Predicting Accident Severity

Syed Muhammad Sherjeel

October 18,2020

1. Introduction

1.1 Background

Accidents are some of the worst disaster one can encounter.It not only cost global economy US$1·8 trillion (constant 2010 US$) in 2015–30.Not limited to this,these occurrence also cause loss of multiple precious lives every year.Many people lose their loved ones due to these unprecedent happenings.Lastly,many travellers also suffer delays in their commute due to uncertain events.Therefore,being able to accurately predict accident occurrence we will not only save economy trillions of dollars but will also be able to save many precious lives.Lastly,we might also be able to make commute safer.

1.2 Problem

Data that might contribute to occurrence of accident or severity of accident are road conditions,light conditions,location,Humidity, weather ,speeding.This project aims to predict what would be severity of accident under different circumstances.

1.3 Interest

Obviously,government leaders,Logistics Companies and those who commute very often would be very interested as it could potentially save them trillions of dollars alongside a great number of humman lives.

2. Data acquisition and cleaning

2.1 Data sources

Data set that we used in this project was provided by Seatle SPD.This data set was hosted on IBM cloud.From there data was downloaded.It was provided alongside Meta data making the process of understanding the data very feasible and convinient.

2.2 Data Cleaning

This data set provided by SPD had a lot of missing values.It contained many redundant things and values that were not of much of our use.So we had a lot of cleaning to do.

First step was to convert all the categorical variable to numerical variable.Machine Learning algorithms operate on numerical data only but fails on categorical data.Therefore,this step was very much important.
First step was to check for how much data is missing and what is the percentage of missing data.This was done to check if we should too much data is missing I.e, >60 % missing data means this feature needs to be dropped.
We also had to check if data was MCAR,MAR,MNAR.If it was missing completely at random then we can impute then using back fill or forward fill.But if it is missing in a systematic manner then this could cause huge trouble in our calculations.
It turned out that variables like speeding,Inattention and few more had more then 75% of data missing.This could cause a lot of bias in our result. These features were dropped.Remaining features were filled using backfill method.

2.3 Feature Engineering

After data cleaning data,there were around 197000 rows and >40 features.Many of these features were irrelevant such as registration ID,Incident Key etc.
Features that were dropped are,

- ST_COLCODEText : code provided by the state that describes the collision.
- ST_COLDESCText : description that corresponds to the state's coding designation.
- SDOTCOLNUM : A number given to the collision by SDOT.
- SDOT_COLDESC : A description of the collision corresponding to the collision code
- SDOT_COLCODE : A code given to the collision by SDOT.
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- LOCATION : Description of the general location of the collision
- SEVERITYDESC : detailed description of the severity of the collision.
- COLLISIONTYPE : Collision type
- INJURIES : The number of total injuries in the collision. Thisis entered by the state.
- SERIOUSINJURIES : The number of serious injuries in the collision. This is entered by the state.

- FATALITIES : The number of fatalities in the collision. This is entered by the state.

Features mentioned above are not of that much value to us.Since most of them are registration ID issued by states and remaining indicate aftermeth of accident which goes against our planned agenda which is to prevent any such incident from occurring.So they are not of that much usage in our case scenrio.

