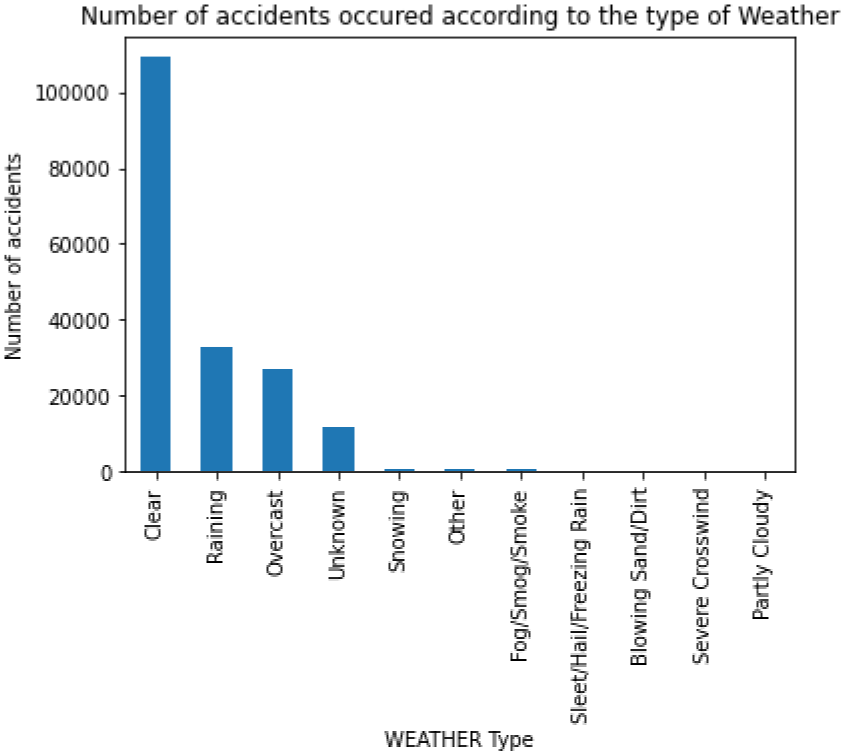
The dataset was filled with inconsistencies in formatting, missing values, repeated features and redundant features. To start off, I got rid of all features which I didn't require for my analysis using my own understanding. Got rid of all the keys and descriptions to the keys given as they would produce no useful insights.

The entire dataset contained 1,100,024 missing values in the beginning. After collecting the useful features, I went about to see which features were categorical and which were numeric and handle them accordingly. SEVERITYCODE was an int64 feature with only 2 unique values(1 and 2) and contained no missing values so no changes were required there. I first decided to tackle my numeric data mainly SPEEDING, UNDERINFL AND INATTENTIONIND by mapping 'Y' to 1, 'N' to 0, strings '1' and '0' to int64 type, and replacing nan values in SPEEDING AND INATTENTIONIND to 0 as there was only 'Y' as an entry in which case it was safe to assume that the nan values represent 'N'. Once that was done, I dropped null rows from the entire dataframe.

The last step was to clean the dates by using 'to_datetime' function from pandas library and creating features 'YEAR', 'MONTH', and 'DAY' for future analysis and then dropping 'INCDATE' from the dataframe.

