PREDICTING SEVERITY OF ACCIDENT

1 Introduction

1.1 Background:

In our daily routine, we use to move outside of our city for some work or other activities. In cities, where there are markets or some entertainment place, people usually travel with care, which results in less accidents. While traveling outside of the city or on some highway, people travel with speed and show less care because of less traffic or less crowd, which results in accidents. One should be more careful on these roads specifically at junctions. Secondly, drivers should be more careful in certain weather or road conditions to avoid serious accidents which can result in injury or fatality. Due to unawareness or carelessness, people may face accidents or stuck on roads due to some accidents, which waste their lot of time and money ultimately. There is a lot of data available with government about the accidents occurred on these roads. This data can be used to alert drivers through radios or mobile alerts when they are approaching on certain junction with specific road and light condition. We must develop intelligent system for this.

1.2 Problem:

Problem is that the data we have is in raw form and is of previous incidents, so how could we use previous data to predict or forecast something could happen in future. We must gather correct information which can help to build some good model for better prediction.

1.3 Interest:

Drivers would be happy to know if they got alert while on road for chances of accident that could occur on certain road in specific environment and can drive more carefully or could change route if feasible. Secondly govt officials would be happy to cater less accidents if they can tell in advance the drivers about chances of specific incident.

2 Data Acquisition and Cleaning

2.1 Data Source:

Data related to accidents and their severity are available through many online links but here we have chosen the data source that is shared by this course from link "http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.csv". I find it enough for my project, since it has lot of data against any feature.

2.2 Data Selection and cleaning:

There is lot of information available in our data set. First, I focused the columns that are much related to provide us required result or which are easy to be segregated into binary information since our algorithms work better on binary information. Data selection is made on factors that are visible to drivers and the information they will get from alarms could be verified and they are willing to act upon.

Lot of information was found missing in the data and there was some information tagged with 'unknown' or 'other'. Th qis tagged information is of no use as they are not providing any information. So, I simply removed the fields carrying this information to have some useful data only.

2.3 Feature Selection:

I used four features for the input to our model. They are: "ROADCOND", which shows the condition of road, whether it was wet or dry. Another feature is "WEATHER", which shows weather was clear, raining or overcast during certain incident. Third feature I selected is "LIGHTCOND", which describes the condition of light during incident happened. Fourth condition which improves our accuracy of prediction the severity is "JUNCTIONTYPE". Last one is not related to any weather or environment condition, but it helps to get better model having better accuracy. Once driver approaching certain type of Junction, he/ she should be extra careful.

3 Data Analysis

3.1 Target Selection:

We have used SEVERITYCODE as our target feature from our dataset. This column values with respect to severity are:

SEVERITYCODE	SEVERITYDESC
1	Property Damage Only Collision
0	Unknown
2	Injury Collision
2b	Serious Injury Collision
3	Fatality Collision

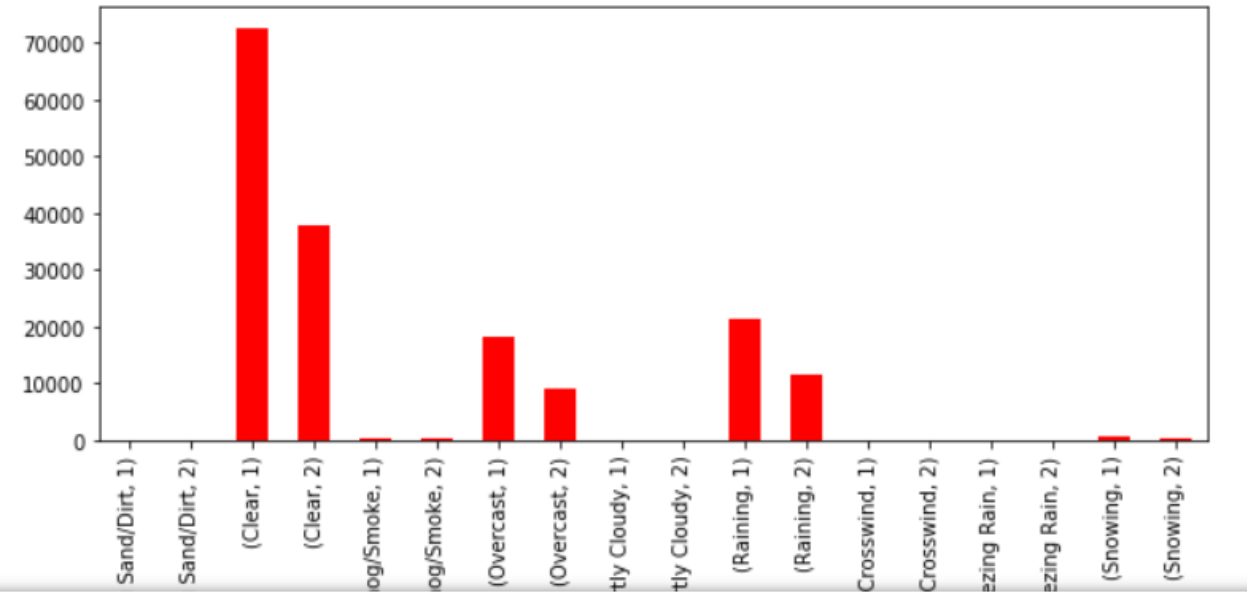
After taking the count of severity in my data, I came to know that some data has very little share in the dataset and putting them for training might underfit our model. They have less than 0.5% occurrence in our overall data. They are “3”, Fatality Collision and “0”, unknown. So, I simply removed these entities from my dataset. Other thing I noticed is injury or severe injury. These both could be put under same umbrella as both shares the information that injury could occur, minor or severe doesn’t matter when there is injury. So, I simply replaced all the values of “2b” with “2” under SEVERITYCODE column. Now we have binary target set, either injury happened during accident “2” or it was just some property damage “1”.

