Predicting Severity of Car Accident

By – Ameya Shrikant Dalvi

Date – 21st Sept 2020

Problem Statement

Car/Vehicle accidents are a significant source of deaths, injuries and property damage. Accidents often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. It is a major concern for public health and traffic safety. Managing of car/vehicle accident is essential to mitigate accident impacts and improve traffic safety and transportation system efficiency. Accurate predictions of severity can provide crucial information for emergency responders to evaluate the severity level of accidents, estimate the potential impacts, and implement efficient accident management procedures.

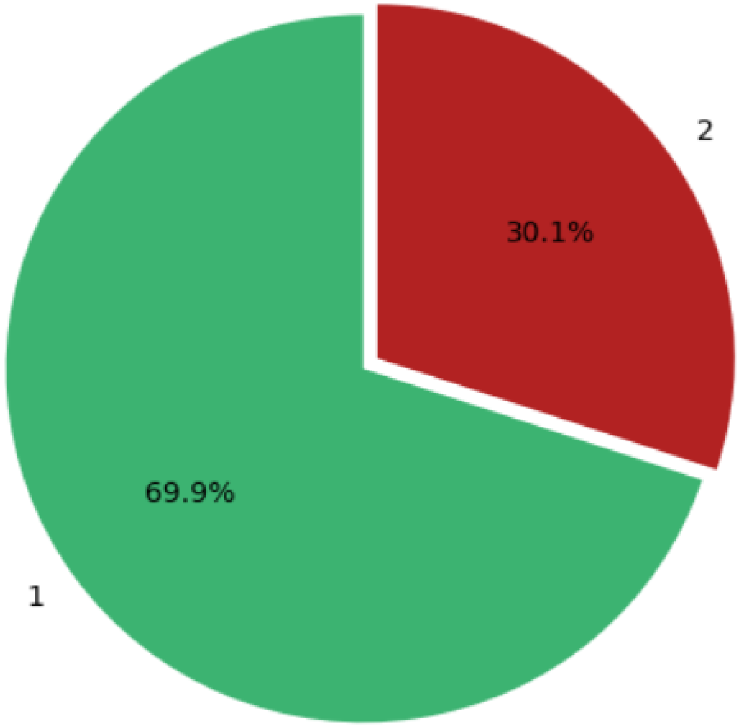
This data is taken from Seattle government and hence audience will be Seattle government, police, rescue groups and insurance companies.

