Predicting severity of accidents based on prevailing conditions

1. Introduction

1.1 Background

It is only a matter of time until all of us are going to hit the road. More so, because we have been deprived of our freedom to move around in the last few Covid19 months. Freeways will be buzzing with traffic, rogue drivers, slow drivers, lane cutters and traffic jams. Due to these, one of not-so-good-things will return in a big way- Accidents. A variety of reasons can be attributed to any accident; ranging from a straight forward one like mechanical issue, driver error to a combination of variables - bad light + snow + fatigued driver.

While we may not be able to prevent every accident from occurring, it will be great if we can get a reliable alert to drive with caution when conditions are not right. Today, we have radio stations or sign boards warning us of bad weather or slippery conditions. It would help if there was an intelligent system on our vehicles which can predict likelihood of an accident based on prevailing conditions at run-time

1.2 Problem

This project aims to predict the severity of accidents by using all the relevant data available on past accidents. This could help provide an alerting mechanism that can fore-warn drivers on the potential risk of accidents.

1.2 Target Audience

A system like this would typically interest officials from Department of Transport and Safety to predict accidents and plan their resources effectively. It can also interest auto industry to create intelligent features to increase their customers' safety

2. Data Acquisition and Analysis

2.1 Data Acquisition

The data used for this project has been procured from Seattle Department of Transportation. It contains all collisions data from 2004 to present along with details like incident type, address, persons involved, road condition, light condition, etc.,

2.2 Data Cleaning

As with any data available in public domain, there were redundant information, missing values and details that cannot typically utilized in a machine learning model.

Below are steps taken to clean and standardize the data.

1. Rows with missing values for columns X, Y, WEATHER, ROADCON and LIGHTCON were dropped since they seem to contain information critical to ascertain severity of the accident.
2. Columns like INATTENTIONIND, UNDERINFL, WEATHER, ROADCON and LIGHTCON needed more cleansing. INATTENTIONIND was filled only when it was true. All other blanks were defaulted to False. UNDERINFL had both 0,1 and Y,N. They were converted to have 0s and 1s.
3. WEATHER, ROADCON and LIGHTCON had "Unknown" as one of the options. If all three had the same values, these rows were dropped. I also dropped rows which had one of ROADCON and LIGHTCON as "Unknown"
4. INCDTTM indicates date & time of the accident. For 20% of the data, time field of the accident was not available. This missing data was tackled by computing the average time (13:00) of all the other accidents and updating the fields with it
5. Columns pertaining to unique identified for collisions & reporting were dropped since they do not any significance to predicting severity

2.3 Feature Selection

Post data cleaning, there were 166705 samples and 21 features in the dataset. Next, another round of analysis was performed to identify redundant information present pertaining to location and impact of the accident.

There were multiple fields describing the location of the accident - X, Y Co-ordinates, INTKEY, LOCATION, JUNCTIONTYPE, CROSSWALKKEY. After a close at the data present in these fields, I chose to retain ADDRTYPE and JUNCTIONTYPE as it contained more specific details of the location

Again, multiple features were present to capture impact of the accident - COLLISIONTYPE, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT, VEHCOUNT, HITPARKEDCAR. PEDCOUNT, PEDCYLCOUNT were dropped since their incidence was relatively less.

Below list of 13 features were shortlisted for exploratory analysis to determine their relationship with accident severity:

