Predicting Car Accident Severity:

IBM Data Science Capstone Report

Introduction to the Business problem

Background

Accidents are a leading cause of death across the world. Due to the increase in population as well as motorization, the number of accidents has been increasing globally. Better predictions of the severity of accidents are crucial to enhance the safety performance of road traffic systems, and reduce such unfortunate incidences.

Though measures are in place to reduce accidents and increase awareness regarding the importance of ensuring safety, accidents continue to occur frequently. In addition to the suffering caused by the loss of human life and resulting disabilities, the economic burden is quite high. The increased expenditure, such as health care, specialized force training, unemployment due to disabilities, all add to this burden. In 2020, WHO reported that road traffic crashes cost most countries 3% of their gross domestic product. Hence, it is vital to find ways to minimize accidents and in turn the adverse effects on human life and society. A literature survey shows that driver injury severity can be classified into a few categories such as property damage, possible/evident injury, or disabling injury/fatality. Hence, modeling accident severity can be addressed as a pattern recognition problem. Such a problem can be solved by deep learning, statistical techniques, or by physical modeling approaches. In a deep learning model, an input vector (attribute) is often mapped into an output vector (target label) through a set of nonlinear functions. In the case of accident severity, the attributes are the characteristics of the accident, such as driver behavior, road and weather conditions, etc. The target label in this case would be the accident severity.

Problem definition:

- To predict the severity of accidents to enhance the safety performance of road traffic systems.

Target Audience

Government, WHO, any company/enterprise involved in marketing road safety equipment. As the aim is to predict the severity of accidents, anyone or any company/organization is a target audience as the success of enhancing safety depends on collaborative work.

DATA

Data Understanding

The data used for this study has been collected and shared by the Seattle Police Department (recorded by Traffic Records). Coursera provides this data for downloading through a link (provided below).

The dataset reports all collisions, which occurred in the period from 2004 to the date issued (updated weekly). The dataset provides information regarding the severity of the traffic accident, location, type of collision, weather, road and light conditions, visibility, number of people involved, etc.

The data set provided for this work allows the analysis of a record of about 200,000 accidents in the state of Seattle. This set is sufficiently large for an accurate prediction of the accident severity. There are 38 labels in the dataset. The *label 'SEVERITYCODE' is the predictor/target variable*, and are defined as follows: •3—fatality •2b—serious injury •2—injury •1—prop damage •0—unknown

The remaining 37 attributes may be used to predict the target variable. Several attributes such as 'UNDERINF', SPEEDING', 'ROADCOND', (i.e. whether or not a driver involved was under the influence of drugs or alcohol, whether or not speeding was a factor in the collision (Y/N), the condition of the road during the collision.) can be analyzed for understanding how gravely they contribute to the accidents. Only those attributed which are essential for the accident severity prediction would be used to train the models. Further, the data is also coded in accordance with the State Collision Code Dictionary, describing various factors such as the direction of vehicles involved, type of vehicle, etc. The attributes, which will be considered, in this case, for predicting the accident severity are:

1. Road Condition - Counts for each road condition are shown below:

- Dry 124510
- Wet 47474
- Unknown 15078
- Ice 1209
- Snow/Slush 1004
- Other 132
- Standing Water 115
- Sand/Mud/Dirt 75
- Oil 64

2. Light Condition - Counts for the different Light conditions are shown below:

- Daylight 116137

- Dark - Street Lights On 48507
- Unknown 13473
- Dusk 5902
- Dawn 2502
- Dark - No Street Lights 1537
- Dark - Street Lights Off 1199
- Other 235
- Dark - Unknown Lighting 11

3. Speeding -

- Y 9333 (Means Yes)
- Remaining are Nan - We will change it to N, and consider it as No.

4. Collision Type - Count for different recorded collision types are shown below:

- Parked Car 47987
- Angles 34674
- Rear Ended 34090
- Other 23703
- Sideswipe 18609
- Left Turn 13703
- Pedestrian 6608
- Cycles 5415
- Right Turn 2956
- Head On 2024

The data will be cleaned and used so that we can determine which attributes are most common in traffic accidents in order to target prevention at these high-incidence points.