3.2 Relationship between input feature and target feature:

The amount of data and their impact on severity of accident form selected input feature set is depicted in below graphs. They also show the ration between both type of incident.

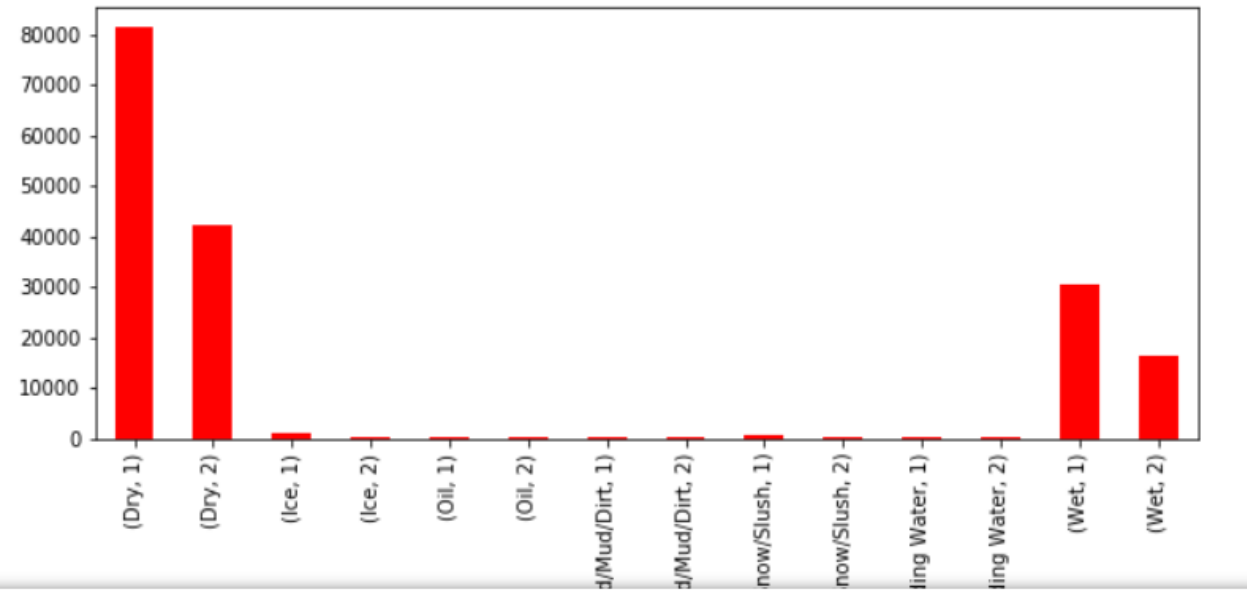
WEATHER:

There is strong relationship between weather and driving norms. Here we have three categories for weather which gave high figures for incident to occur and almost 50 percent of the values are where injuries happened, or they were more severe. This is particularly the reason where people are more careless in these types of weathers.