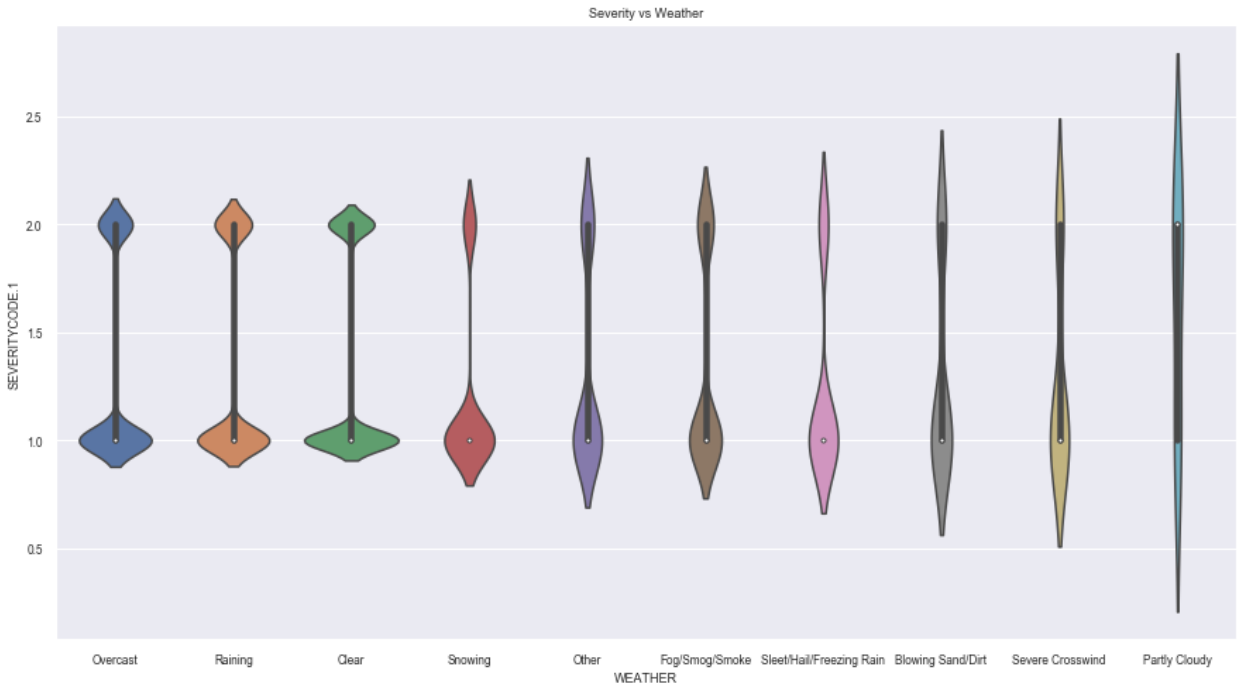
COLLISIONTYPE, PERSONCOUNT, VEHCOUNT, JUNCTIONTYPE, INATTENTIONIND, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND, SPEEDING, HITPARKEDCAR, INCDTTM

The final set of features identified was then utilized in building various machine learning models in order to achieve the objective of this exercise