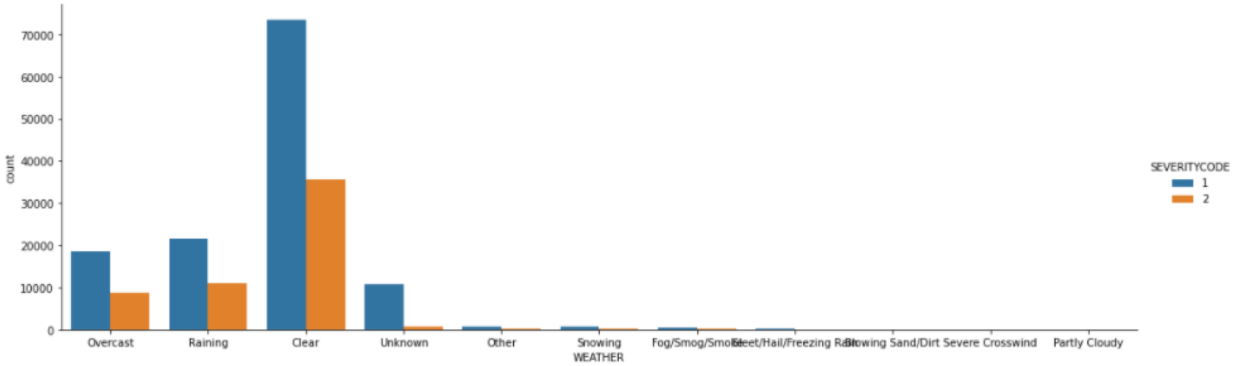
4. Data Visualization

Finding insights and trends is crucial to understanding the data and figuring out how to model ML algorithms. Below we will look at some of the insights.

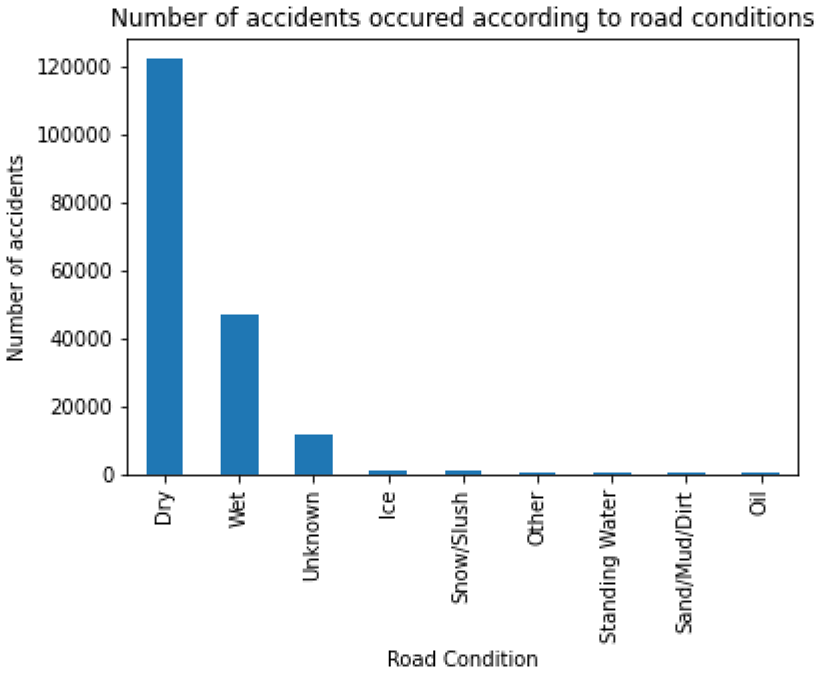


A staggering amount of accidents occured when the weather was relatively clear. This suggests bad weather barely has anything to do with the cause or severity of the accident. Although rain and overcast types of weather are responsible for between 27,000 to 33,000 accidents in the last 15 years which is still significant.

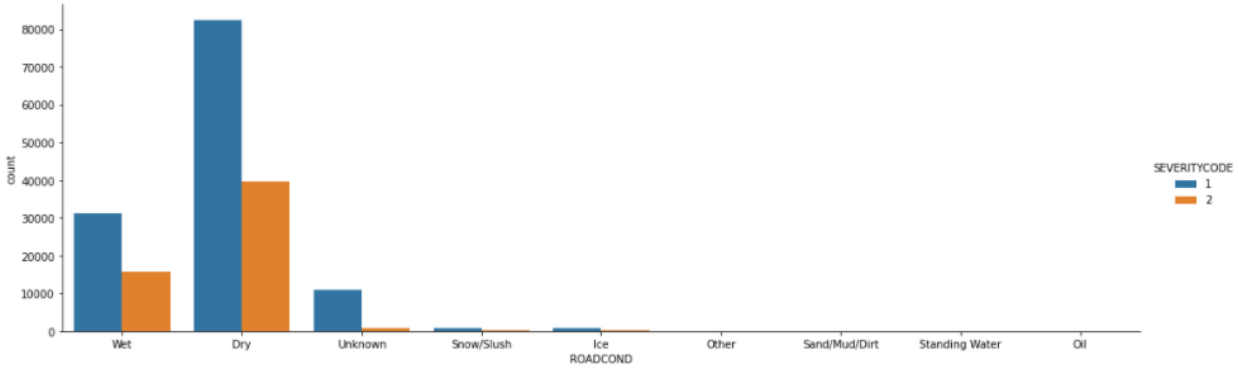
Now lets see how severe the accidents were for each of the weather conditions



