Accident Severity

# Introduction

## Background

Roads are essential in our everyday lives for driving, riding, walking, travelling, and obtaining goods and services. Unfortunately, road accidents involved fatality and serious injury bring unmeasurable suffering to human and social life.

Worldwide more people are killed in road accidents every year, which influences global health and contribute to causes of poverty. Actions will need to be taken to halt the rising in road injury. This is done by identifying causes of road accidents, and what contributes to road fatalities, then implementing road safety measures at global, regional and national levels. The severity of accident injuries can be reduced by demonstrating effective road safety strategies that are based on comprehensive studies.

## Problem

Accident severity analysis is done to identify factors of accidents which are important for estimating accident injury cost. This will assist in determining comprehensive safety strategies and accordingly, increasing road safety. Factors to use in the analysis might include ……..

This project aims to predict the severity of an accident based on analyzing these factors

## Interest

Seattle Traffic Management Division along with many other Traffic Management division are interested in accurate prediction of accident severity to improve road safety strategies. The study can also raise awareness to the general public population by highlighting which factors to focus on while driving. In addition, road engineers and consultants will have better insights on road conditions when designing new roads and associated traffic lights.

# Data acquisition and cleaning

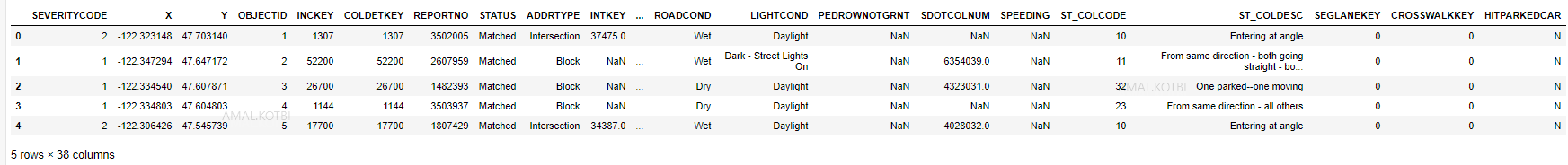
## Data Sources

The used Dataset has been provided by Seattle Traffic Management Division (SDOT) under link: [Data-Collisions.csv](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv)

Data is updated weekly by SDOT and contains observations from 2004 to Present timeframe

Full explanation for Dataset can be found at link: [Data-Collisions-Metadata.pdf](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf)

Below is quick look on our data set which contains 194673 observations and 38 variables



Our target is to predict the level of severity caused by an accident under (SEVERITYCODE) label. Hence, (SEVERITYCODE) will be our dependent variable Y

We will use different combinations of remaining independent variables X to identify which factors lead to the level of severity caused by an accident.

To accomplish our target, we will need to clean our data set. first, by removing unrelated columns and handling missing values.

## Data Cleaning

Dataset provided contained a lot of missing values. First, I dropped all rows which has empty values under target variable “SEVERITYCODE”, as we are concerned with training our model and predicting “SEVERITYCODE”, we have no interest in non-existing values under “SEVERITYCODE”. Second, I calculated missing values under the remaining attributes. And below is summary:

* "X": 5334 missing data
* "Y": 5334 missing data
* "ADDRTYPE": 1926 missing data
* "INTKEY" : 65070 missing data
* "LOCATION": 2677 missing data
* "EXCEPTRSNCODE": 109862 missing data
* "EXCEPTRSNDESC": 189035 missing data
* "COLLISIONTYPE": 4904 missing data
* "JUNCTIONTYPE": 6329 missing data
* "INATTENTIONIND": 164868 missing data
* "UNDERINFL": 4884 missing data
* "WEATHER": 5081 missing data
* "ROADCOND": 5012 missing data
* "LIGHTCOND": 5170 missing data
* "PEDROWNOTGRNT": 190006 missing data
* "SDOTCOLNUM": 79737 missing data
* "SPEEDING": 185340 missing data
* "ST\_COLCODE": 18 missing data
* "ST\_COLDESC": 4904 missing data

Therefore, I decided to drop unrelated columns and columns that contain mostly missing values. Dropped columns are: ('X','Y','OBJECTID','REPORTNO','STATUS','LOCATION','INCKEY','COLDETKEY','PEDROWNOTGRNT','EXCEPTRSNCODE','EXCEPTRSNDESC','INATTENTIONIND','INCDATE','SPEEDING','SEVERITYCODE.1','SDOTCOLNUM')