2.4 Dealing with imbalanced data

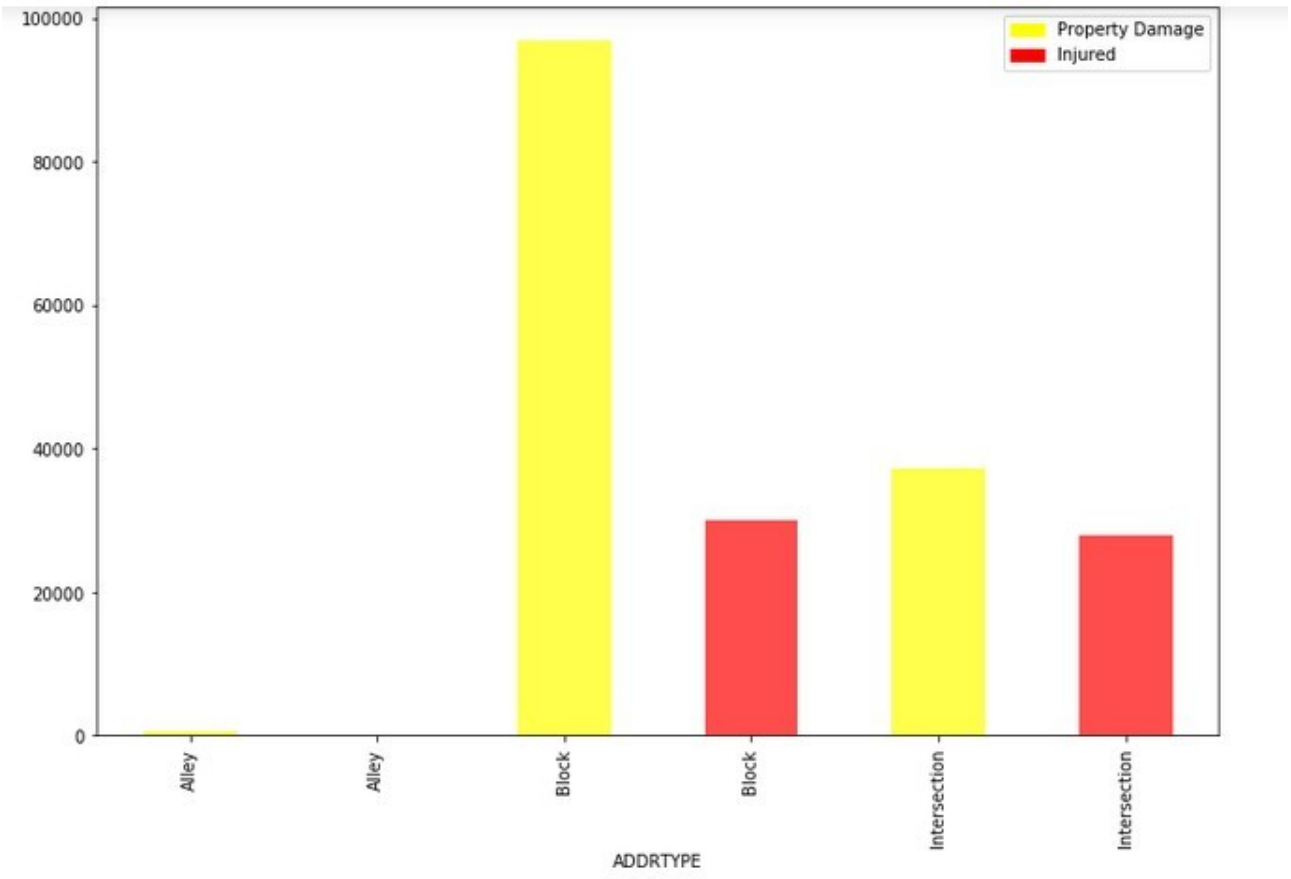
Dataset we got from SPD was very imbalance.The ratio between one class to other was almost 1:3.Initially when we developed a model,it produced a very biased result in favour of majority class.Obviously,this was meant to happen with such a biased dataset.Now there were limited options in regards to dealing with imbalanced dataset.Outcome was to down sample majority class to match the minorty.This imporved the overall performance of dataset.

3. Exploratory Data Analysis

3.1 Relationship between Accident and location

While analyzing data to find relationship between accident and location where accident took place in order to check the correlation between the two,following conclusion was drawn.