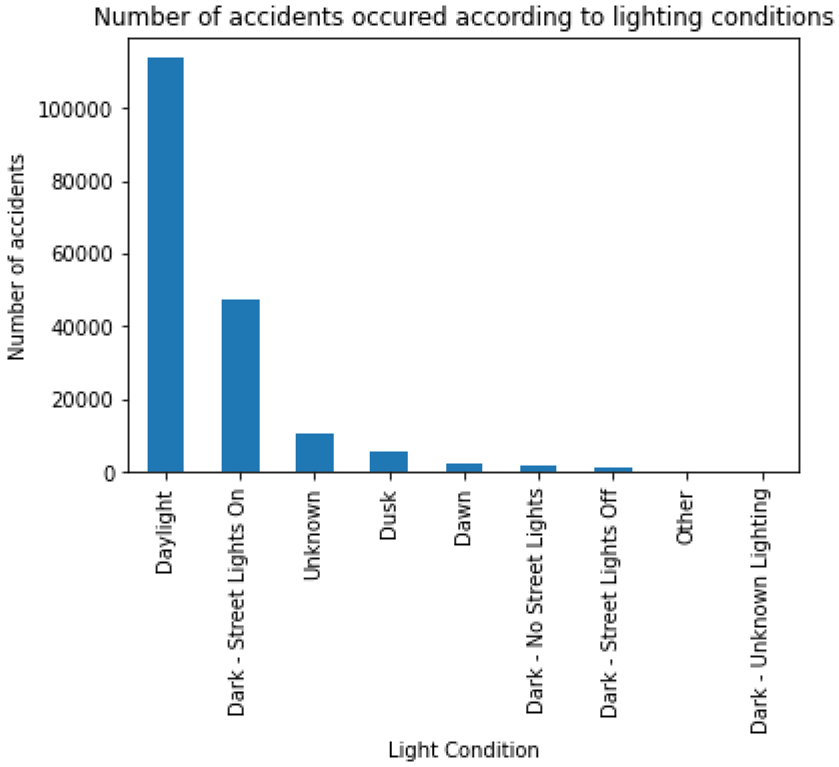
The above graph shows the number of sever accidents to take place during each type of weather. Mostly the accidents were at SEVERITYCODE 1 which according to the metadata means they were prop damage so nothing to worry about. But SEVERITYCODE 2 means there was injury although not serious as that is SEVERITYCODE 2b which none of the accidents are.

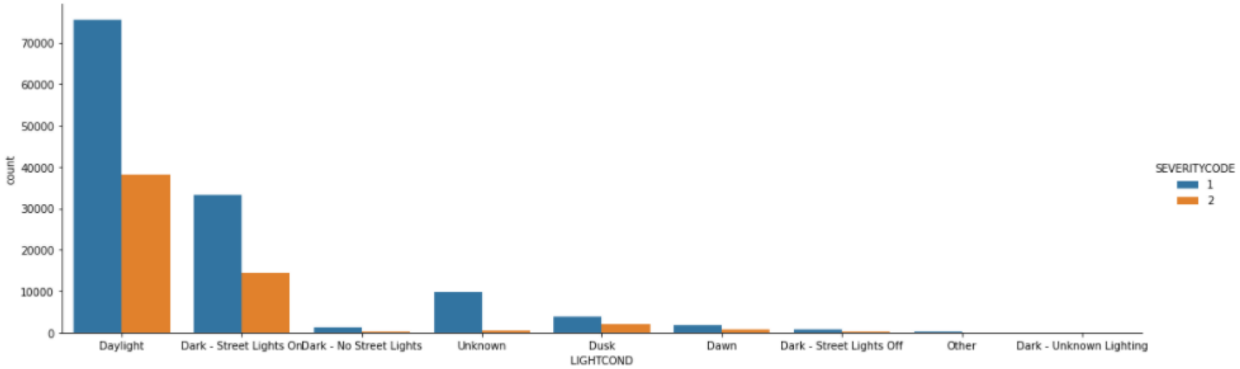


Again it seems like the road conditions were mostly ideal. Bad road conditions weren't as big of a factor. But which accidents according to each type of road conditions were code 1 or code 2?