- Data Source: These data have been collected and shared by the Seattle Police Department (Traffic Records) and are provided by Coursera for downloading through a link.
- Data Location: https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset Data set name: Data-Collisions.csv

This analysis should help anyone interested, to focus their resources on handling the most contributing factors and determine a prevention strategy

CLEANING DATA

The data needs to be cleaned, before the different ML models are used for predicting accident severity.

1. Columns not required for predicting accident severity such as OBJECTID (ESRI unique identifier), INCKEY (A unique key for the incident) etc., should be removed. A new Data Frame was created including the target variable, and the attributes used for prediction.

2. The features used for prediction are of object type. These were converted to numerical type using label encoding. SPEEDING can have two values, 0 & 1. Hence LabelEncoder was used. The other three features can take multiple values (object type), hence OneHotEncoder is additionally used to convert the data to a binary response.

3. To remove null values, after converting Nan values in SPEEDING to No, nan values in other features are dropped.
4. Dataset Balancing: The target variable: SEVERITYCODE is not balanced. The code representing 'property damage' (136485) is represented more than twice than code 2, representing injury (58188). The class distribution must be equal or at leats close to equal. If not, when an algorithm needs to learn, it will not recognize the minority.

   Hence, two balancing methods were employed: upsampling of minority group and downsampling of majority group. They are separately checked by the accuracy, recall and f1 score metrics (see discussion). In upsampling, randomly observations from the minority class are duplicated, to increase its sample size.For this the resample module from Scikit-Learn was used.

   Models have been also employed on the unbalanced data, which shows:

   a) How accuracy score can be high, becuase of unbalanced data. However, this is misleading as the recall of minority group will be low.

   b) RandomForestClassifier works better with unbalanced data as compared to other models used in this project.

Once cleaned, the data was analyzed with different ML models.

Methodology

We have labeled data. Hence, four supervised ML models are employed for prediction. Their efficiency is checked with accuracy, f1, and recall score metrics.

The ML algorithms employed are:

Logistic Regression

The data is binary, predicting two possible outcomes for the SEVERITY CODE (Though SEVERITY CODE can have 5 different values, in the dataset used, it has only 2 unique values: 1 & 2). Hence, logistic regression can be used for predicting accident severity.

K-Nearest Neighbor (KNN)

KNN will help us predict the severity code for a set of features by finding the most similar to data point within defined k distance. KNN is a simple algorithm, which is easy to implement and interpret. Compared to other supervised models, it is known to provide higher accuracy.

•

Decision Tree

The decision tree is a ML supervised model. The results of tree model are easy to represent (visually) and interpret. It makes a series of decisions based on a set of features, and these decisions lead to a specific result. Though these are simple to understand, a single tree is usually not sufficient for producing satisfying results. And hence called, Random Forest

In order to understand the effect of different features on accident severity, we have plotted the data using bar plots.

RandomForest

Random forest is a tree-based algorithm, using multiple trees to make the decisions. This makes it much more powerful. In RF method, decision trees are randomly created, with each node working on a random subset of features (hence the name)

Results & Evaluation

To check the accuracy of the four models used, the f1-score, accuracy score and recall score metrics is employed. Data has been analyzed in three forms: unbalanced data, upsampled data, and downsampled data.

A) Unbalanced Data

With unbalanced data, the accuracy scores for all models are high. This is because the data is unbalanced with the majority having a SEVERITYCODE =1. Hence, the accuracy is for predicting the majority class.

	Method of Analisys	F1-score	Accuracy	Recall
0	KNN	0.841758	0.744393	0.972805
1	Decision Tree	0.846881	0.749640	0.990706
2	LogisticRegression	0.845307	0.749688	0.990706
3	RandomForest	0.846881	0.846881	0.990706

If we individually check the classification reports of each, we see the problem of using data without balancing first. In all models (except RF) the number of correct positive predictions made out of all positive predictions (recall score) made for SEVERITYCODE =1 is high (0.98 with LR model), whereas for SEVERITYCODE =2, which was in minority, is low (0.22 with LR).