3. Exploratory Data Analysis

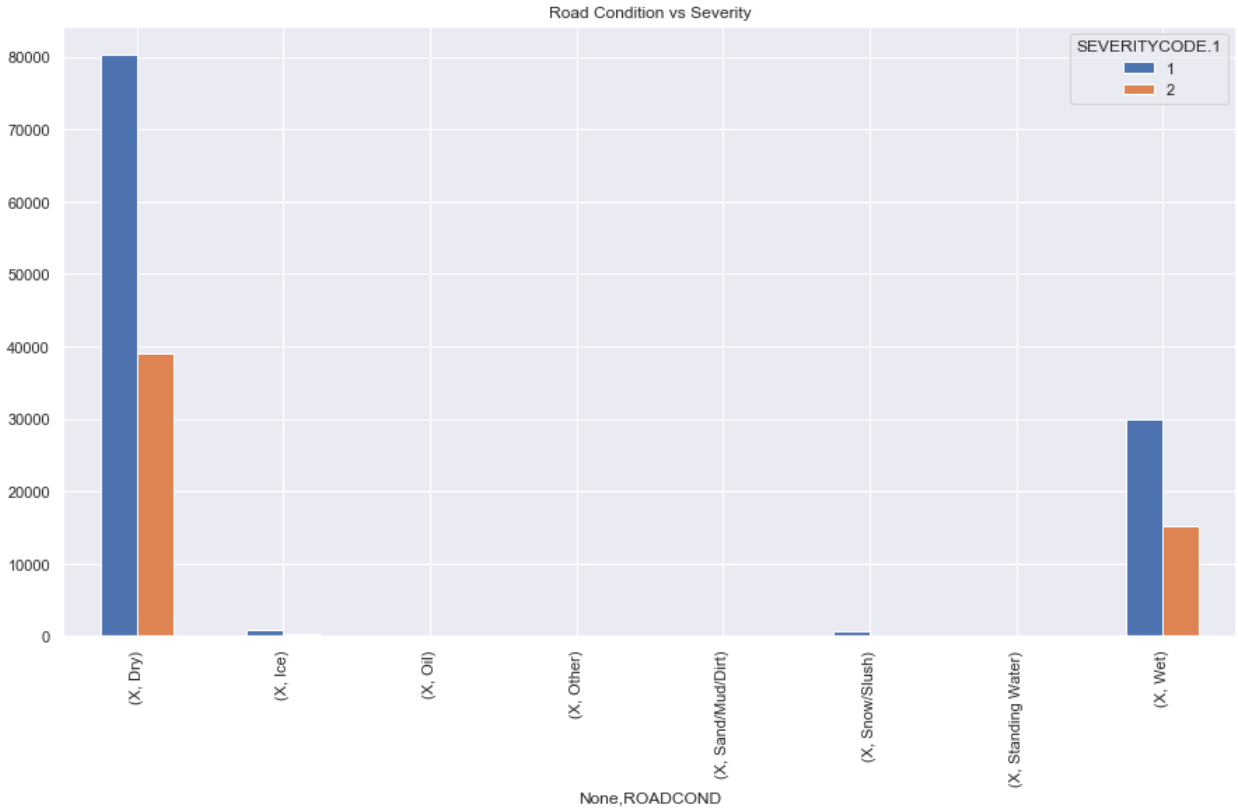
3.1 Relationship between weather and severity

Weather should play a significant role in the occurrence of accidents as well as its severity. Hence, I created a simple violin plot of weather conditions vs severity of accidents. This plot clearly showed an increase in severity of accidents whenever there was inclement weather like Raining, Snowing, Overcast conditions, etc. However, the plot also showed that lot of accidents did occur even when the weather conditions were clear.



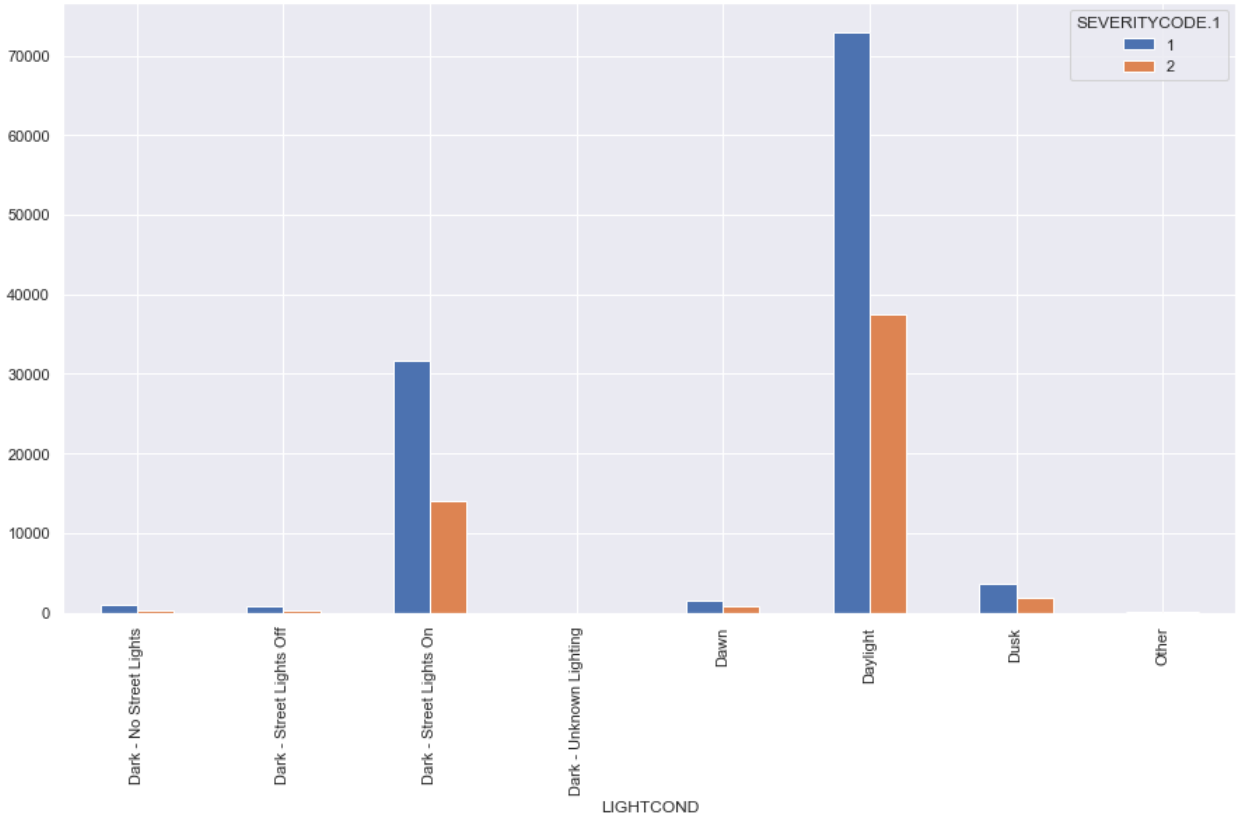
3.2 Relationship between road conditions and severity