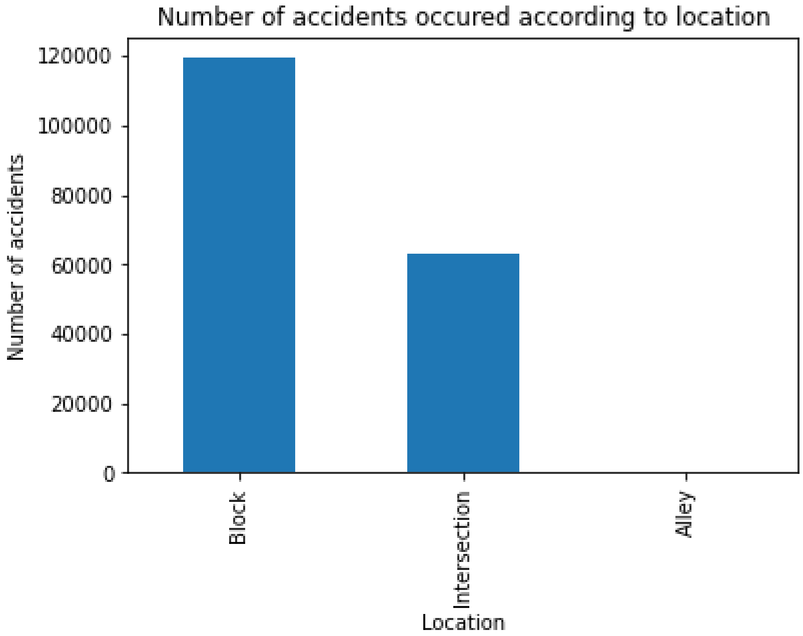


Again we see the trend of most accidents being code 1 rather than code 2



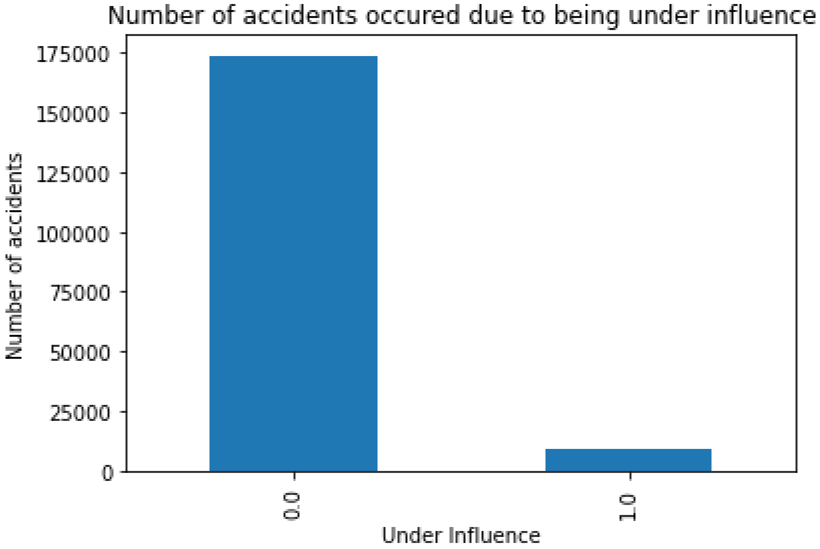


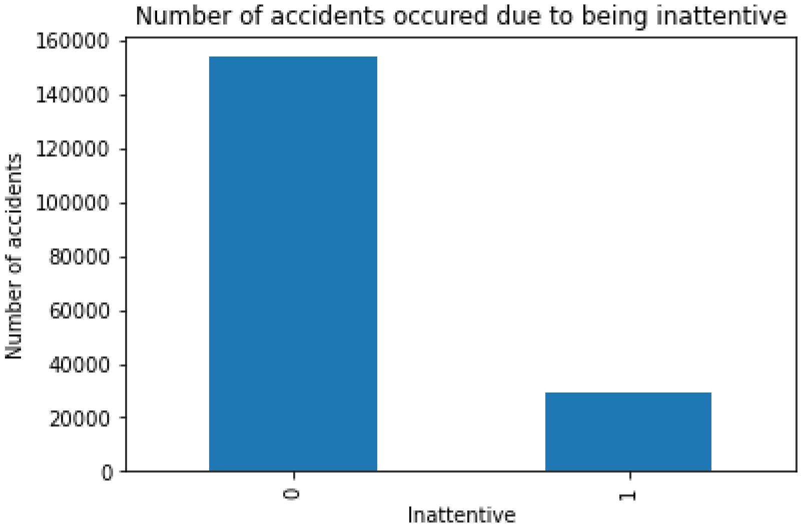
Daylight seems to be where most accidents occur followed closely by dark but with street lights on. So far it looks like outside conditions are not relevant to the causing of the accidents. But before we come