Data Acquisition and Cleansing

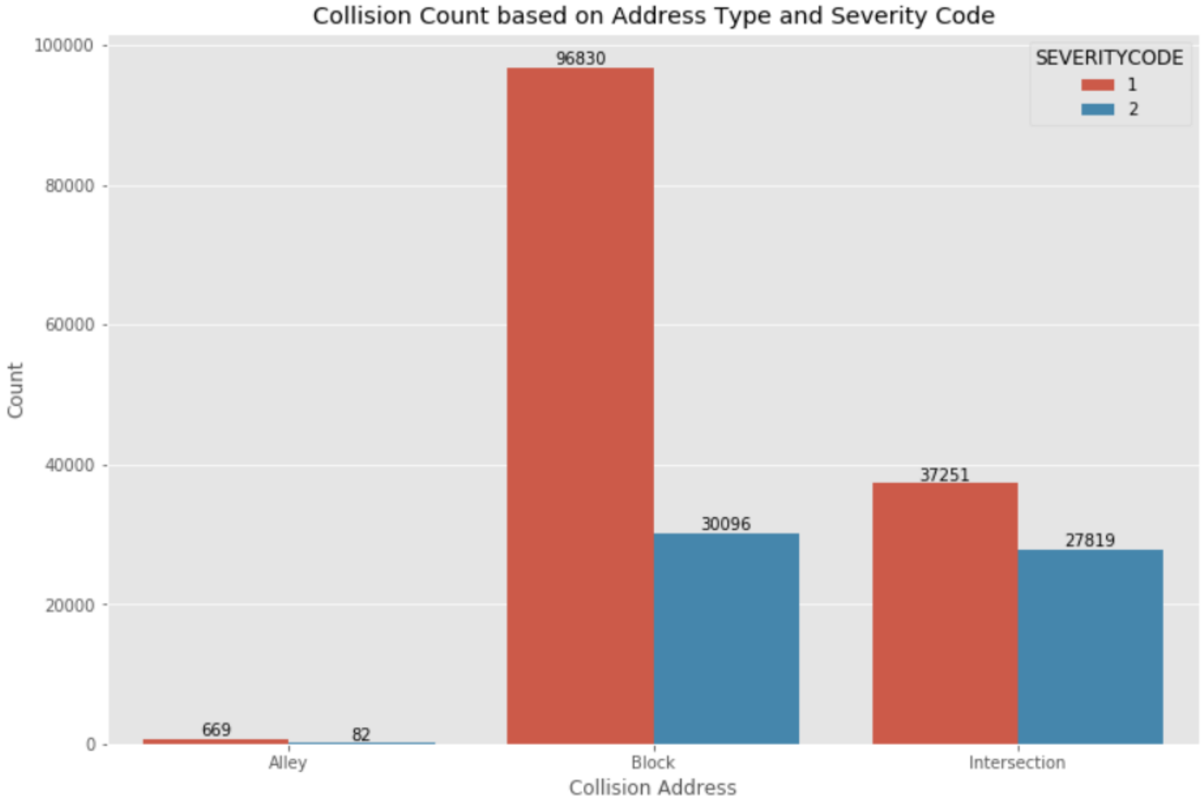
- Data was taken from the site - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
- It was observed that lots of data fields are empty.
  - SPEEDING fields has data for only 9333 records out of 194673 total records.
  - UNDERINFL fields is not correctly maintained we have data as – **N, 0, Y, 1.** So we are not sure if **N** means **no** and **Y** means **yes**. Or **0** means **no** and **1** means **yes**.
  - Hence fields where we have lots of missing value or fields where data is not correctly maintained are not considered when developing the model.
- Skewed data is given to us where cases of non sever incidents >>> cases of sever incident. Hence down sampling was done to remove the skewness.