Another independent variable that should have a direct relationship with accidents is operating "road condition". I used a pivot table to visualize the relationship between different conditions against the severity of accidents. As expected, there were high degree of severity 1/2 accidents when road conditions were "wet". There were also high number of accidents when the conditions were "dry".



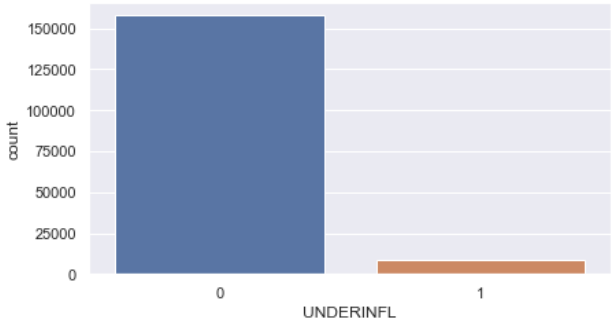
3.3 Relationship between light conditions and severity

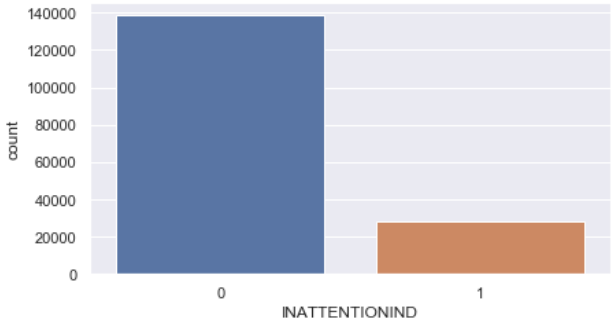
Light conditions should ideally affect the incidence of accidents. Hoping, there was a direct co-relation between poor light conditions and accidents, I plotted them together as below. However, there were higher number of accidents when there was either daylight or when street lights were on. The only possible reason I could deduce was that there should be higher traffic during day time or there were street lights on busier roads.