to our conclusion lets look at ADDRTYPE to see where most of our accidents occurred.

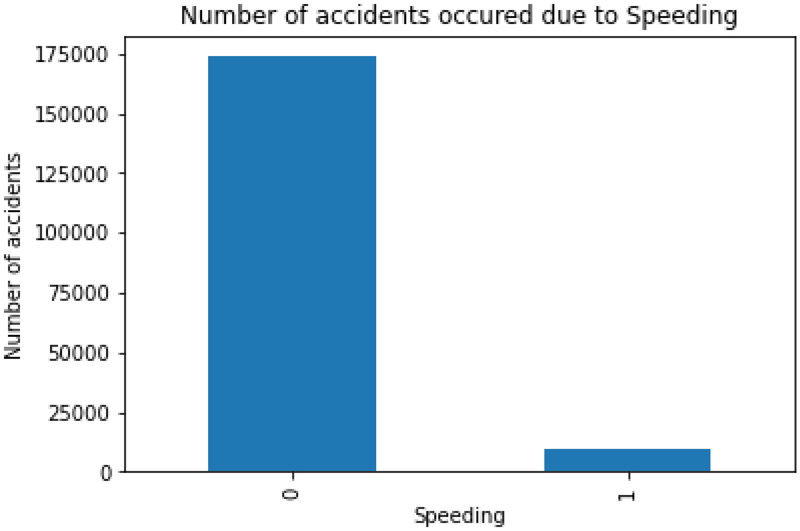


The data well informs us that outside factors aren't as significant as we thought in relation to the occurance of accidents. So then what is?

What is left for us is to analyze human factors of such accidents such as inattention and driving under influence and see how they relate.