Finally, to handle other missing data, I decided to follow below:

* "ADDRTYPE": replace by most frequently address type “Block”
* INTKEY: replace by 0 for "no intersection"
* "COLLISIONTYPE": remove rows as collision type is important for severity prediction
* "JUNCTIONTYPE": replace by most frequent “"Mid-Block (not related to intersection)"”
* "UNDERINFL": convert "N" to "0" and "Y" to "1", and replace by most frequent value “N”
* "WEATHER": replace by "unknown" for empty unknown values
* "ROADCOND": replace by "unknown" for empty unknown values
* "LIGHTCOND": replace by "unknown" for empty unknown values
* "ST\_COLCODE": replace by most frequent Collision code
* "ST\_COLDESC": will leave as is ,it describes ST\_COLCODE

Now our dataset is cleaned up and ready for analysis. As a result of that analysis, we will select attributes that have more impact on (SEVERITYCODE)

# Exploratory Data Analysis

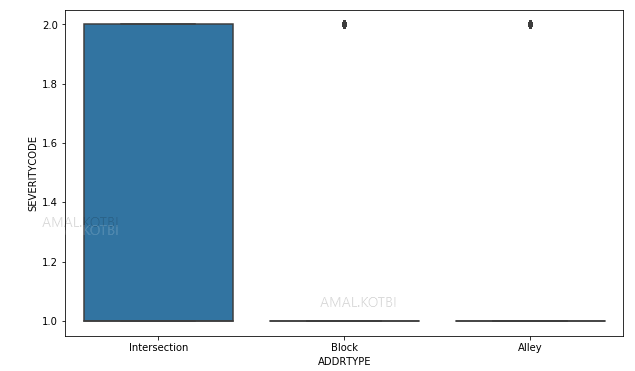
I analyzed each feature against our target variable “SEVERITYCODE” to select the feature that has impact of severity of an accident. Thus, generated visualizations such as box plot and regression plot

## ADDRTYPE vs SEVERITYCODE

Most accidents happen at “Block” Address Type

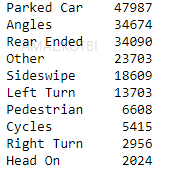


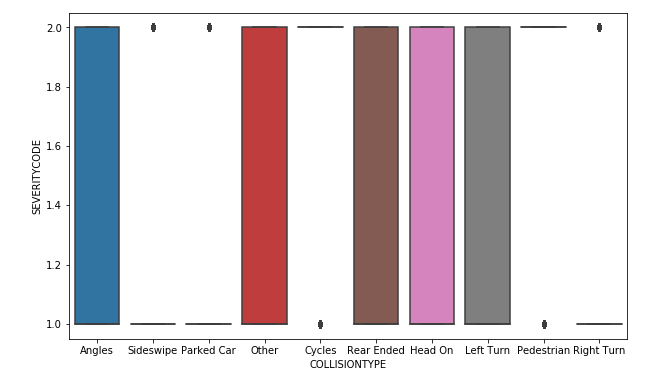
In addition, from below boxplot we can see that Address Type affects severity of an accident



## COLLISIONTYPE vs SEVERITYCODE

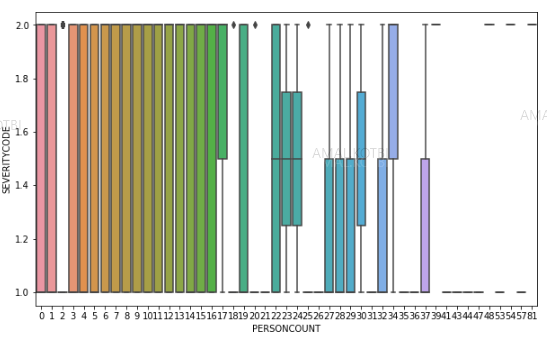
Most accidents happen at “Parked Car” Collision Type

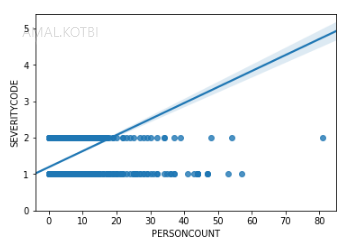


In addition, from below boxplot we can see that Collision Type affects severity of an accident