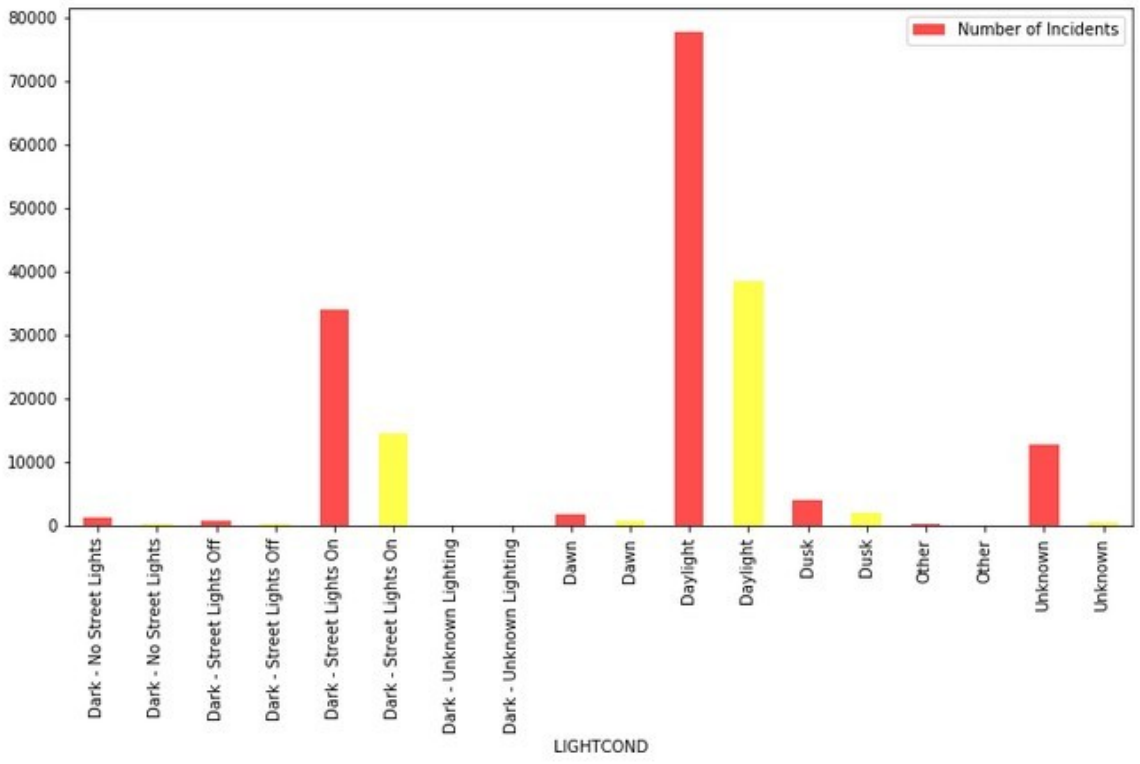
- If one is traveling through block he is more likely to face unprecedent circumstances then while traveling through intersection or alley.
- However same is opposite for someone traveling through Alley.Probability of him being involved in an accident are minimal
- If involved in an accident,one is more likely to face minor property damage then injuries.
- Accident severity is greater in intersection then in Blocks,mainly due to high speed and multidirection traffic in intersection.On the other hand,blocks have certain speed limits and therefore damage is less fatal.



3.2 Relation between Light condition and accident

It is widely accepted that more fatal accidents are likely to occur in poor light condition and less accidents will occur in good light conditions.But while analyzing the data,we got opposite insights.Following were our conclusion,

- More fatal accidents occur in Daylight then in dark or poor lighting conditions.This is mainly due to the fact that drivers become careless when light conditions are good
- However less accidents occur in poor light conditions,this can be explained by the fact that drivers drive more carefully when light conditions are poor.
- One interesting fact was that accidents associated with sun light are more dangerous compared to accidents due to location type.
- As graph indicates that probability of such accidents ending up in fatalities is high then minor property damage.