	precision	recall	f1-score	support
1.0	0.74	0.98	0.85	43685
2.0	0.81	0.22	0.34	18825
accuracy			0.75	62510
macro avg	0.78	0.60	0.59	62510
weighted avg	0.77	0.75	0.69	62510

However, RandomForest is known to work better with unbalanced data. With RF model, the recall for SEVERITYCODE =1 is lowered (0.59), but the recall for SEVERITYCODE =2 (minority) is high (0.82), and overall the model works much well than other employed models on an unbalanced data.

	precision	recall	f1-score	support
1.0	0.88	0.59	0.71	43685
2.0	0.46	0.82	0.59	18825
accuracy			0.66	62510
macro avg	0.67	0.71	0.65	62510
weighted avg	0.76	0.66	0.67	62510

B) Balanced data: Two methods of balancing are employed:

1) Upsampling: This method is usually employed when the sample size is small. The accuracy dropped in comparison to the accuracies observed with unbalanced data. However, if we individually check the classification reports for all models the number of positive predictions have increased for SEVERITYCODE=2.

	Method of Analisys	F1-score	Accuracy	Recall
0	KNN	0.672878	0.687416	0.646935
1	Decision Tree	0.596140	0.687701	0.463828
2	LogisticRegression	0.667145	0.704253	0.596422
3	RandomForest	0.596140	0.596140	0.463828

2. Downsampling: This method is usually employed for large data. The metrics indicates that downsampling shows similar accuracy and recall scores as upsampling.

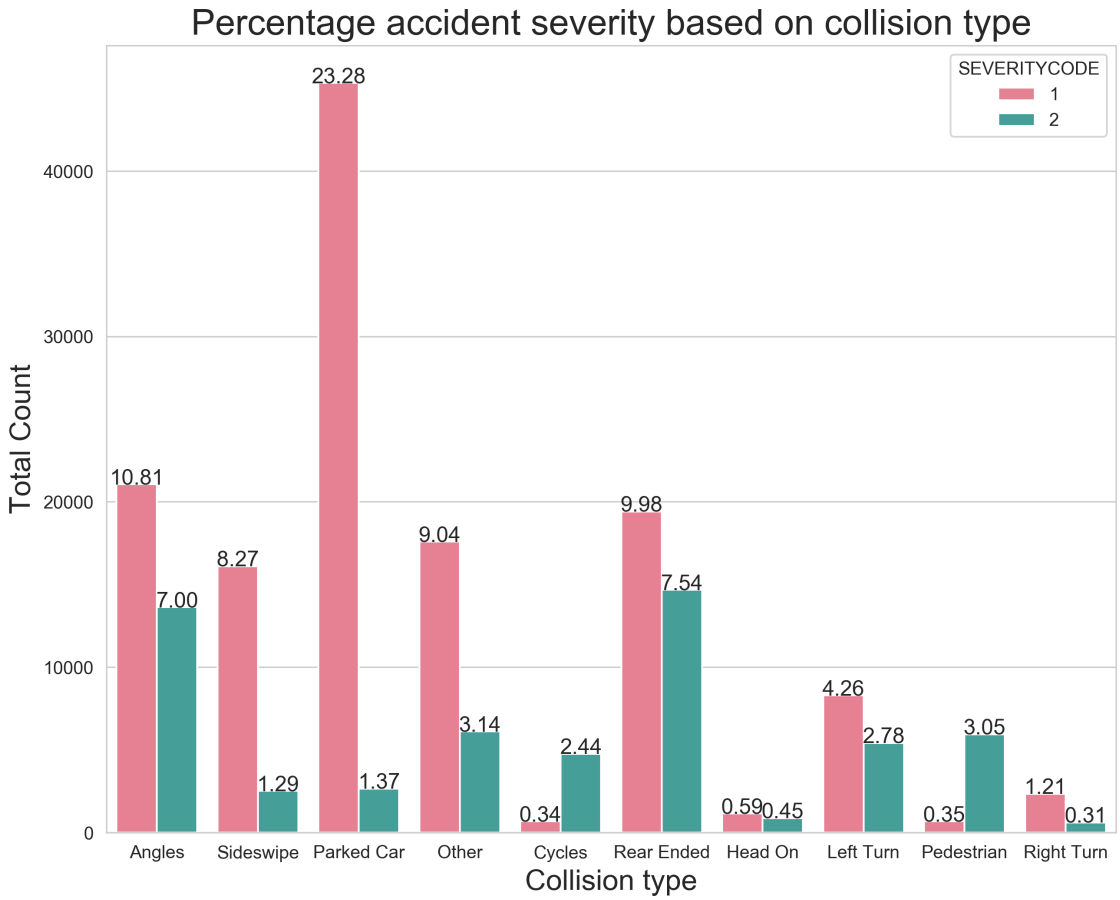
	Method of Analisys	F1-score	Accuracy	Recall
0	KNN	0.674533	0.659697	0.711645
1	Decision Tree	0.602063	0.691981	0.470227
2	LogisticRegression	0.671544	0.708670	0.601013
3	RandomForest	0.602063	0.602063	0.470227