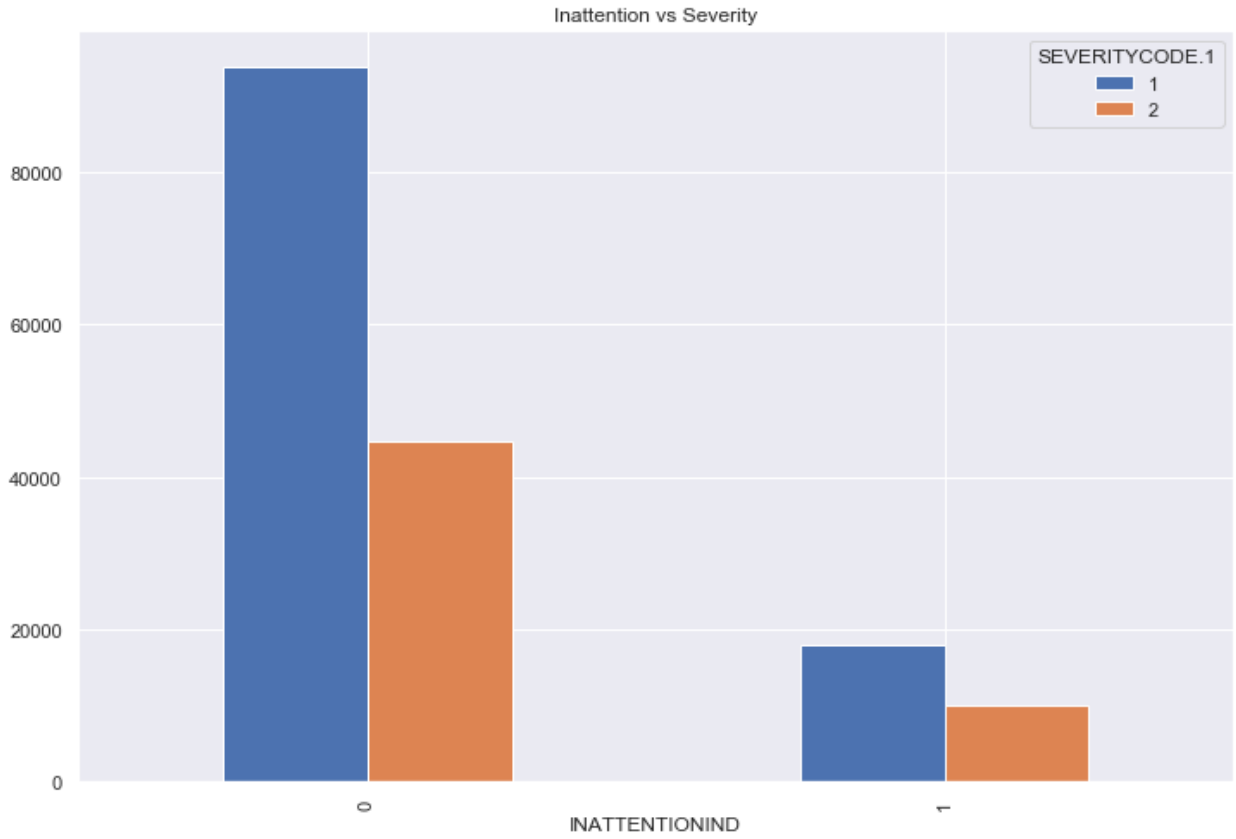


3.4 Relationship between inattention, under the influence and severity

Next, I assessed the frequency of accidents due to 2 indicators - INATTENTIONIND and UNDERINFL. There was very limited number of incidents due to inattention per the below chart. Hence, I dropped this variable as a feature as it does not seem to be significant. However, there were a decent amount of accidents due to the other indicator - UNDERINFL.