Exploratory Data Analysis

- Skewness of data.
  - Non sever cases (1) – 134750
  - Sever cases (2) - 57997

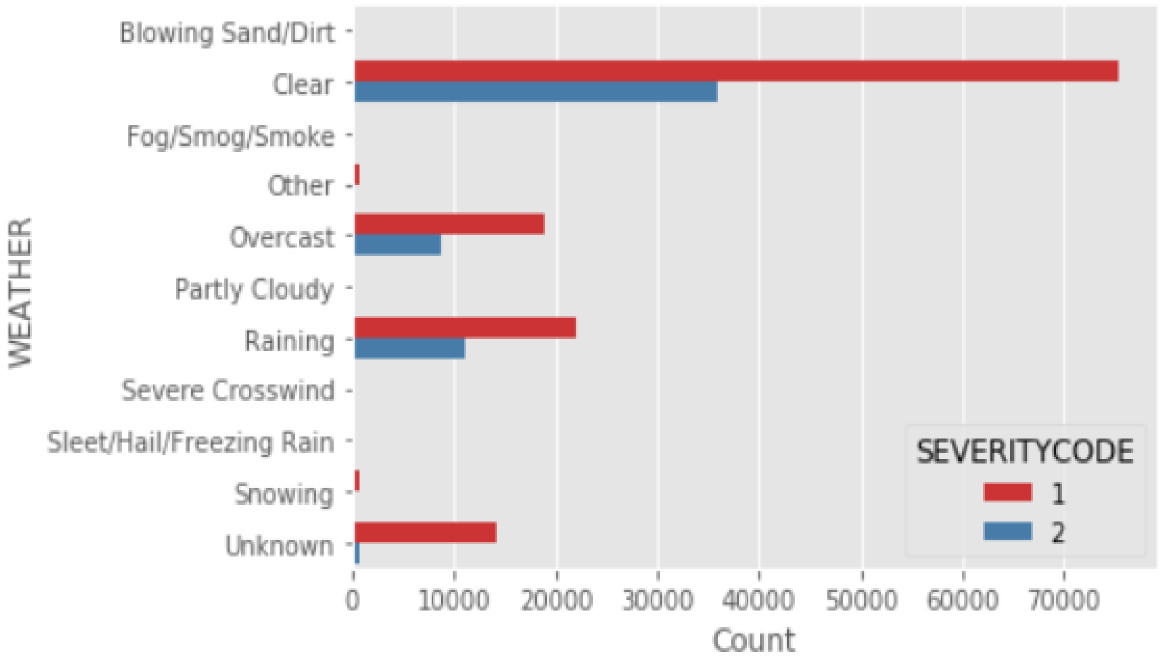


Exploratory Data Analysis…



- ADDRTYPE – Address type where accidents occur
- Majority of accidents happen at address type – **Block**
  - If accident happens at **Block** there is *~25%* probability that it will be sever
- Next address where accidents occur the most is at – **Intersection**.
  - There *42.8%* chance that an accident at Intersection will be sever

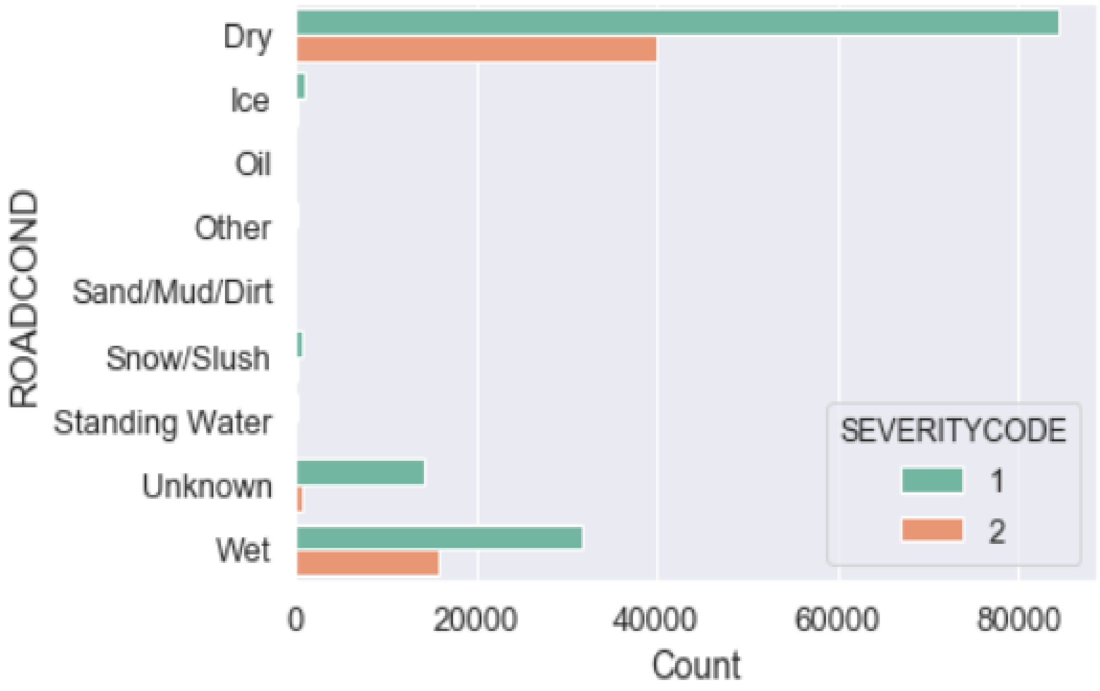
Exploratory Data Analysis…



Based on Weather we can say that most of the accidents takes place on

- **Clear Weather - 58%**
- **When it is Raining - 17.5%**
- **Overcast Conditions - 14.6%**