## PERSONCOUNT vs SEVERITYCODE

Most accidents happen with two people involved. We will plot PERSONCOUNT using box plot and regression plot to get a clear understanding:

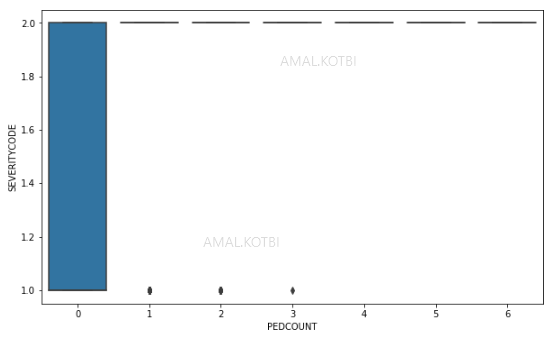


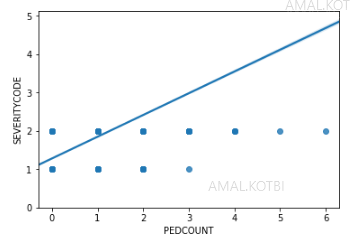


From above figures we can see that severity is affected by the number of people involved in an accident

## PEDCOUNT vs SEVERITYCODE

Most accidents involve no pedestrians, yet we will generate both boxplot and regression plot to get a clear understanding

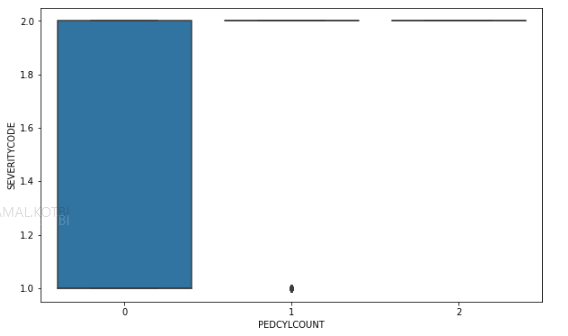


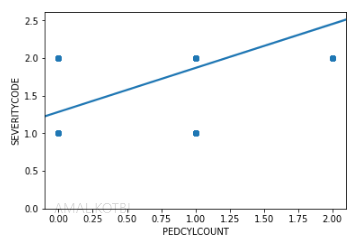


We can see from both figures that severity is affected by the number of pedestrians involved in an accident

## PEDCYLCOUNT vs SEVERITYCODE

Most accident involve no cyclists. Yet we will generate both boxplot and regression plot to get a clear understanding

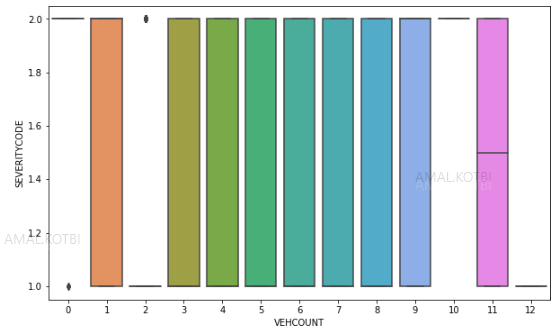


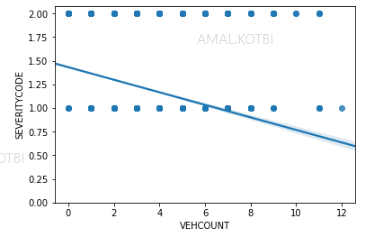


We can see from both figures that severity is affected by the number of cyclists involved in an accident

## VEHCOUNT vs SEVERITYCODE

Most accident involve two vehicles. Will generate both boxplot and regression plot to get a clear understanding

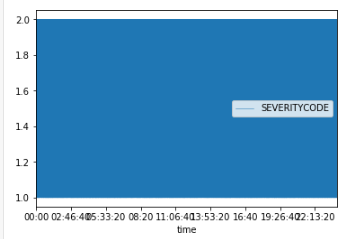


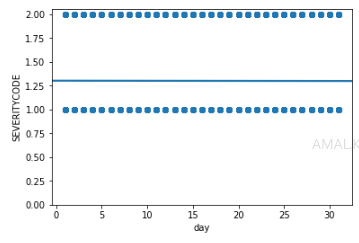


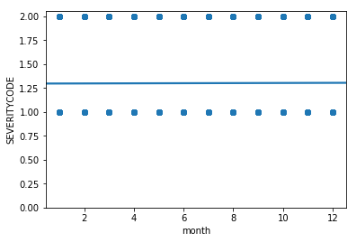
We can see from both figures that severity is negatively affected by the number of vehicles involved in an accident. The more the vehicles, the less the severity