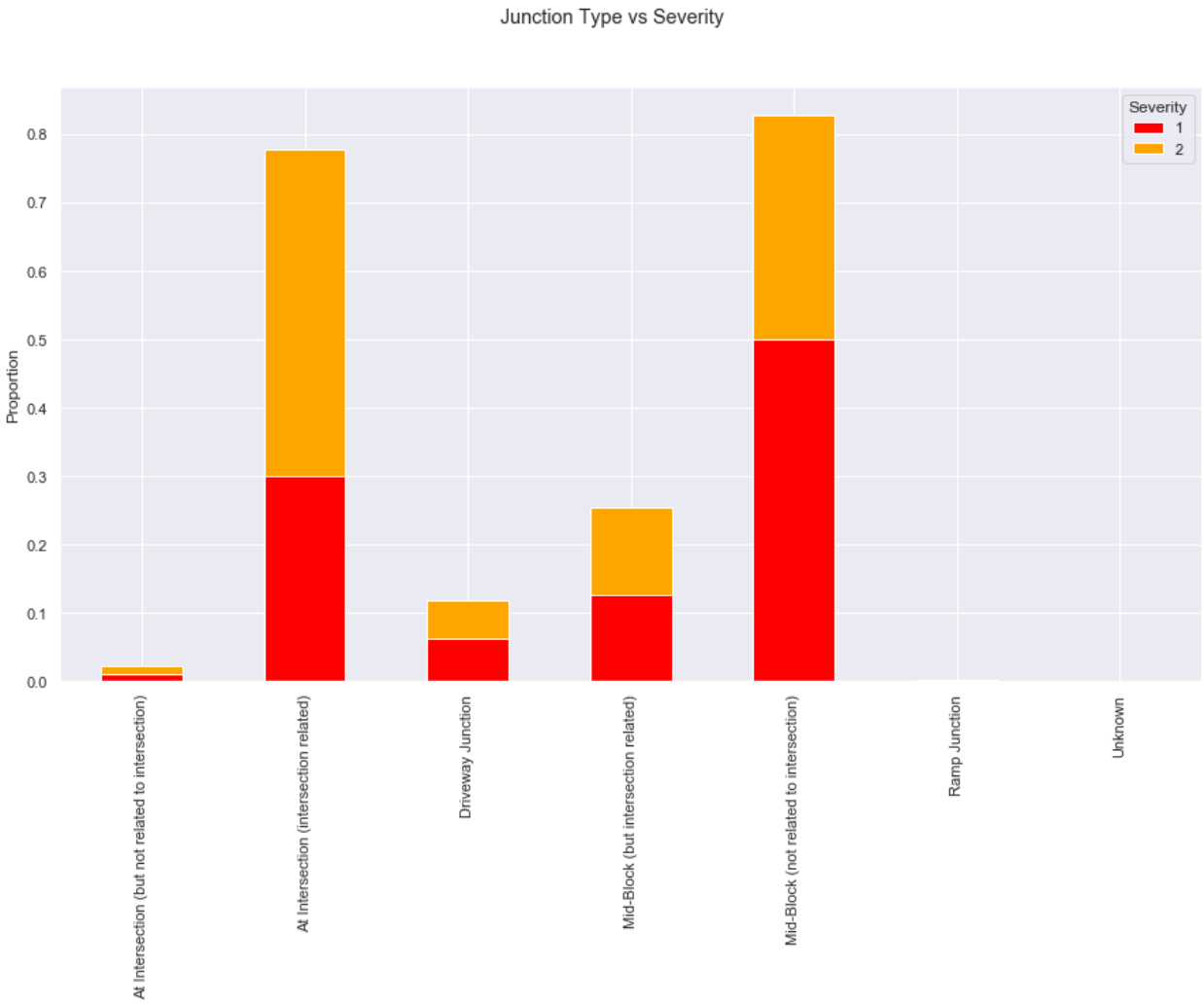




Hence, I plotted the total number of accidents due to inattention as against severity. As expected, there seems to be a direct relationship between them. This feature was included in the machine learning models to be created next.

