Accident Prediction Report

Bikas Kar

Introduction

In this capstone we are going to solve the problem of finding the severity of the accident based on the weather condition or lighting some other conditions which can lead to accident. In this capstone we will also try to predict the probability of the accident which can be avoided people travelling on the road.

Background of this project is to find the causes of the accident and if we can figure out way to predict an accident and thereby saving precious lives. This prediction will help not only the commuters to be alert while driving, it can be used by traffic management officials to flash important messages using billboards.

Data Description

In this capstone we are going to use the data which is available in this site which contains the various attributes of each accident. there are 37 attributes for each accident. In this the most important one is the Severity code which is what we have to predict Based on other attributes . Some of the attributes which we Can use our location of the accident, lighting conditions, road conditions, Weather conditions, speed of the vehicle etc

We can create a model Which will predict of accident and also the probability and train the model using training data and then further test it with test data which will be a portion of the current data which we have.

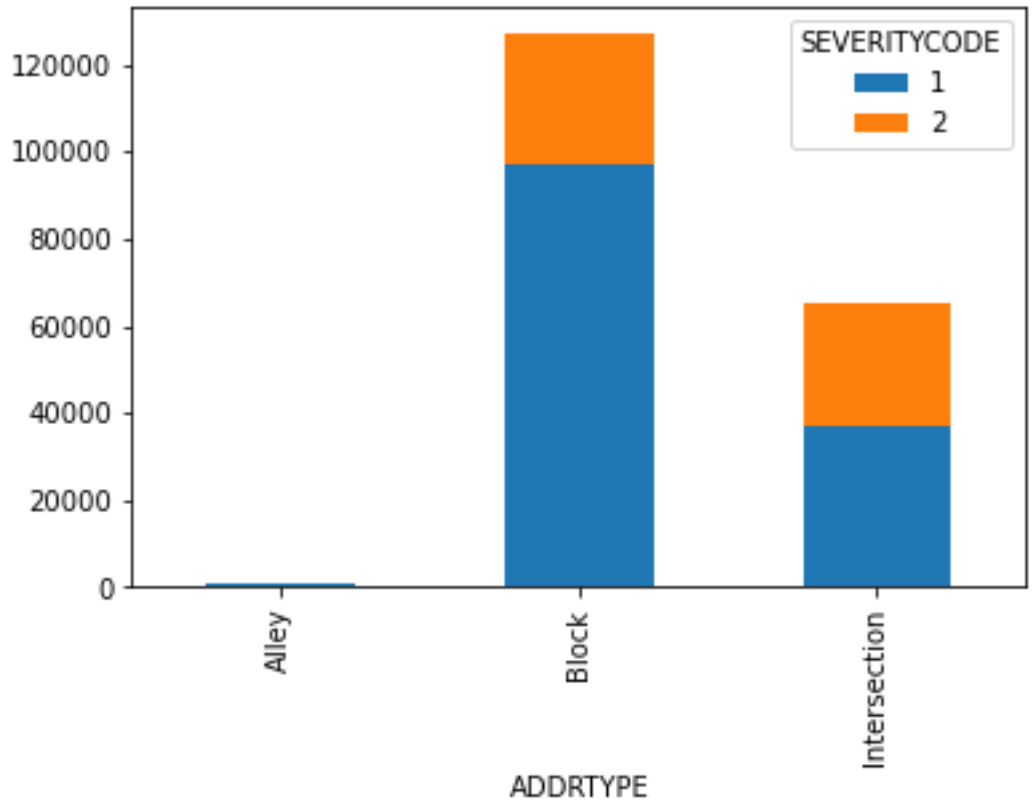
After having a quick look at the data we can see that most of the accidents are either of severity code 2 or 1. For severity code 2, which is an injury most of the accident are happening at the intersection so this is also be a key factor 2 predict an accident.