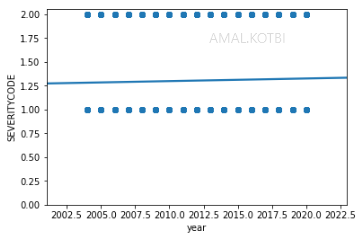
## INCDTTM (Date and Time) vs SEVERITYCODE

We have analyzed day, month, year, and time. And found out that most accident happen at mid-day. And month October showed the highest number of accident. And year 2006 has the most number of accident as well. We will plot all of them against SEVERITYCODE to get a clear understanding of the relationship





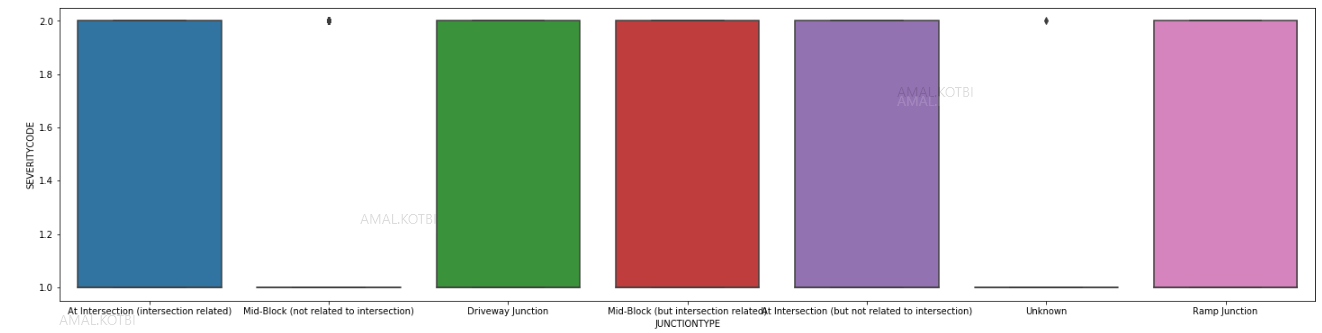




From plots above, we do not see a clear impact of time, day, month, year over severity of accidents. Therefore, I have decided to drop these columns from our feature selection

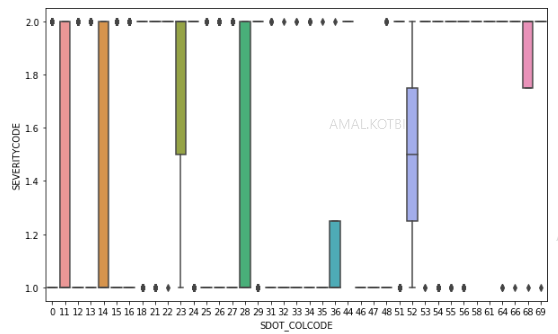
## JUNCTIONTYPE vs SEVERITYCODE

We can see that most accidents happen at “Mid-Block (Not related to intersection)”. And from box plot below, we can see that accident severity is affected by the type of the junction



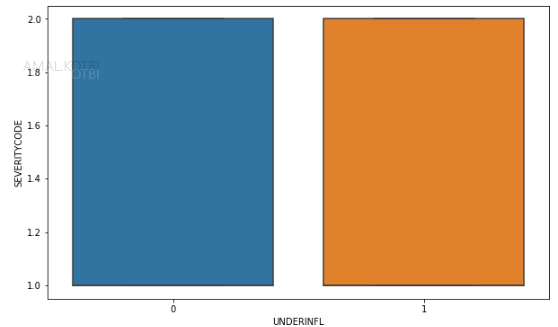
## SDOT\_COLCODE vs SEVERITYCODE

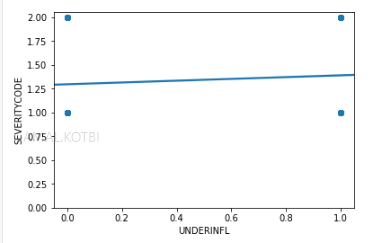
We can see that most accident collisions are under “11” code. And from box plot below we can see clear impact on the collision code on accident severity



