Predicting the Accident Severity from Sample Dataset

Shalesh Nath Sharma

August 22, 2020

I. Introduction

1.1 Business Problem

In this project we will predict accident severity based on different features and we will try to detect locations where the accident frequently happen. This report will be targeted to Police officer and driver to make them more wary for these particular places.

2.1 Data

In this project we will use sample dataset "Data-Collision" to solve the problem.

2.2 Data Cleaning

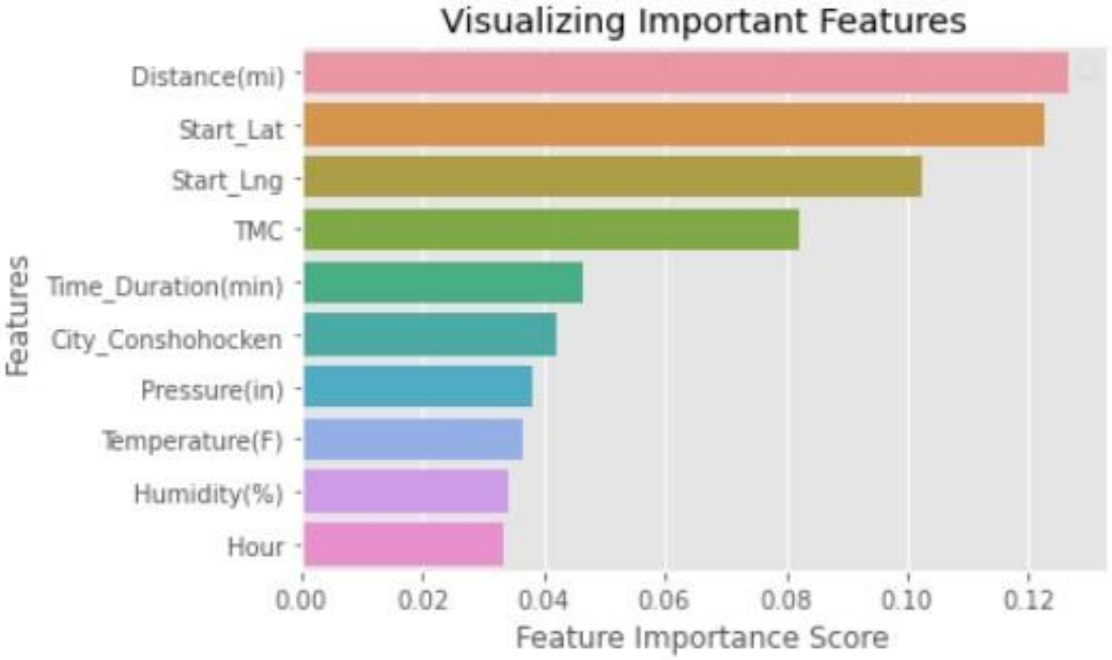
First we drop unrelevant column for our prediction dataset, here is some column that I dropped and the reason I dropped it.