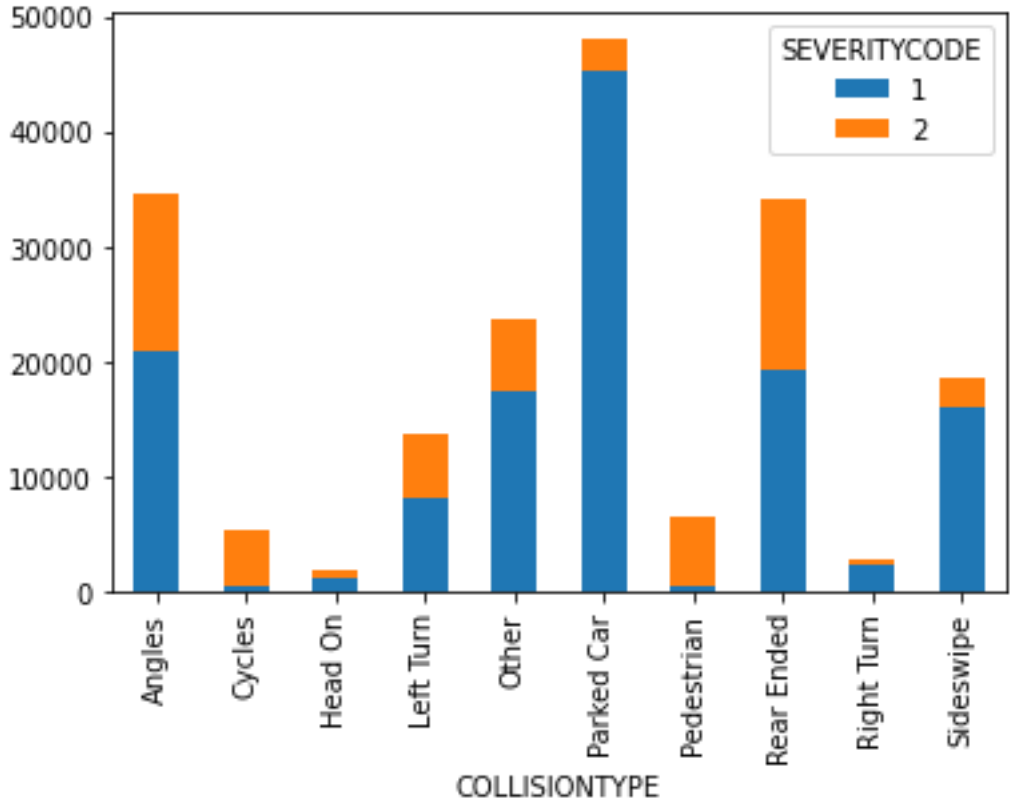
Also we can see that most of the accident at least 50% are happening when they weather condition is not clear. So as you we can have a insight in the current data and predict the severity of the accident