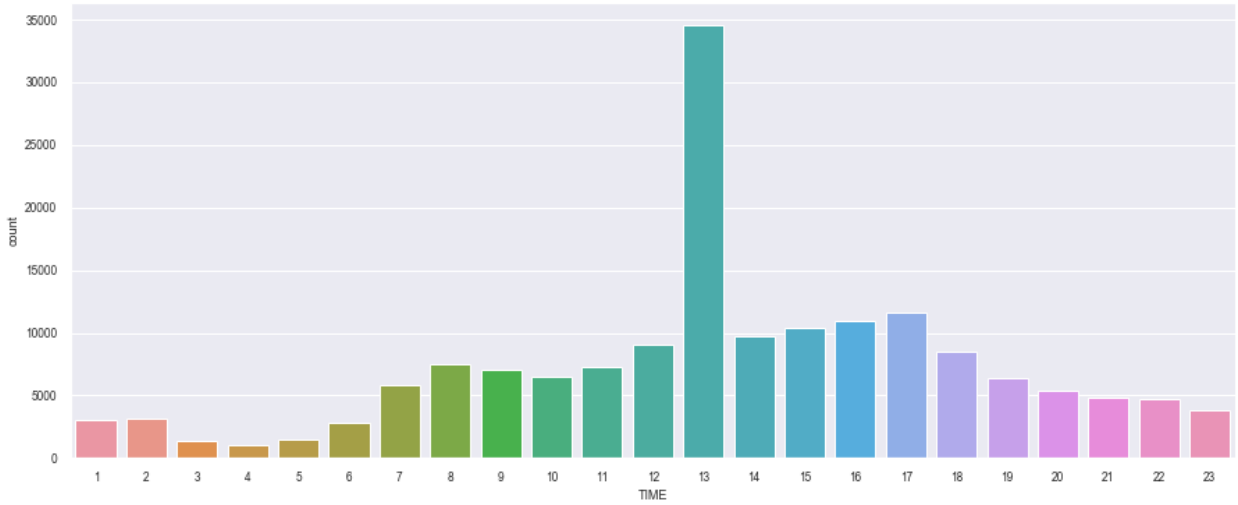
3.5 Relationship between junction type and severity

Junction type was one of the variables from the data set that I had included based on intuition. Based on plotting its relationship with severity of accidents, I was able to ascertain that maximum numbers of accidents occurred at mid sections (not intersection related) or at intersections. Typically, severity of the accidents was on the higher side in accidents occurring in mid-section as compared to intersections. Based on these observations, this feature was included in the machine learning models to be developed