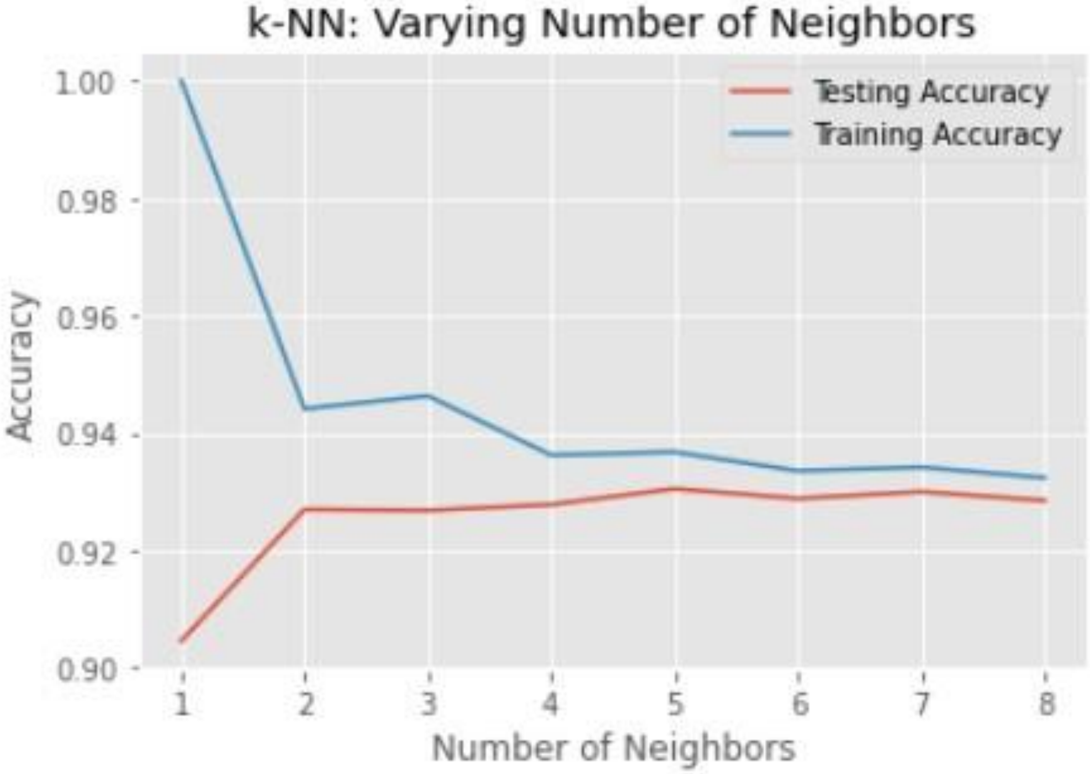
- INCKEY, COLDETKEY, STATUS : I dont know the importance of matching unique key and secondary key, would be appreciated if someone could tell me as I can't find it on internet.
- OBJECT ID, REPORTNO, INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, SDOTCOLNUM : Not relevant to predict Severity
- SEVERITYCODE.1 : Duplicate of SEVERITYCODE
- SEVERITYDESC, SDOT_COLDESC,ST_COLDESC: Description from another columns
- INCDATE,INCDTTM : For simplicity
- SEGLANEKEY, CROSSWALKKEY : Too many unique Value

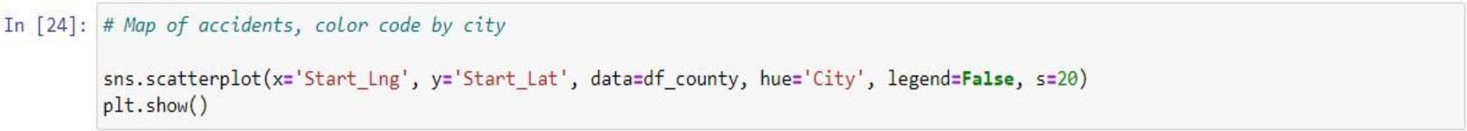
2.3 Feature Selection

We assume Location where collision happen will be the most important feature to predict accident severity. Followed by Road Condition, Light Condition, Weather and Description of collision. Based on our assumption We will focusing on location (Junction type, Collision Type)