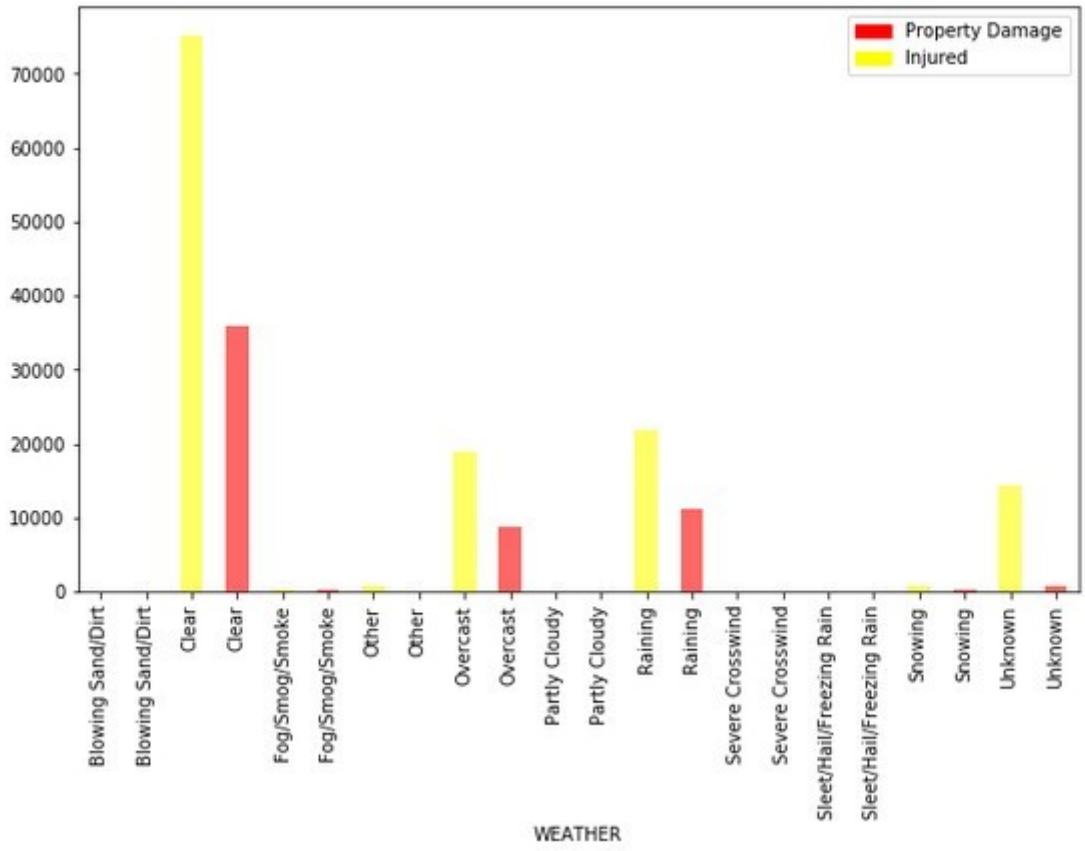


3.3 Relationship between weather and Accidents

A more general perception is that bad weather leads to pilling up of accident cases or impact the severity of accident.while it does suit the common sense but it is not always true.

Following were our conclusion,

- In a severe weather,weather and radio stations issue a lot of warnings which make drivers cautious and drivers deploy all sort of safety manners to protect them selves.
- In a clear weather,there is no such thing as warning being issued by state or stations.So, drivers become less cautious compared to terrible weather.
- However in case of weather accident severity is opposite to what was in case of Light conditions.
- One involved in an accident during clear weather is more likely to end up injured then in a property damage.
- However,two factors that were also involved were Overcast and Raining weather.These two situation contribute to some extent to accident severity.



4. Predictive modeling

As per our main objective,we want to develop a system that can predict state and severity of accident and factors that impact them.This will not only allow us to take precaution under these certain circumstances but will allow us to prevent any damage from happening in first place.
There are various statistical technique to develop mathematical model for different circumstances.We will try and develop few models from wide array of options to check feature importance.

4.1 Classification model

The term classification as the name states refers to associating objects with certain class in order to identify them in accordance with certain attributes.In statistics,the term classification refers to classifying objects with respect to certain features or dependent variables.
Mathematically,classification model refer to developing probabilistic relationship between one or more explanatory variables(also called independent variables) with dependent variable.

There is a wide array of classification models in statistics.Each model has certain strength and lag behind in some other aspect.We will try and use following models in this project.

- Decision Tree Classifier
- Support Vector Machine
- Random Forest Classifier
- Ada Boost Classifier