Next, let us discuss how different features affect accident severity, and how can we solve the challenges presented by analyzing the data.

Discussion

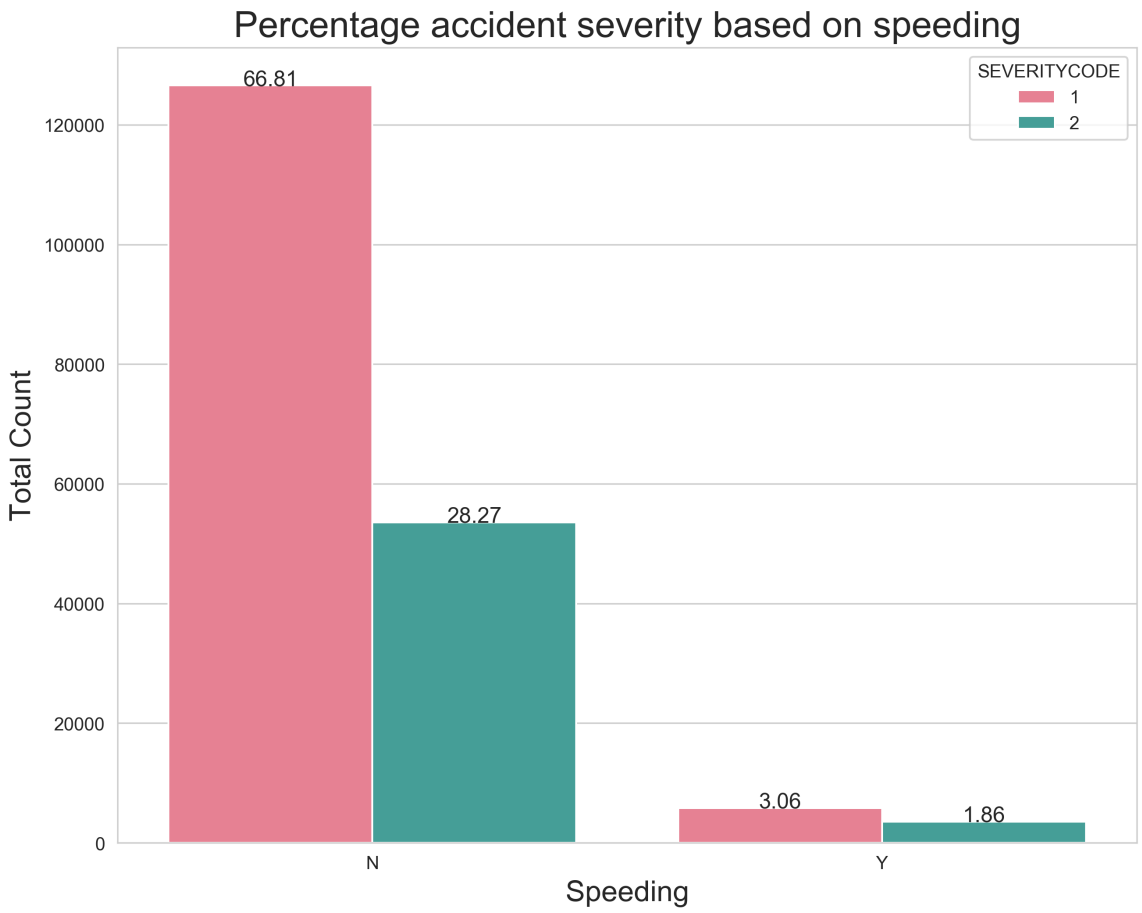
In this project, we first defined the problem we needed to solve, which is to understand (predict) how different factors influence the severity of accidents. We retrieved the needed data, cleaned it, and followed by using different models to predict the severity of a car accident based on the knowledge of various factors influencing accidents.