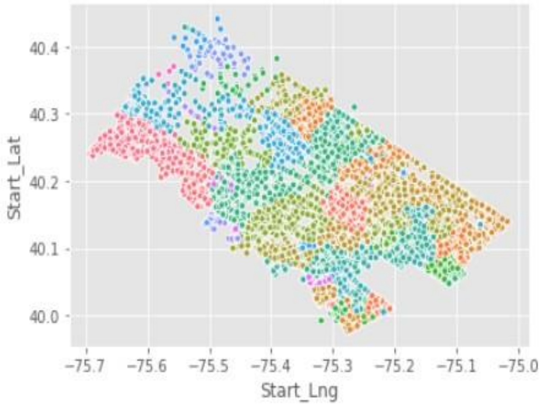
Exploratory Data Analysis

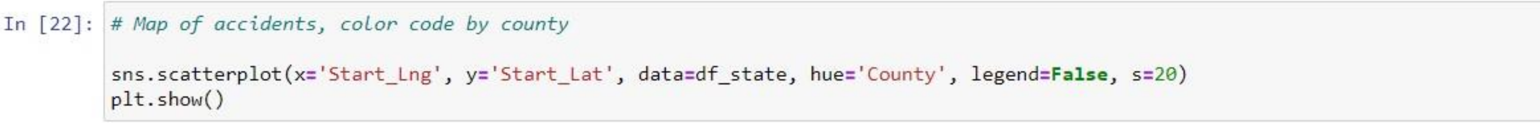

Visualizing Important Features
Distance(mi)
Start_Lat
Start_Lng
TMC
Time_Duration(min)
City_Conshohocken
Pressure(in)
Temperature(F)
Humidity(%)
Hour
Features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Feature Importance Score


k-NN: Varying Number of Neighbors
Testing Accuracy
Training Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
Accuracy
1
2
3
4
5
6
7
8
Number of Neighbors

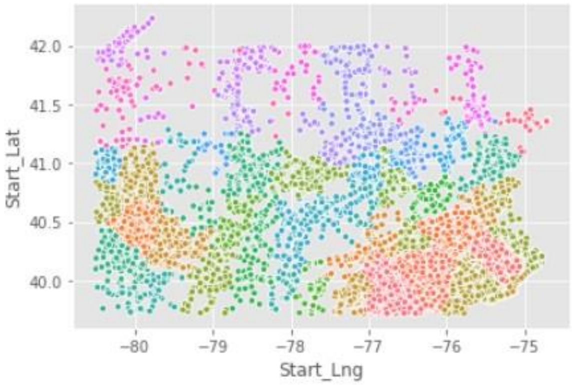


```
In [24]: # Map of accidents, color code by city

         sns.scatterplot(x='Start_Lng', y='Start_Lat', data=df_county, hue='City', legend=False, s=20)
         plt.show()
```





```
In [22]: # Map of accidents, color code by county

         sns.scatterplot(x='Start_Lng', y='Start_Lat', data=df_state, hue='County', legend=False, s=20)
         plt.show()
```



2.4 Target Feature : Severity

This is the variable we want our Machine Learning Model to predict
Note : 1=Property Damage Only Collision, 2=Injury Collision

2.5 Address Type

Collision address type is whether it is Block or Intersection

From the graph property damage collision more likely happen in Block, whilst Injury collision more likely happen in Intersection

2.6 Vehicle Count

From the graph we can see that if only 1 Vehicle involved in collision most likely it was a injury collision, while there is 2 most likely a Property only collision, this is probably true when 2 vehicle collision it might be happen because someone crashed a parked car, while if only 1 vehicle involved there is probability that another person involved in crash.

2.7 Inattention ID

Whether or not collision was due to inattention. (Y/N)

From the graph I can't see any insight as the number of inattention people when collision happen more or less have the same ratio.

2.8 Under Influence

Same from the graph before, I can't see any insight as the number of people under influence of drugs or alcohol when collision happen more or less have the same ratio.

2.9 PEDROWNOTGRNT

Whether or not the pedestrian right of way was not granted. (Y/N)

From the graph we can see that when the pedestrian right of way was granted and there is a collision its most likely an injury collision.

2.10 ST_COLCODE

A code provided by the state that describes the collision.
From the graph we can see that number 32 contribute a lot to property damage only collision while number 1,2,3, 45 contribute a lot to injury collision, what this number supposed to be? From description on here, 32 = One Parked - One Moving ; 1 = Vehicle Going Straight Hits Pedestrian, 2 = Vehicle Turning Right Hits Pedestrian, 3 = Vehicle Turning Left Hits Pedestrian, 45 = Bicycle. And this is make sense because injury collision happen when vehicle hits pedestrian or Bicycle accident while property damage only collision because someone crashed a parked car.