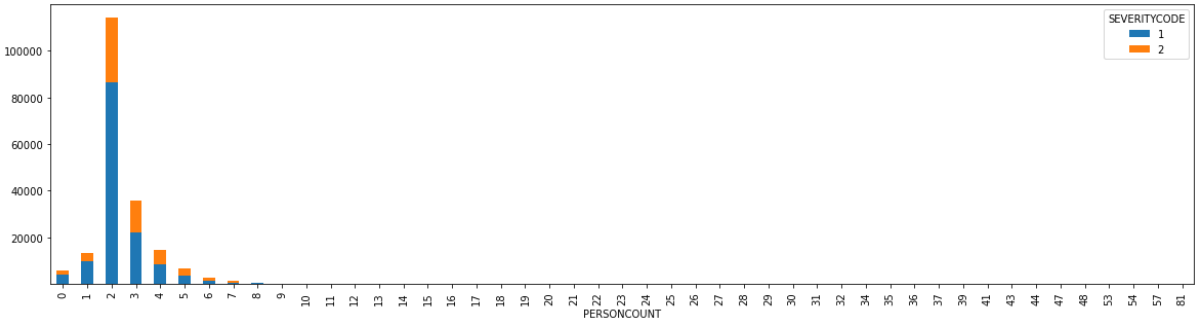
Data Analysis

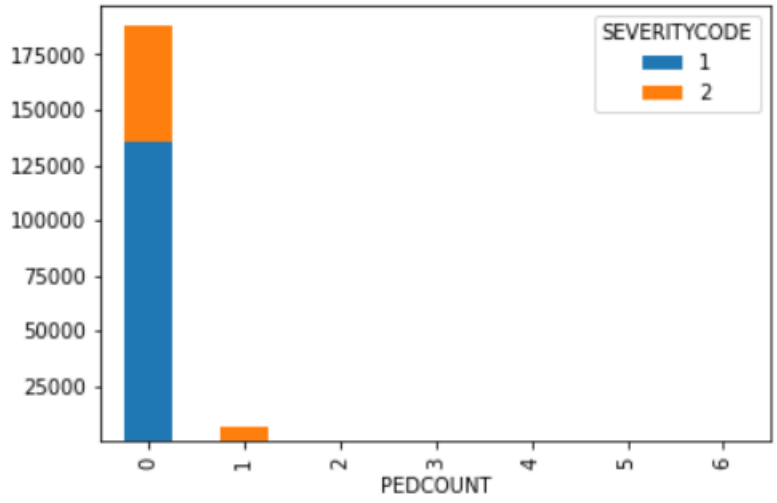
In this section, we will try to understand dataset and look into each column and its values. After this process, we will come to know which columns will be useful for us to use in our model.

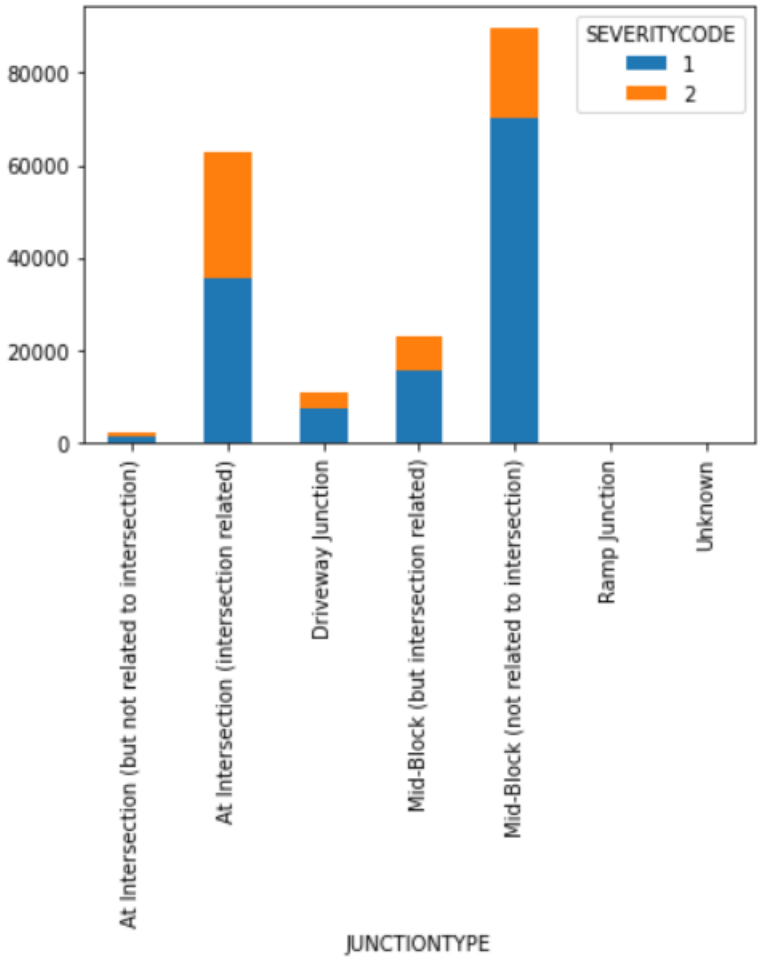
Below is the exploratory analysis done:

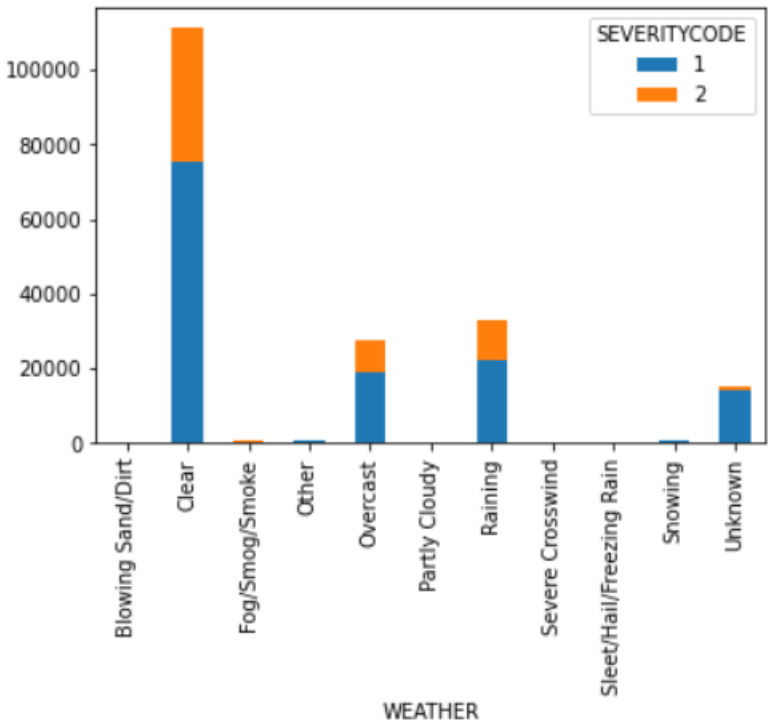


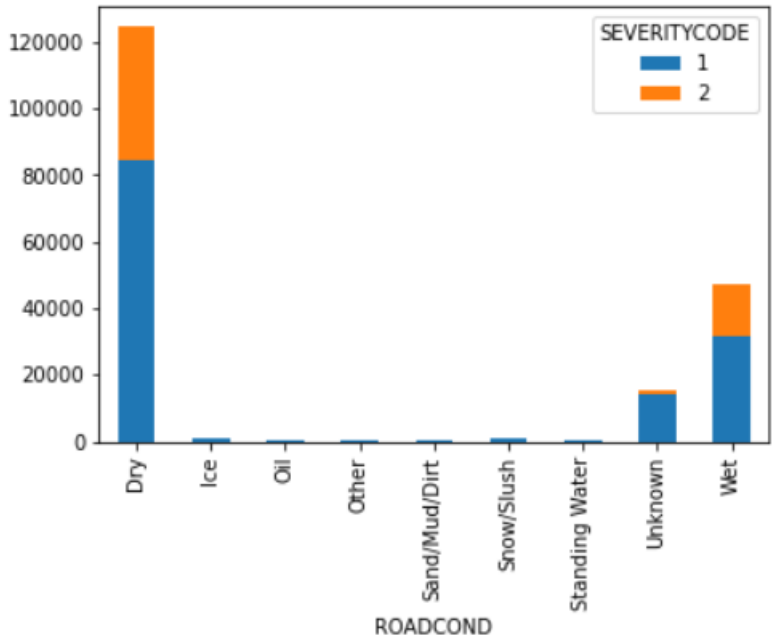

SEVERITYCODE
1
2
50000
40000
30000
20000
10000
0
Angles
Cycles
Head On
Left Turn
Other
Parked Car
Pedestrian
Rear Ended
Right Turn
Sideswipe
COLLISIONTYPE


SEVERITYCODE
1
2
100000
80000
60000
40000
20000
PERSONCOUNT


SEVERITYCODE
1
2
175000
150000
125000
100000
75000
50000
25000
PEDCOUNT


SEVERITYCODE
1
2
80000
60000
40000
20000
0
At Intersection (but not related to intersection)
At Intersection (intersection related)
Driveway Junction
Mid-Block (but intersection related)
Mid-Block (not related to intersection)
Ramp Junction
Unknown
JUNCTIONTYPE


SEVERITYCODE
1
2
100000
80000
60000
40000
20000
0
Blowing Sand/Dirt
Clear
Fog/Smog/Smoke
Other
Overcast
Partly Cloudy
Raining
Severe Crosswind
Sleet/Hail/Freezing Rain
Snowing
Unknown
WEATHER



After looking at all the data, I arrived at following attributes which can be used to predict the severity of accidents:

- ADDRTYPE’,
- 'COLLISIONTYPE’,
- 'PERSONCOUNT’,
- 'PEDCOUNT’,
- 'PEDCYLCOUNT’,
- 'VEHCOUNT’,
- 'WEATHER',
- 'JUNCTIONTYPE’,
- 'ROADCOND’,
- 'LIGHTCOND’,
- 'SEVERITYCODE'

Data Cleaning

Once we understand the data, I need clean the data. I have done the following to clean the data:

1. Replacing Null value with Frequently occurring instances

2. Dropping rows for following attributes having null values or having values “Other” or “Unknown”

Data Encoding

Data needs to be converted into numerical values for ML to work. We have used two methods for data encoding :

1. Label encoding ➔ for the attributes which are having values ordered
2. One Hot encoding ➔ for the attributes whose values do not have any order.