1. Collision Type



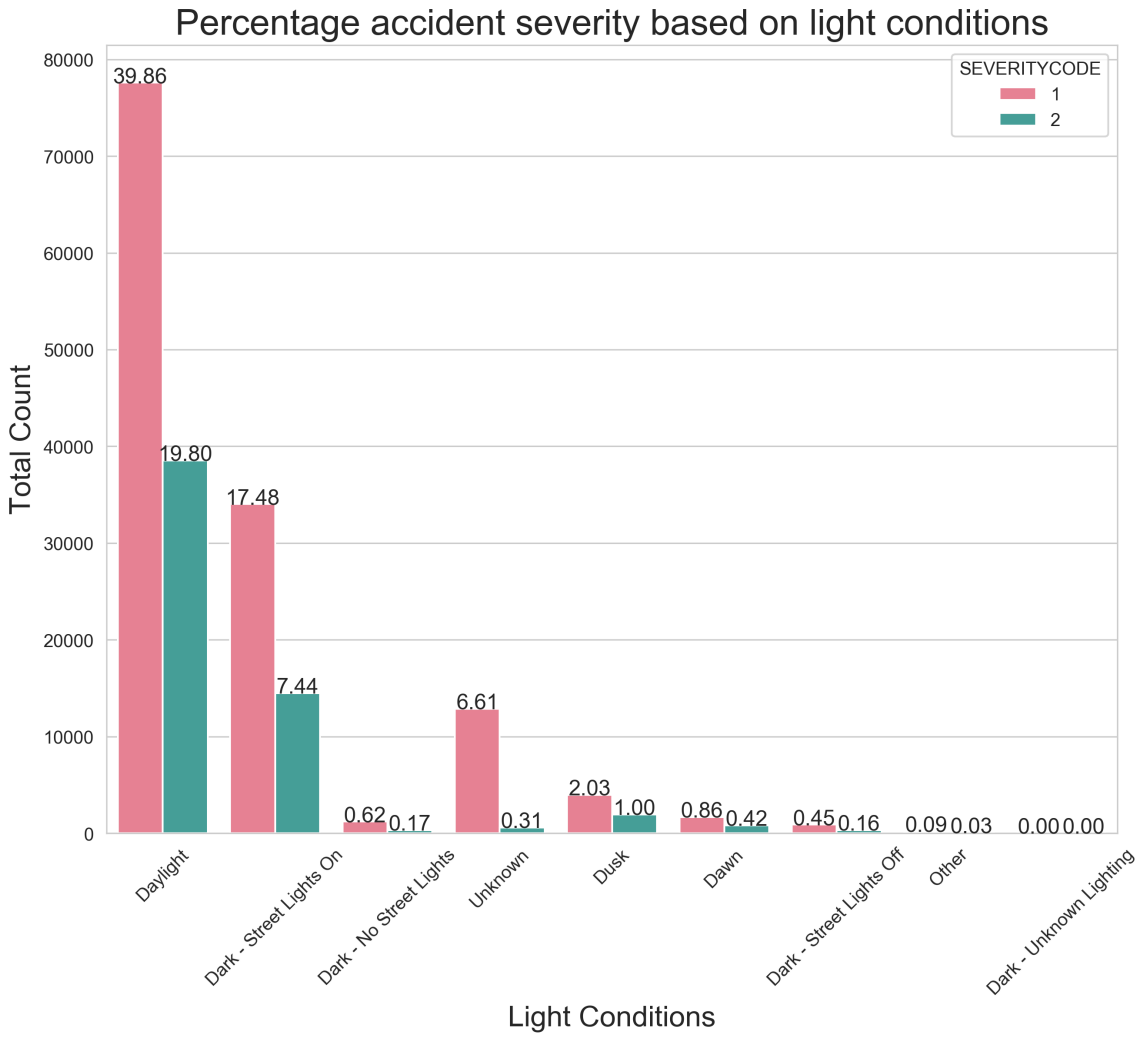
The collision involving cycles and pedestrians has a high chance of causing injury. This can be also because other categories include heavy vehicles such as cars, trucks that can cause more property damage. Further, a pedestrian and cyclist are also less protected from impact. Nevertheless, the security of a cyclist and pedestrian has to be considered vital while constructing roads, traffic lights etc. In case of vehicles, the chance of injury is high if it collides, either at angles, from the back, head on and at left turns. Special attention needs to be paid to the distance between vehicles and the speed of vehicles. Drivers, themselves also need to be cautious and vigilant about the movement, speed and proximity (rear end) of other vehicles. Parked cars result in lower injury, due to high chance of the car having no occupants.

