Capstone Project for Car Accidents

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.

Data acquisition

I have used the data given as part of the capstone project - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

Metadata can be found on this link - https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf.

Data Cleansing

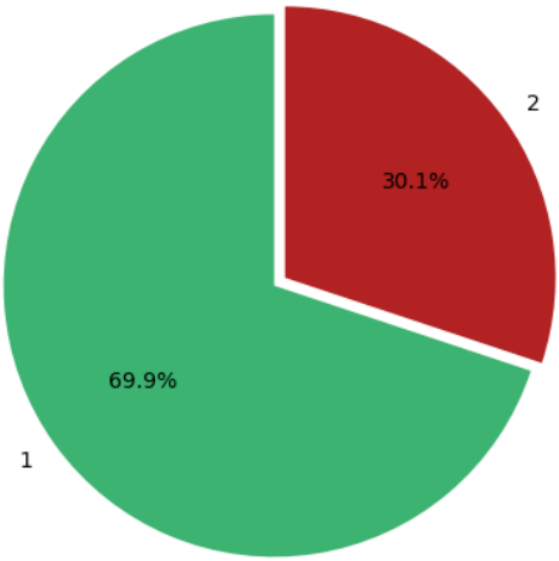
We found lot of discrepancies between the data and the metadata mentioned. As per metadata we have lots of fields available, but very few fields are available in the csv file. It is also observed that some columns have very few data and lots of null values eg – SPEEDING field, where we have data only for 9000 records out of 190000 records. Similar cases are observed with few other fields and hence these fields were dropped

```
COLDETKEY          194673 non-null int64
REPORTNO           194673 non-null object
STATUS             194673 non-null object
ADDRTYPE           192747 non-null object
INTKEY             65070 non-null float64
LOCATION           191996 non-null object
EXCEPTRSNCODE      84811 non-null object
EXCEPTRSNDESC      5638 non-null object
SEVERITYCODE.1     194673 non-null int64
SEVERITYDESC       194673 non-null object
COLLISIONTYPE      189769 non-null object
PERSONCOUNT        194673 non-null int64
PEDCOUNT           194673 non-null int64
PEDCYLCOUNT        194673 non-null int64
VEHCOUNT           194673 non-null int64
INCDATE            194673 non-null object
INCDTTM            194673 non-null object
JUNCTIONTYPE       188344 non-null object
SDOT_COLCODE       194673 non-null int64
SDOT_COLDESC       194673 non-null object
INATTENTIONIND     29805 non-null object
UNDERINFL          189789 non-null object
WEATHER            189592 non-null object
ROADCOND           189661 non-null object
LIGHTCOND          189503 non-null object
PEDROWNOTGRNT      4667 non-null object
SDOTCOLNUM         114936 non-null float64
SPEEDING           9333 non-null object
ST_COLCODE         194655 non-null object
ST_COLDESC         189769 non-null object
SEGLANEKEY         194673 non-null int64
CROSSWALKKEY       194673 non-null int64
HITPARKEDCAR       194673 non-null object
dtypes: float64(4), int64(12), object(22)
memory usage: 56.4+ MB
```

Exploratory Data Analysis

Skewed Data

Firstly, it was observed that we have skewed data set where the number of records for non-sever accidents were more than sever accidents.

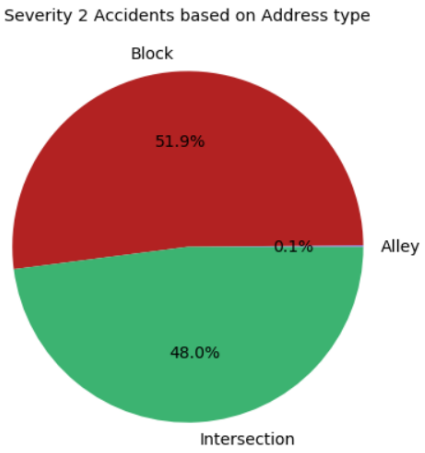


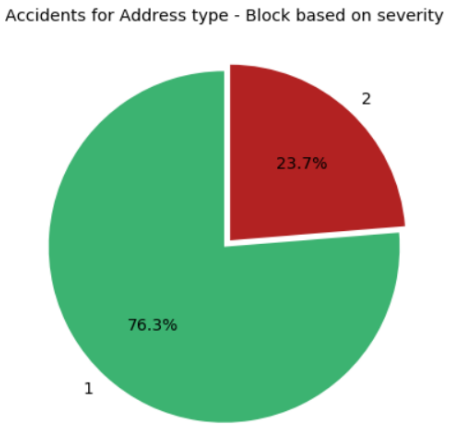
Secondly, majority of accidents happen at address type – **Block**