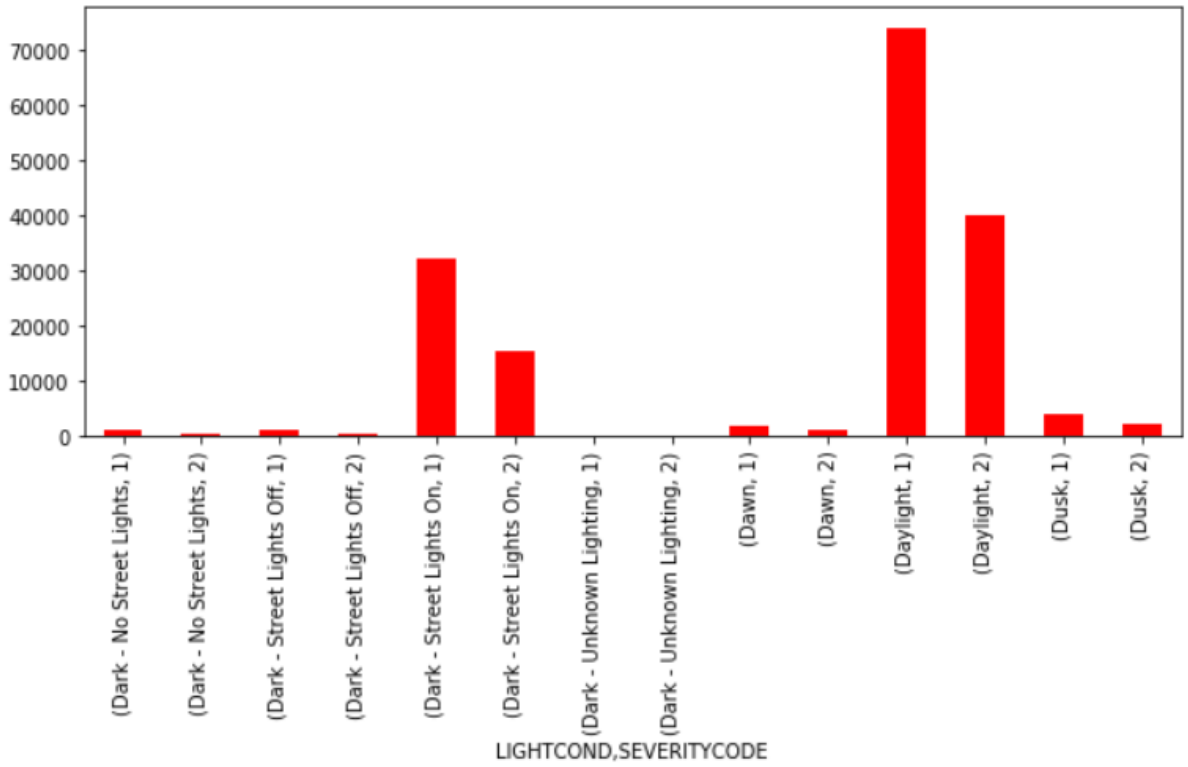
ROADCOND:

Road condition is also a critical factor that drivers should care about. Below stats shows some high accidents ratio for minor or severe categories particularly when road is dry or wet.