## UNDERINFL vs SEVERITYCODE

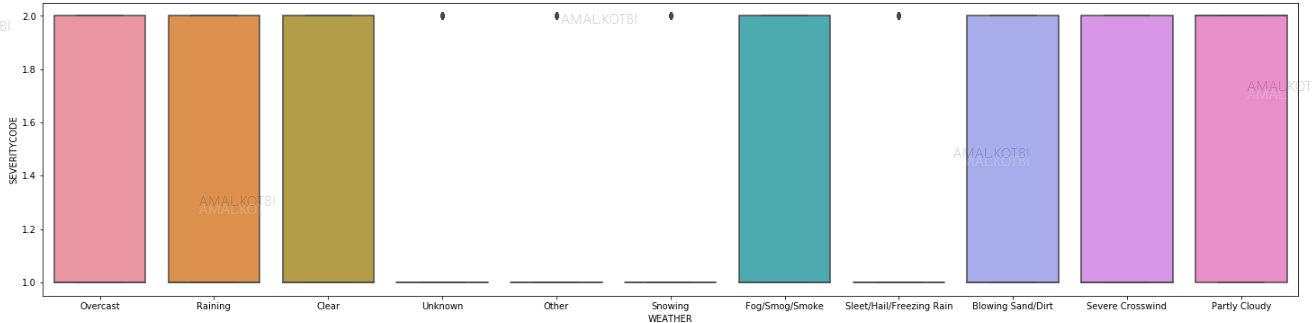
We can see that most accidents happen with driver not under influence. And from box plot and regression plots below, we see no clear impact on the severity of an accident. So we will drop column





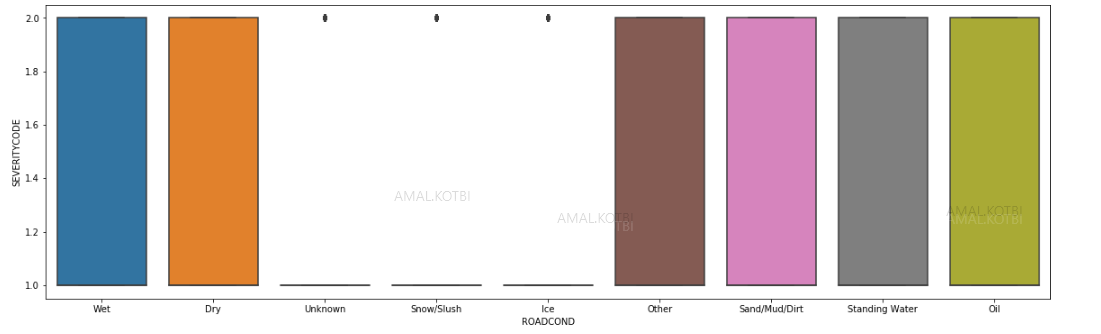
## WEATHER vs SEVERITYCODE

Most accidents happen on “Clear” weather. From box plot below, we see the weather impact on severity



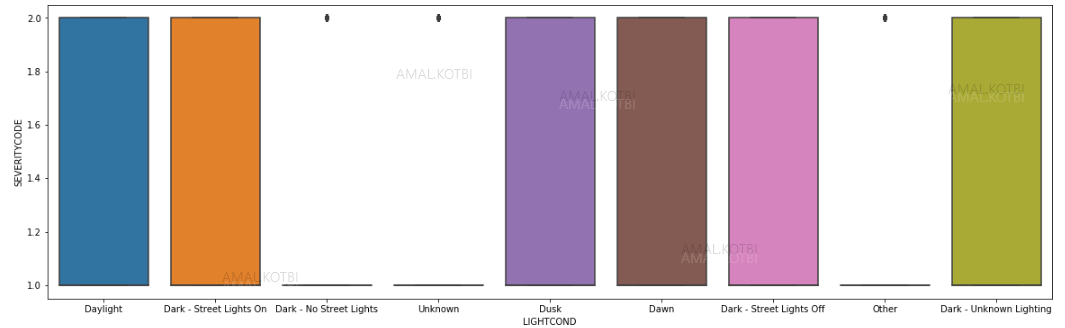
## ROADCOND vs SEVERITYCODE

Most accidents happen on “Dry” road condition. From box plot below, we see the impact on severity