- If accident happens at address type 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

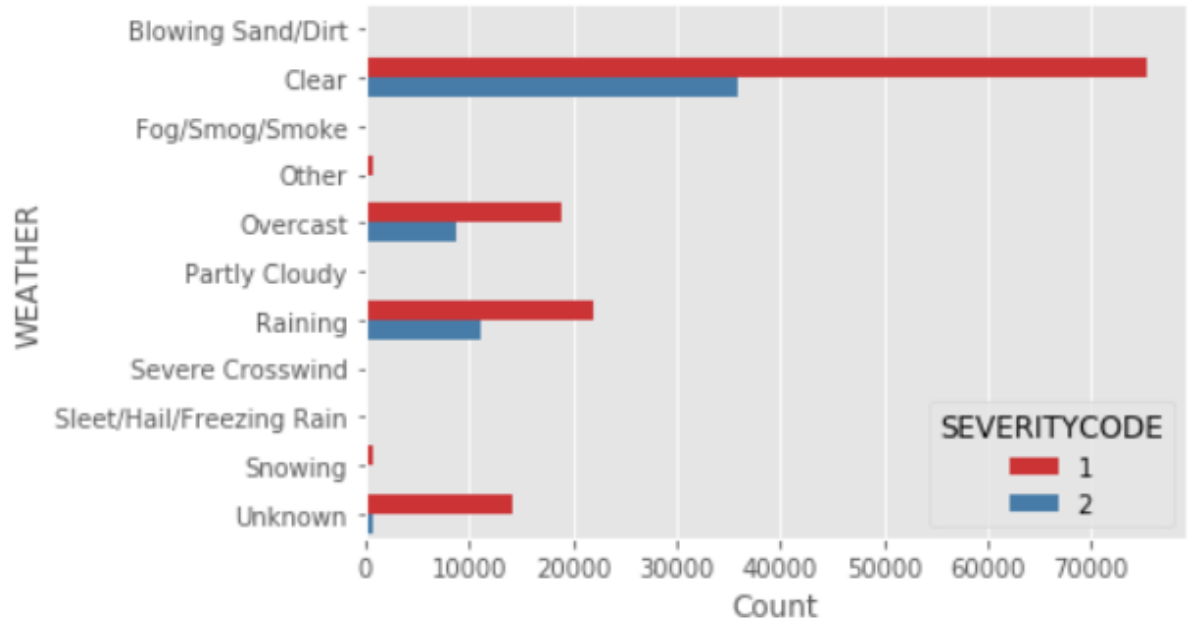




Weather conditions

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

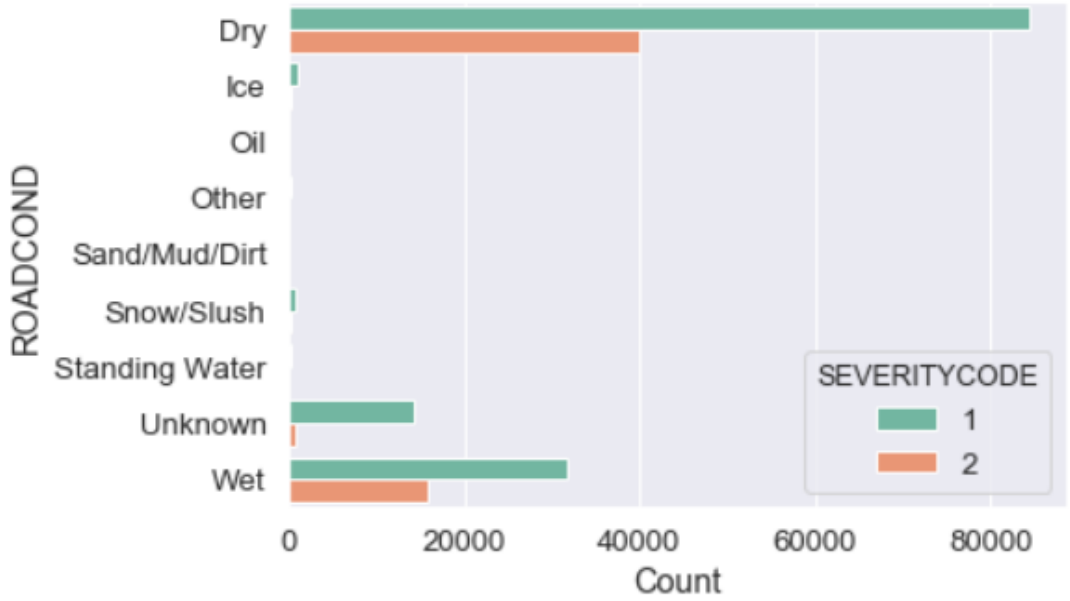
1. **Clear Weather - 58%**
2. **When it is Raining - 17.5%**
3. **Overcast Conditions - 14.6%**