2. Speeding:



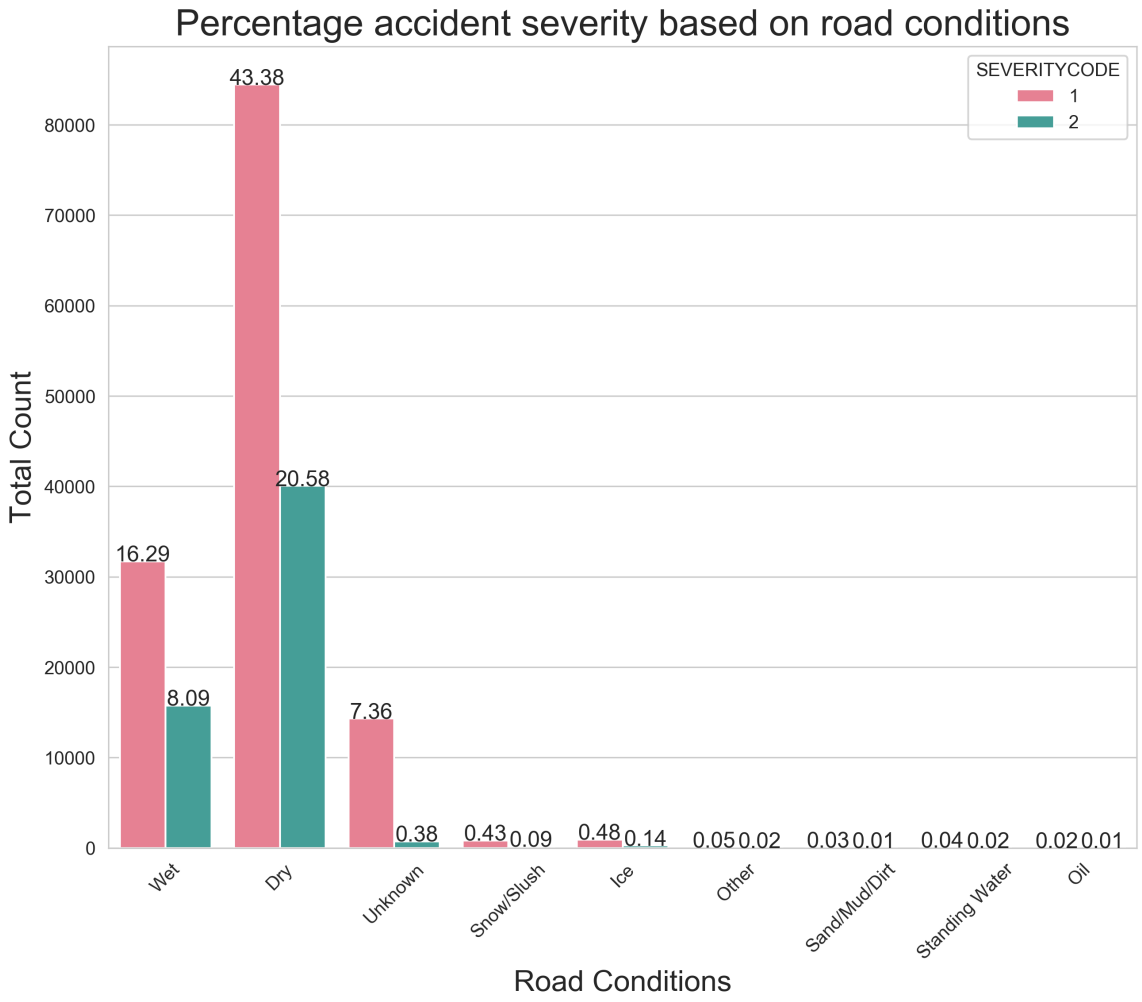
About 95% of the reported accidents occurred when speeding was not involved. The number of drivers speeding might be lower and hence the low numbers. Speed check need to be implemented and monitored regularly, particularly, in locations where speeding is observed. Lowering speed limits can boost safety.