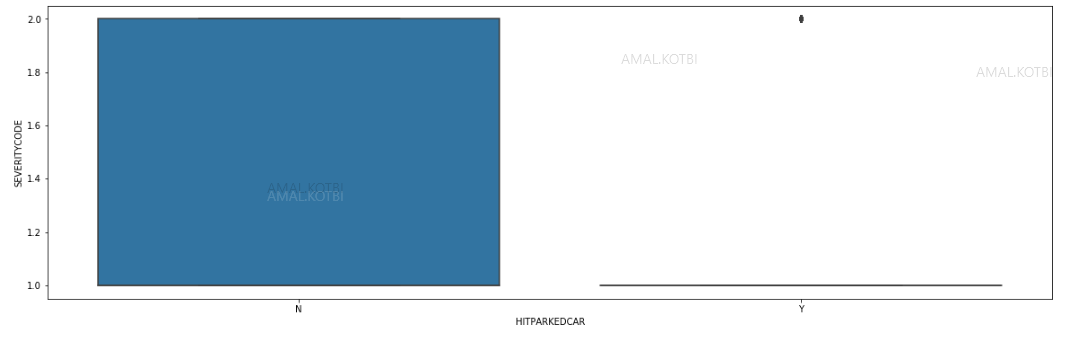
## LIGHTCOND vs SEVERITYCODE

Most accidents happen on “Daylight” light condition. From box plot below, we see the impact on severity



## HITPARKEDCAR vs SEVERITYCODE

Most accidents do not involve hitting parked car. Yet we can see the impact on severity from box plot

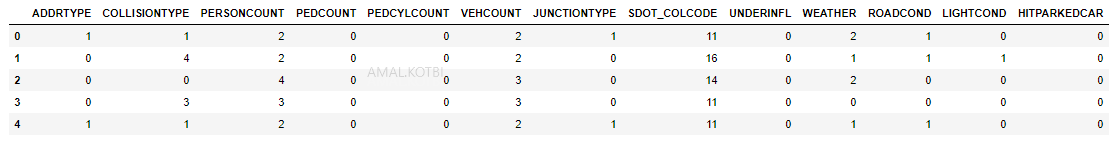


# Feature selection

Based on Data exploratory analysis, we have selected independent features below to build our model

['ADDRTYPE','COLLISIONTYPE','PERSONCOUNT','PEDCOUNT','PEDCYLCOUNT','VEHCOUNT','JUNCTIONTYPE','SDOT\_COLCODE', 'WEATHER','ROADCOND','LIGHTCOND','HITPARKEDCAR']

In order to build our model, categorical features have been converted into numerical so it can be fitted into the model



# Classification model

I used supervised machine learning classification models to predict the class of accident severity. The models will learn the relationship between selected features above and target variable “SEVERITYCODE”. I have built different models using different classification algorithm (KNN, Decision-tree, SVM, Logistic regression) and calculated model accuracy.

## Dataset Train/test Split

In order to build the model and test its accuracy, our dataset has been divided into Train data (to train the model), and Test data (to test accuracy of the model). Train dataset size=151815, while Test dataset size= 37954

## Models results and Accuracy

I have built the models using (KNN, Decision-tree, SVM, Logistic regression) and calculated model accuracy using jaccard-similarity and f1-score. Below is result of each model:

|  |  |  |
| --- | --- | --- |
|  | jaccard-similarity | f1-score |
| KNN | 0.744 | 0.715 |
| Decision-Tree | 0.7494 | 0.686 |
| SVM | 0.7609 | 0.72 |
| Logistic regression | 0.75 | 0.7 |

The models performed similarly. Models accuracies evaluated and their values appear to be very close to each other. SVM algorithm has shown best results to predict accident severity

# Discussion

The models built have room for improvement, and accuracy can be improved by utilizing a better data set that contains (Speeding data, driver under influence or not, driver didn’t pay attention..etc)

These factors contained a lot of missing data and either were removed or showed no value.

# Conclusion

In this study, I have analyzed the relationship between accident severity and other factors that will impact the severity of an accident. I identified address type, collision type, and the involved number of people, pedestrians and vehicles count, also junction type, and conditions of weather, road and light, and finally whether the accident involved hitting a parked car. These factor were used to build a classification model to predict the accident severity. The model is useful in helping the general population being aware while driving, the model also assists Traffic Management departments to plan and implement any new traffic regulations, in addition, the model shed light for road engineers and consultants in designing new roads