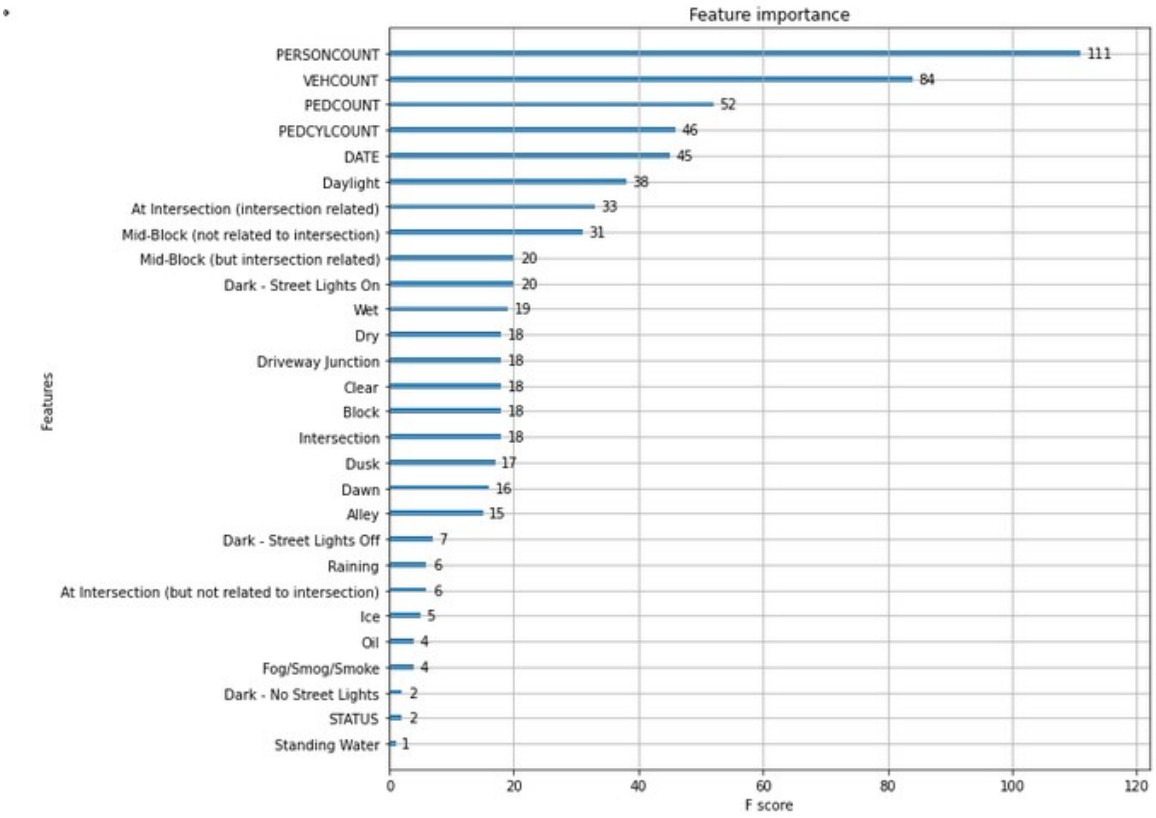
Following is F1,accuracy score,precision and recall of different models.

- Decision Tree Classifier
  - Accuracy Score 0.66
  - Precision Score 0.66
  - Recall Score 0.66
  - F1 Score 0.66
- Support Vector Machine
  - Accuracy Score 0.66
  - Precision Score 0.66
  - Recall Score 0.66
  - F1 Score 0.66
- Random Forest Classifier
  - Accuracy Score .66
  - Precision score .66
  - Recall Score .66
  - F1 Score .66
- Ada Boost Classifier
  - Accuracy Score .65
  - Precision score .66
  - Recall Score .65
  - F1 Score .66
- XG Boost Classifier
  - Accuracy Score .67
  - Precision score .67
  - Recall Score .67
  - F1 Score .67

5. Result and Discussion

Although,all the models gave similar results and were interpretable but there were certain differences.So In the end,we decided to use XGBoost model as it provided best accuracy,f1 score,precision and recall.All in all it was the all in one package we had.

After performing complex statistical techniques,we had to analyze which factor impact road accident more then other.This is major objective of this course as mentioned in the introduction.
Outcome was that greater number of persons,vehicles,pedestrians involved,greater the severity of accident and vice versa.
This was expected from beginning but thing that was surprising is that on some day accidents occur more then the other.On these dates there is usually either start of work week or end of work week.Mainly either Monday or Friday.
Precautions can help us avoid these situations from occurring in the first place.More caution needs to be advised on normal days as travelers become careless when situation is normal and this leads to a disaster.
Further more we need to lower the number of vehicles of street to prevent other traffic related issues.
On alley,there needs to be more strict speed control laws and drivers should be penalized more,



6. Conclusion

Although we got ~67% accuracy using classification models using XG Boost,however it can be further improvised.Many important features had to be dropped to due the fact that they has 75~90% missing data.Had we tried to fill them up using something we might have biased our dataset and

eventually we would have had biased our dataset.If Dataset had all the variables in place then it would have been much more accurate.