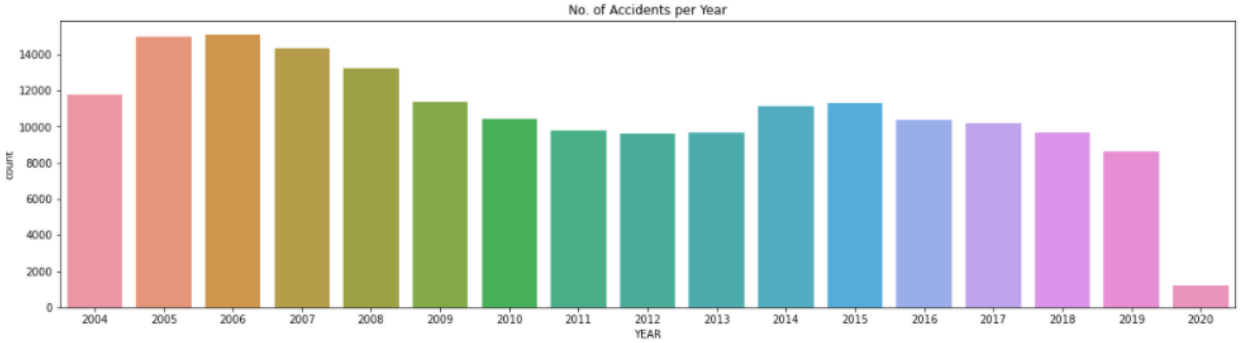


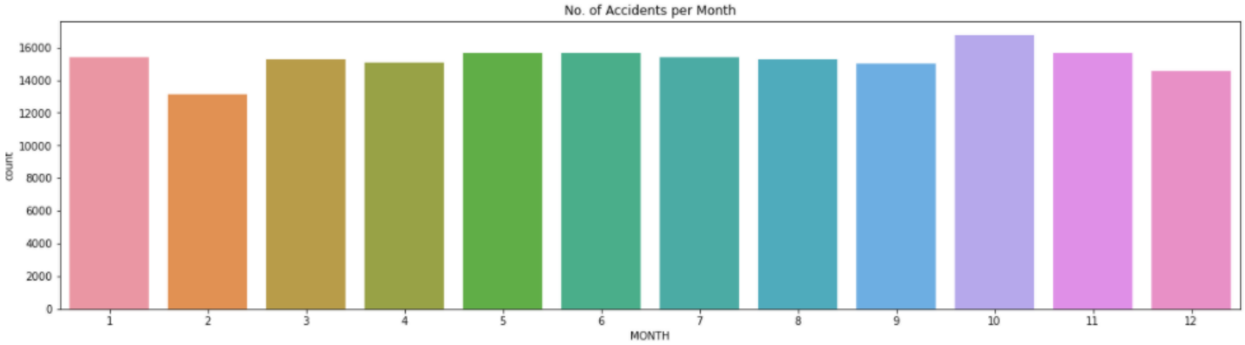


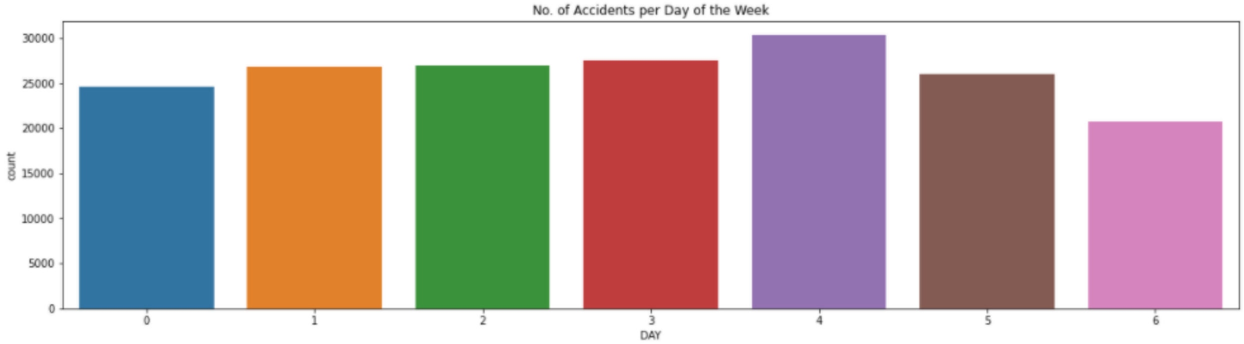


From this we can conclude that there isn't a single greatest factor in determining the cause of accidents but we can say that in terms of when we have bad weather, rain and wet roads are the greatest cause of accidents.

Lets look at the trends of road accidents that have happened overtime.







According to year, the number of accidents where highest in 2005 and 2006 but recently have seen a steady decline.

According to month, for the last 15 years, October has generally seen the largest number of accidents.

According to the day of the week, the highest number of accidents were on a Friday, the end of the weekday where everyone is in a rush to go home or enjoy their weekend.

5. Data Preparation

The first part of this section is using one-hot encoding on the categorical data of our dataset. This in turn gives us a lot more features to work with but this way the weightage given to each category of data will be equal while using machine learning models. This is important to remove biases in our data and models. Each unique category in each feature was given dummy values by using (get_dummies) function in pandas and then renamed to make more sense.