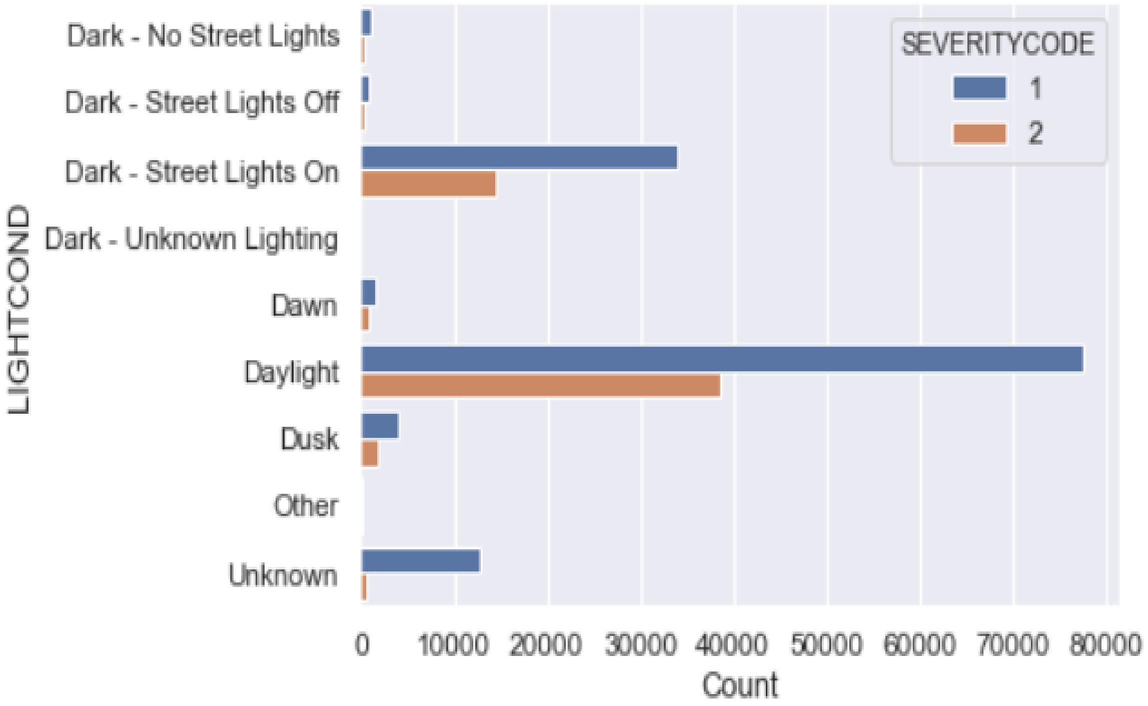
Exploratory Data Analysis…



Road Conditions

- Majority of accidents takes place on **Dry** road conditions followed by **Wet** and **Unknown** road conditions. If we consider only these three road conditions we can see that - **Dry** and **Wet** Road Conditions constitute 98% of all the sever Road accidents.
  - 70.8% of Severe accidents takes place on **Dry** road conditions.
  - 27.9% of sever accidents occur on **Wet** road conditions.

Exploratory Data Analysis…



Light Conditions

Similar to Data fields that we have seen till now majority of accidents are taking place in only under certain conditions. For Lighting Conditions majority of accidents are taking place in **Daylight** followed by **Dark - Street Lights on**.
I was expecting that more sever accidents would take place when light conditions are bad or not favorable like - *Dark - No Street Lights* or *Dark Street Lights off*. But from data we do not find any strong evidence suggesting the same.

Predictive model

For this problem I have used classification models to predict if the car accident will be sever or not. Also, I have used f1 score to determine which model is the best.