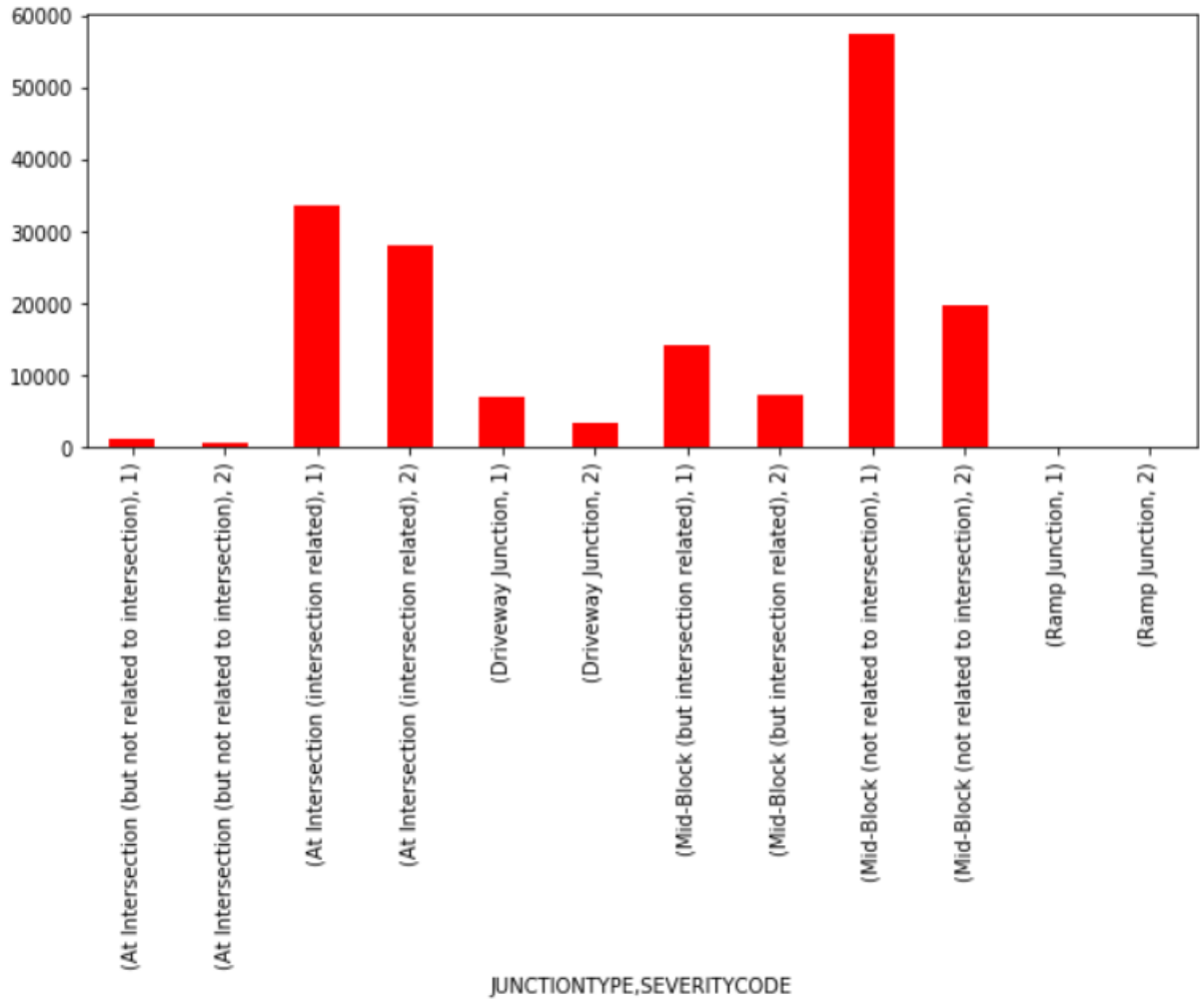
LIGHTCOND:

Light plays vital role in accidents if drivers are not careful. In bright daylight or darkness where there are street lights on, might make drivers more relaxed and they get into accidents. So, one should be more careful in these conditions.



JUNCTIONTYPE:

This feature is not environment related but at certain points there are more accidents as compared to others. This feature with other mentioned above will be helpful for drivers to decide to go through the same rout or might change some.



3.3 Predictive Modeling:

Since the targeted data (SERVERITYCODE) is in binary form, either this or that, so I have used algorithms which works better on two value output. Therefore, I have used two algorithms for modeling. Logistic regression and Decision Tree. For finding accuracy between our predicted output and the output of test sample we used Jacquard Similarity technique, F1_score and accuracy_score technique. I have observed that both models somewhat predicted the target output with 60% accuracy. Prediction has been made whether the result after accident could be minor (property only) or it could be more severe like injury. I used standardized data for logistic regression model while original data used for decision tree as it was in the binary form.

3.4 Results and Observation:

	Jaccard_similarity	f1_score	accuracy_score
Logistic Regression	0.585	0.619	0.585
Decision Tree	0.586	0.597	0.586

Results of all accuracy measures shows almost same values for both models. This shows that there is almost 60% accuracy in the selected data and the target data. Initially I have chosen only weather, road condition and light as independent variables but I observed that with the addition of fourth variable i.e. Junction type we have got some better results.

4 Conclusion:

With the results and data, we have for analysis, it is observed that accident count due to normal weather and normal road condition with normal light status are more as compared to other reasons. This shows that in these circumstances, drivers may show more carelessness and relaxed while driving and face accident. For other reason behind accidents are darkness and wet road with rainy weather. When we put junction types into analysis, we observed that in all these weathers, accidents occurrences are more at some junctions as compared to others. So, drivers should be more careful while arriving at these points to avoid accidents. Or drivers can change their route, if feasible, to avoid certain junctions.