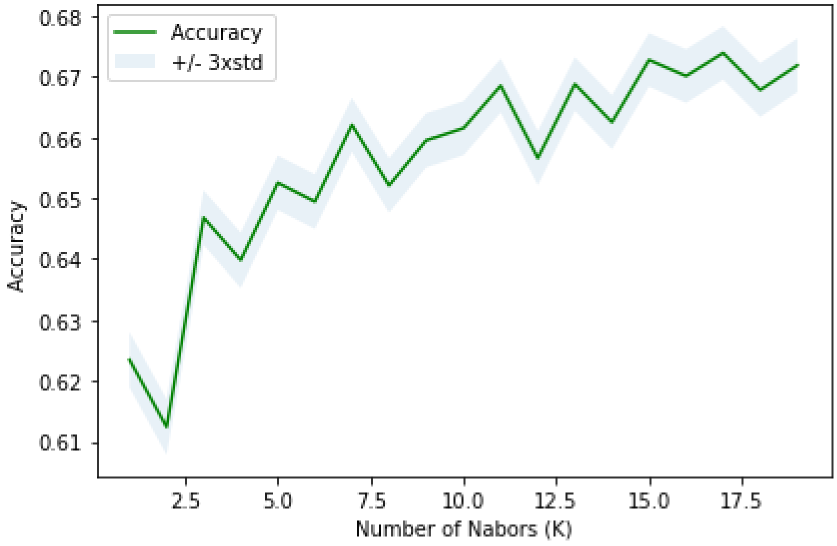
The second part is down sampling of the data. After looking closely at the SEVERITYCODE feature we see that code 1 consists of 69% of the total values in the feature. This will again train a biased model if left unchecked. My solution to this problem was to down sample the data by using resample from the sklearn.utils library in order to bring both severity code entries to the same value. We end up losing some of the data, but if we leave the SEVERITYCODE feature as it is, the model will mostly only predict code 1 and almost never code 2.

The last step was to split our data into training and testing sets. I decided to keep the test size 10% of the data in order to give the models sufficient room to train properly and effectively.

6. Machine Learning Models

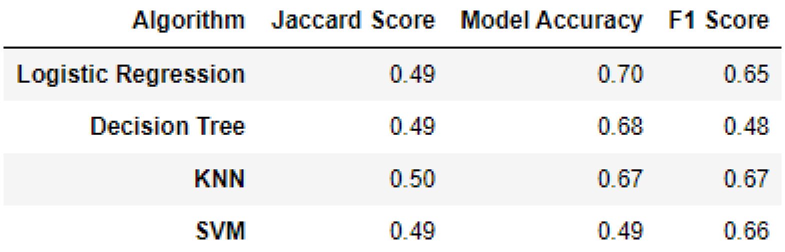
I trained the data on multiple models. The chosen models were Logistic Regression, Decision Tree, K-nearest Neighbor and Support Vector Machine. I previously face the problem that all my models were only predicting severity code 1 and completely neglecting code 2. Later I figured out that it was due to the bias in my SEVERITYCODE feature having data mostly for code 1 which was then adjusted by down sampling as shown in the data preparation step.

I also evaluated the optimal number for K for the KNN model and graphing it as shown below.



The accuracy generally keeps increasing as we keep going but I capped the value of K at 20

The metrics for evaluating accuracy was accuracy_score and f1 score for each of the models and the report is shown below.



Algorithm	Jaccard Score	Model Accuracy	F1 Score
Logistic Regression	0.49	0.70	0.65
Decision Tree	0.49	0.68	0.48
KNN	0.50	0.67	0.67
SVM	0.49	0.49	0.66

From our readings we can see that Logistic Regression has shown the highest accuracy and a fairly good f1 and Jaccard score. Therefore, we can conclude that Logistic Regression is what works best for this model.

7. Conclusions

In this project I have analyzed multiple features of the dataset to determine what are the causes of accidents and what factors relate to severity of an accident. It is clear that bad weather or road and light conditions are not major factors to the causation of accidents as most accidents were observed in ideal environmental conditions. Nor were most accidents caused by the drivers' inattention, speed or them being under influence. As there was no clear feature that had more influence on severity more than the other, upon using One-Hot Encoding we gave each categorical variable the same weightage and allowed the model to determine the trends. As for the machine learning model, logistic regression seemed to be the most accurate model amongst them.

8. Future Work

Although I had determined a fair number of basic trends, there were many other analysis that could be determined from the feature set as well as some of the dropped features which were too complex for me at this stage. For future analysis, determining which feature had most influence on the outcome of severity from the model could also be determined. And hotspots of where accidents most occurred in terms of location on the map could also be looked at.