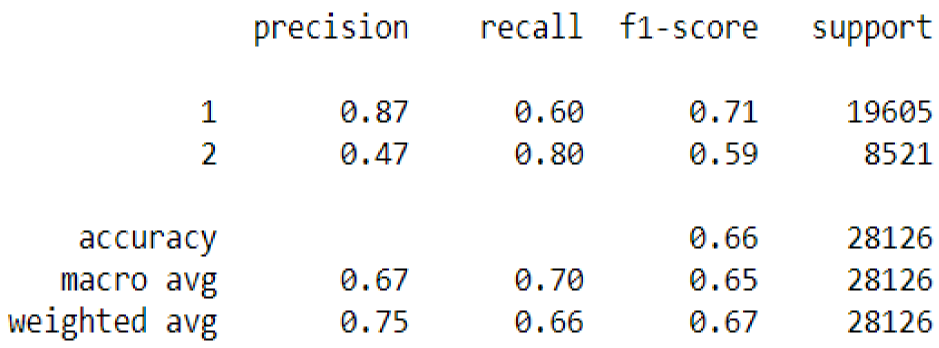
I have used 4 models to predict the outcome of the accident.

Predictive model..

- **Logistic regression** – It is one of the simplest model for classification. With Logistic regression I was getting a f1 score of 0.66 for accuracy.
- **SVM model** – I thought that SVM should give me the best results, but that was not happening. Also when using grid search with SVM model I faced lots of issues and my system also crashed. Hence I did not presume much with SVM.
- **Random Forest Model** – This is an ensemble model and gave me the best result while consuming less computer resources. The improvement was there in f1 score of accuracy, but it was not that significant when compared to logistic regressions.
- **XGBoost** – I tried using XGBoost, but again I faced the same issue that I have observed with SVM. System crashed a lot and it consumed lot of computer resources. Also I did not see any improvement in the f1 score of accuracy when compared to Random forest

Predictive model..

- Logistic Regression

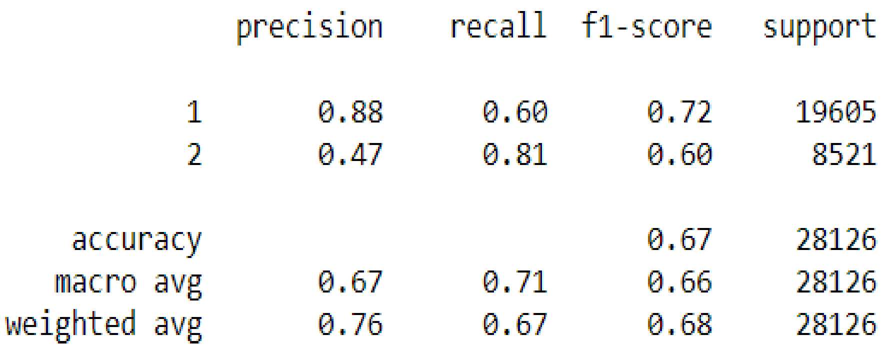


```
              precision    recall  f1-score   support

           1       0.87      0.60      0.71     19605
           2       0.47      0.80      0.59      8521

    accuracy                           0.66     28126
   macro avg       0.67      0.70      0.65     28126
weighted avg       0.75      0.66      0.67     28126
```

- Random Forest



```
              precision    recall  f1-score   support

           1       0.88      0.60      0.72     19605
           2       0.47      0.81      0.60      8521

    accuracy                           0.67     28126
   macro avg       0.67      0.71      0.66     28126
weighted avg       0.76      0.67      0.68     28126
```

Conclusion

- Based on the data that I had, it is difficult to predict accurately if an accident will be sever or not. The model that I developed has an accuracy of 67%.
- Also some of the important factors that we assumed to be cause of road accidents like driving under influence of drugs or alcohol, inattention of driver, over speeding etc. where not recorded correctly or were missed completely.
- In order to improve the accuracy of model more features would be needed, and hence more data needs to be captured.