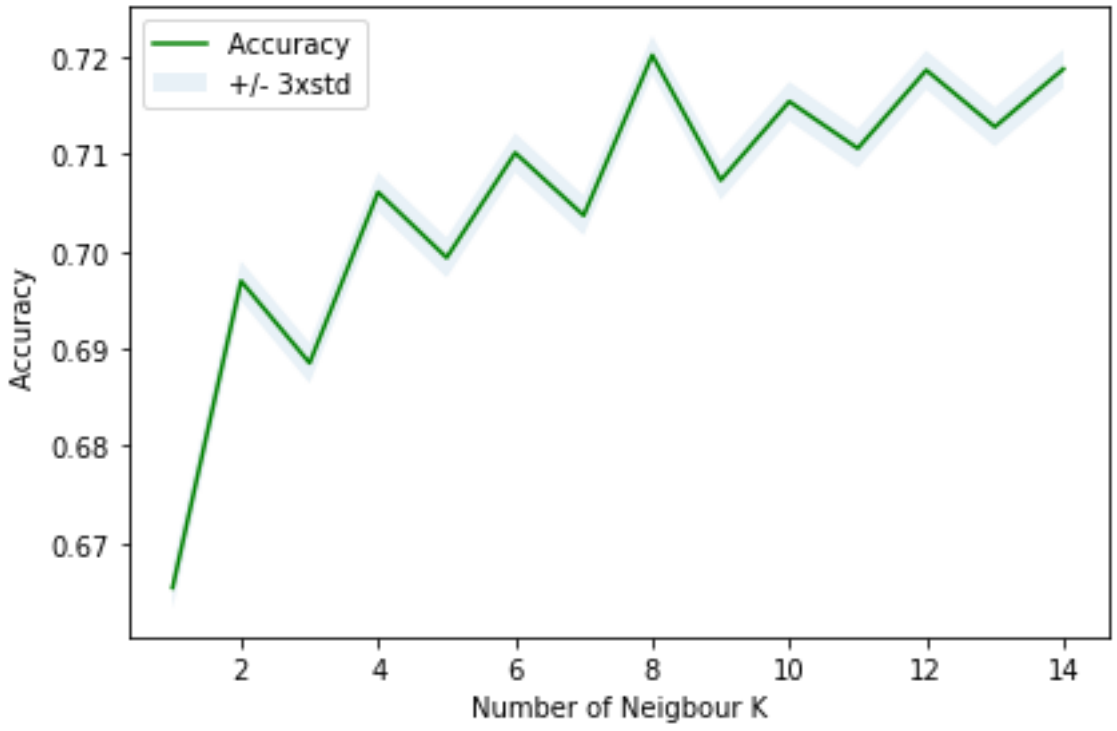
Data Normalization

Data normalization was performed using sklearn. Post that data being split into Training & Test data in 70-30 ratio. In this case, since I have to predict the severity of accident,

Modelling

Following modelling method has been used:

- KNN modeling
    - Best accuracy was 0.7201904306601331 with k= 8



Accuracy & Score

Train Accuracy: 0.7320425481056532
Test Accuracy: 0.7201904306601331

KNN F1-score: 0.69
Jaccard Similarity: 0.7201904306601331

- Decision Tree Modelling

  In this model we have taken the following options:

  - criterion="entropy"
  - max_depth = 400

Accuracy & Score

Train Accuracy: 0.7519336167597193
Test Accuracy: 0.7246922433369188

Decision Tree F1-score: 0.69
Decision Tree Jaccard Similarity: 0.7246922433369188

Result

After model evolution, we for the below are the scores:

KNN F1-score: 0.69
Jaccard Similarity: 0.7201904306601331

Decision Tree F1-score: 0.69
Decision Tree Jaccard Similarity: 0.7246922433369188

Conclusion

Based on results we can say that both models have similar score and accuracy. However decision tree model is taking less duration to process data compared with KNN. So we will go for **Decision Tree modeling** and predict with more accuracy. We hope this will help traffic personals , government officials to reducing accidents and save lives.