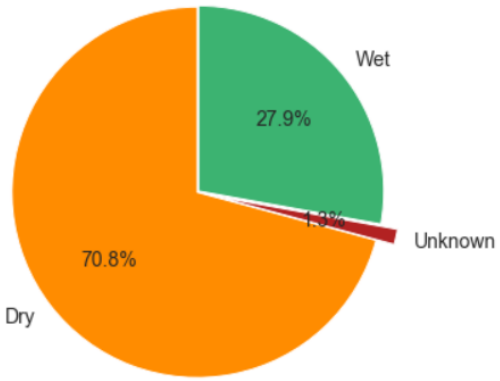


Road Conditions

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.

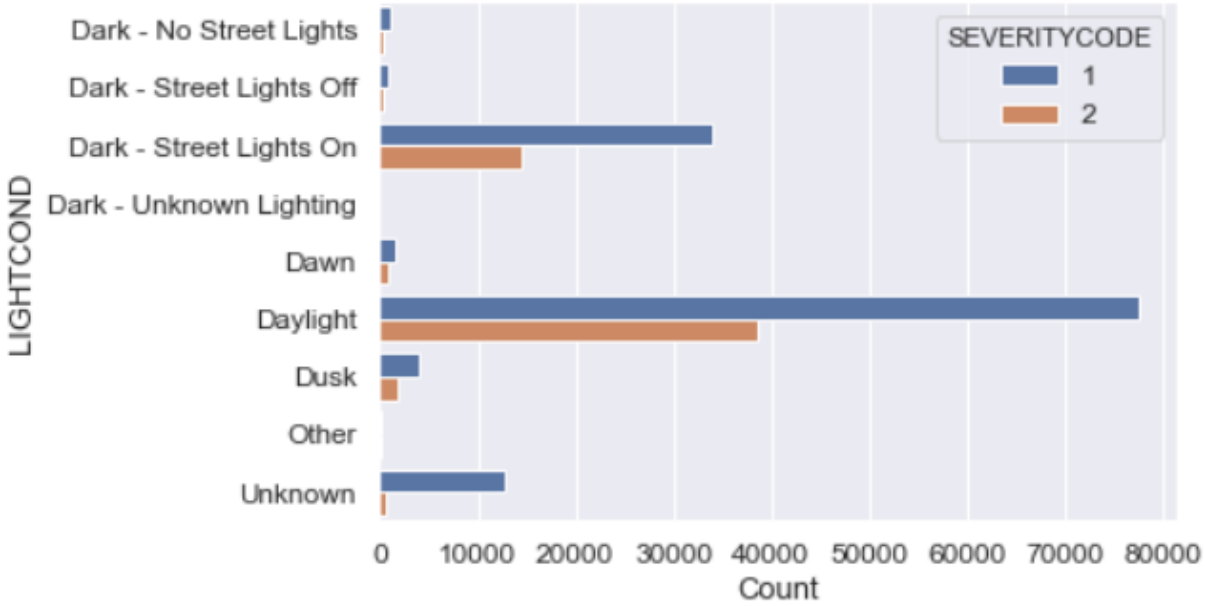
1. 70.8% of Severe accidents takes place on **Dry** road conditions.
2. 27.9% of sever accidents occur on **Wet** road conditions.





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 Modeling

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.

1. 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.

	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

2. 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.
3. 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.

	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

4. 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.

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.