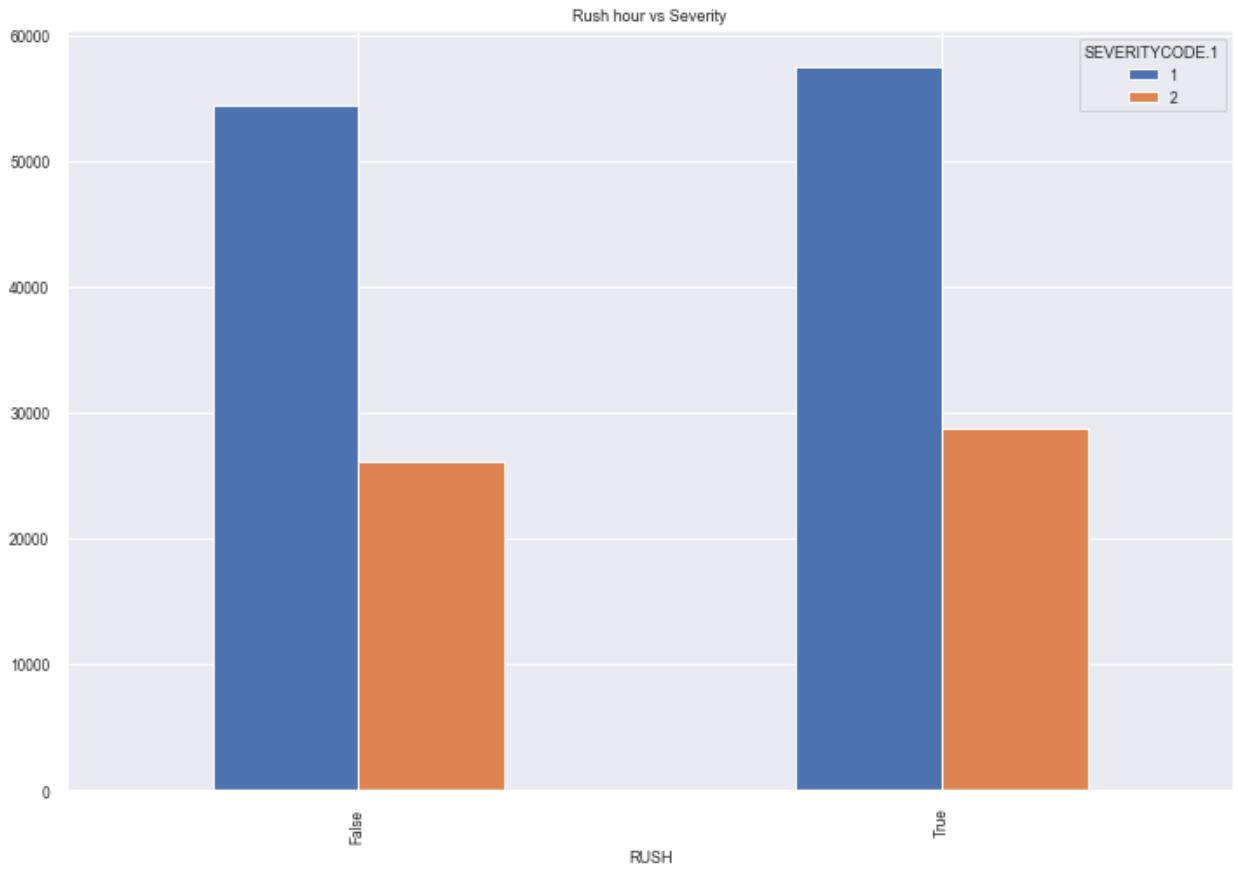


3.6 Relationship between incident time and severity

Traffic should play a critical role in the occurrence of accidents - more the people on the road, more the chances of these encounters. A count plot of INCDTTM should that frequency of accidents was highest between 1200 and 1700 hours. A new field "RUSH" was derived to capture this rush hour time



Plotting RUSH vs Severity yielded the below chart. The incidents with 1/2 severity was higher during this rush hour period as expected. This new field was also included in the machine learning model to help with the prediction.



3.7 Descriptive statistical analysis

Since, all of the significant features are all of categorical in nature, there was limited scope of descriptive statistical analysis. The value-counts of all these variables were performed and this was in-turn utilized to determine the significance of a particular variable to severity. Based on this, I eliminated SPEEDING, HITPARKEDCAR as relevant features.

4. Methodology - Predictive Models

We are intending to build a system that will be able to predict the severity of the potential accidents based on features identified so far. This is essentially a classification problem.

I chose to create KNN, Decision Tree, SVM and Logistic Regression models for building this prediction system. F1Score, Jaccard score and Log loss were computed to determine the performance of all of these models

KNN model was tested with nearest neighbor parameter(k) as 1 to 9. Based on the accuracy score, k was set to 2. Decision tree model was tested with depth 1 to 50. Final value of 2 was selected based on accuracy. Various classifier like linear, poly and sigmoid was evaluated for SVM model. "poly" performed better than the others considering F1 and Jaccard scores.

5. Results

There were only minimal differences in performance between all 4 models created. On an average they had a Jaccard score of 66% and F1 score of 56.5%. Log loss scores are applicable only to Logistic regression model.

Overall, Logistic regression model performed the best (F1 ~57%, Jaccard ~66%, Log loss ~62%)

The performance scores for each of the models created are listed below (Table 1).

Table 1. Performance of classification models. Best scores are highlighted in red

	KNN	Decision Tree	SVM	Logistic regression
F1 Score	0.57	0.54	0.54	0.57
Jaccard Score	0.65	0.67	0.66	0.66
Log loss	-	-	-	0.62

6. Discussion

All of the classification prediction models built reinforces the hypothesis that there is a strong correlation between prevailing conditions and occurrence of accidents. Of the features available in the dataset, weather, road condition, light condition and time of the day played a significant role in the occurrence and severity of the accidents. One of the key observations is the severity of accidents were higher whenever road conditions were wet. Another unexpected observation was the incidence of accidents on well-lit road conditions compared to a poorly lit one.

7. Conclusion

The model, in its current version, will definitely help in predicting and warning drivers of a potential mishap. DOT & Auto industry can further develop these models to provide an alerting system that can help make our roads safer. The current dataset contained only severity 1 and severity 2 data. It will definitely useful to tune these models further with data containing all levels of severity since severity and impact rises for higher severity (2b, 3) cases.

The model can be further developed by integrating live local weather data like intensity of rain, inches of snowfall, windspeed. This can make this model more holistic and have a higher rate of accuracy in predicting unsafe conditions.