2. Light Conditions:



The number of accidents observed are higher in daylight, as driving occurs largely during daytime. The data shows, accidents during daylight have as much chance of causing an injury as an accident occurring during dark. When the streetlights are off, the chances of drivers taking the road are low. This might be leading to the low numbers in these categories. 7% of reported accident occurred in dark, where there were no streetlights. Drivers have to be extremely cautious when driving in such situations. Streetlights should be installed on all commutable pathways.

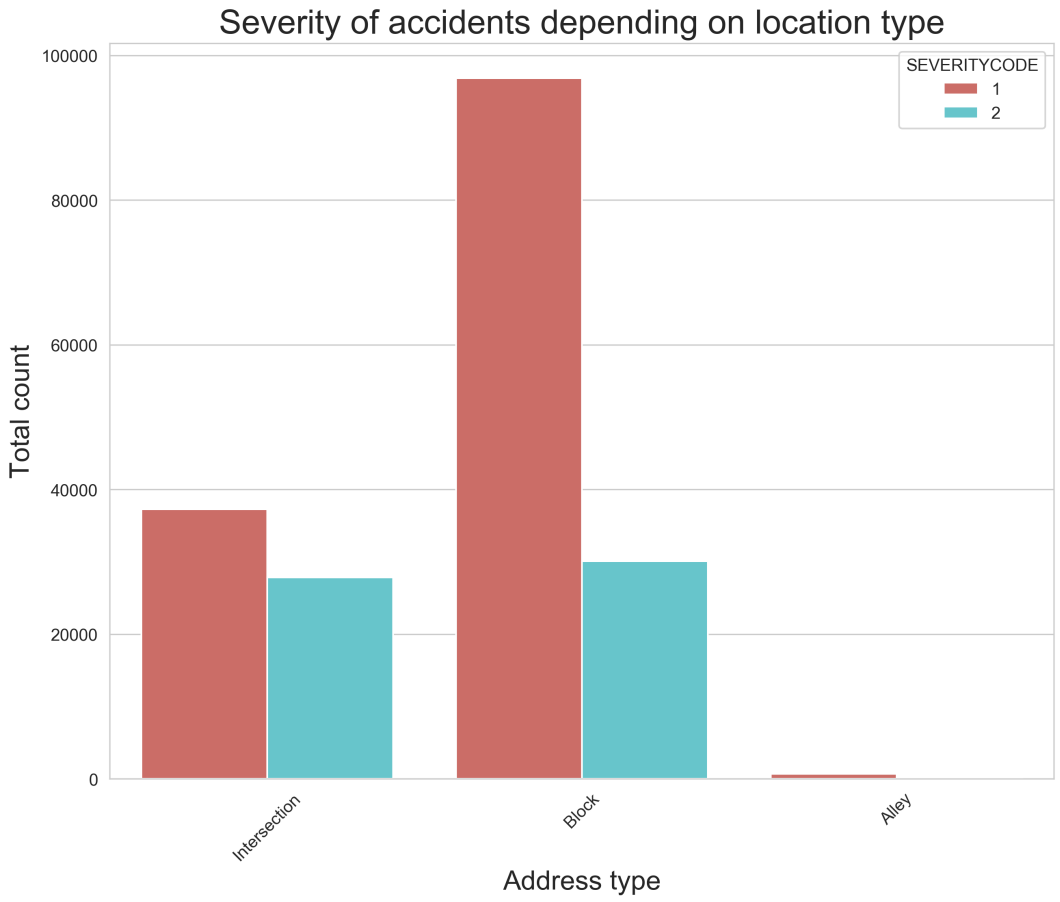
4. Road conditions:



The plot shows that number of accidents is reported most in dry conditions. This is followed by accidents in wet conditions. Looking at the climate of Seattle gives a better picture of this data. The US average is 28 inches of **snow** per year. In contrast Seattle averages 5 inches of **snow** per year. It is reported that on average, Seattle enjoys 152 sunny days per year. The lower snowing average of Seattle directly reflects on the number of accidents occurring in Snow/Slush conditions. It is mostly dry or wet in Seattle. In either condition, the chances of injury are half of incurring property damage.

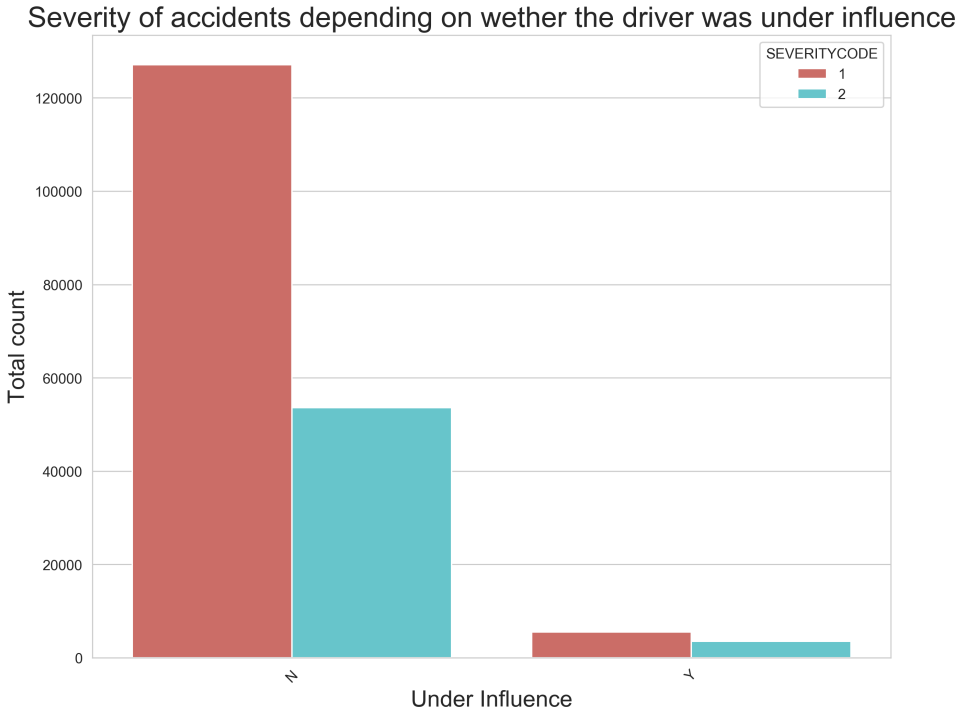
Apart from the features used for modeling, the effect of features, which may result in higher accident severity were plotted.

A) Address type



A large number of accidents are reported from blocks and intersections. The speed limit in alleys is low. This might be the reason behind the no. of reported accidents in alleys being low. Most of the accidents are occurring at the mid-block of a segment causing property damage and injury.

B) Under influence



Surprisingly, the driver being under influence is not shown to particularly affect the severity of the car accident.

Conclusion

Based on data, it can be concluded that several factors needs to be considered to lower the occurrence and severity of accidents. Firstly, cyclists and pedestrians are shown to be more susceptible to getting injured during the event of an accident. Safety of both these categories should be considered while constructing roads, including presence of good footpaths, traffic lights, crossways, bridges to cross large roads, etc. The number of accidents also depends on where it occurs. At the mid-block of a segment, the no. of accidents reported are higher. Observing the affect of light conditions, it is suggested that streetlights should be installed at all commutable paths, to lower the no. of accidents in the dark. Lowering the speed limit might also boost safety.