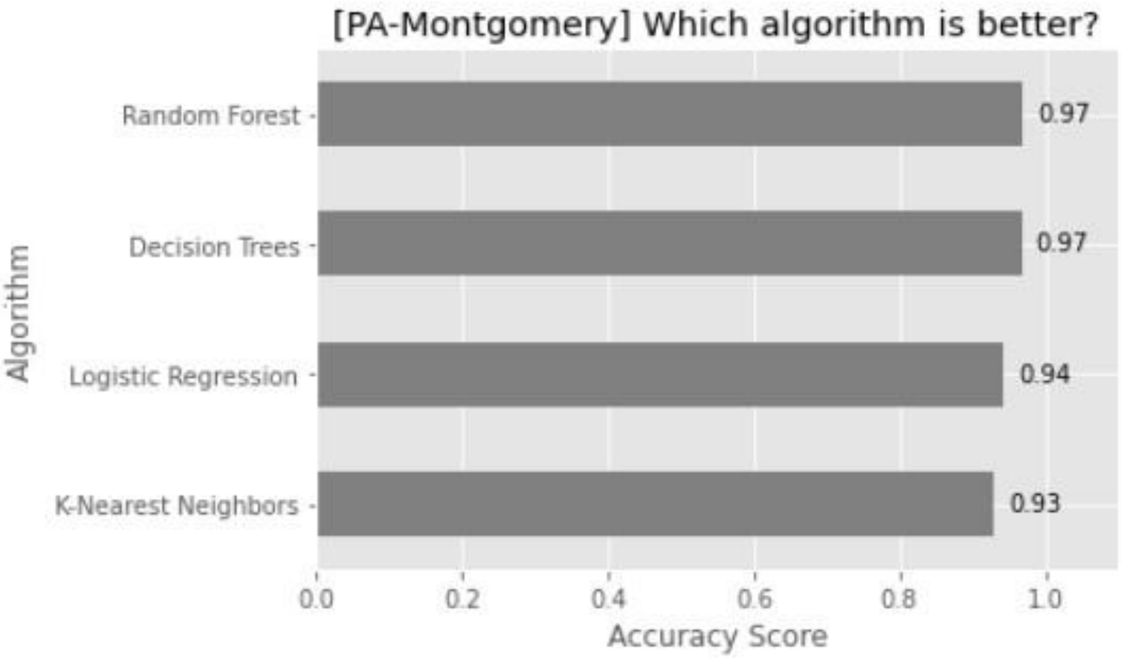
2.11 HIT PARKED CAR

From the graph I can't see any insight as the number of Parked car that was hit when collision happen more or less have the same ratio.

3 Predictive Modelling

For our Model we tried to use these algorithm to get best accuracy : KNN, Logistic Regression, Naïve Bayes, Stochastic Gradient Descent, Linear SVC, Decision Tree, Gradient Boosting Trees, CatBoost. From our Cross Validation Result we get Cat Boost get the best accuracy (75.77%)

Besides accuracy we are showing which feature are the most important in the model



From the graph we can see that COLLISION TYPE feature is the most important feature in this dataset.

4 Results and Discussion

our ML model manage to predict severity of car accident with 76% accuracy, with the most important feature is Collision type with parked car this is make sense because when there is accident with another parked car most likely it will be a property damage collision.

some feature that I dont use such as INCDATE,INCDTTM could be used to predict whether the accident more likely happen at night, day, weekend, weekdays, etc. We could also group up Person Count when the number reach more than 5, and if the target feature has more than 2 variable it probably more useful in real life (such as 3 = fatality, 2 = serious injury, 1 = injury, 0 = property damage).

5 Conclusion

Purpose of this project was to predict severity of car accident from sample data that I get from Coursera Capstone, This project is targeted to police officer or any interested stakeholder, Algorithm that I used was CatBoost Algorithm which give our model accuracy of 76%, top 3 feature that is important is Collision Type,State Collision Code and Person Count.