Coursera

IBM Data Science

Professional Certificate

Capstone Project

Predicting the possible Accident Severity

Mr. Mantake Singh

September 2020

Index

1. Introduction

Road accidents are a major world economic and social problem as shown by the report of loss of lives and properties in many countries around the world. Reporting indicated the number of fatalities from road accidents per year of about 1.3 million and 50 million injuries were recorded or an average of 3000 deaths/day and 30,000 injuries/day. Furthermore, its consequences have an impact on economic and social conditions in terms of health care costs of injuries and disabilities. The World Health Organization (WHO) estimated the economic costs derived from road accidents reached 518 billion USD per year in high income countries and 65 billion USD per year in medium and low-income countries.

Several studies have been conducted by the authorities of various countries regarding the actual reasons behind the accidents and enormous data has been gathered. This data can be easily used, by adopting appropriate Machine Learning model, to predict the severity of accident on the basis of input data. The Model can be Trained and Tested with the help of available data.

Business Problem

The objective of this capstone is to build a model to analyze the available data as well as predict the severity of possible accident given the inevitable and independent circumstances as inputs.

Target audience for the project

This project is particularly useful to Seattle Police Department, Traffic Controllers and the local residents using the road for travel. The model has been built using the general observations only so is completely independent from the location of the traveler. This model should be able to predict the severity of the accident provided appropriate inputs are given. This will highlight the problems during the road travel and help the authorities make required arrangements. For the travelers, this should make the driver more cautious during the drive by raising the severity alert.

2. Approach

There are 6 steps to build a model –

1. **Business understanding**

   The initial phase is to understand the project's objective from the business or application perspective. Then, you need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.

2. **Data understanding**

   In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

3. **Data preparation and cleaning**

   The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

4. **Modelling**

   In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. You can select k Nearest Neighbor, SVM, XGBoost, decision tree, or any other techniques. You can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required.

5. **Evaluation**

   Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

6. **Deployment**

   The deployment phase requirements vary from project to project. It can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment. In this environment, the customers or end-users can utilize the model in different ways such as API, website, or so on.

3. Data Preparation and Cleaning

To solve the objective, we need following data: -

1. List of all the accidents occurred and recorded in the Seattle.
2. Severity of every recorded accident.
3. Weather condition during the travel for every recorded accident.
4. Lighting condition on the Road for every recorded accident.
5. Road condition for every recorded accident.
6. Before the accident, whether the victim was under the influence of alcohol/substance.
7. Whether it was violation of speed limit.
8. Whether the driver was attentive.

Sources of Data

The data has been made available by the IBM online certification team. The collisions data is provided by Seattle Police Department and recorded by Traffic Records which is being updated on weekly basis since 2004.

Data Preparation and Cleaning

The first step to Data Analysis and Machine Learning is acquisition of data from the reliable sources and converting it into the required format. This format should be complete in every aspect and should help the Scientist make his/her case. It should narrate a story that the target audience should listen to with interest. The case should be made on the basis of data rather than instincts. The required dataframe should not contain any unwanted values or null values which can significantly affect the final model accuracy. The null values should either be dropped from the dataframe or appropriate values should be putted to make it more sensible.

After building a dataframe, the values are checked for null and appropriate corrections are made. There were 11 types of input in the WEATHER columns which were all converted to integers for further analysis as follows –

Serial Number	WEATHER original column values	WEATHER replaced numerical values
1.	Unknown	0
2.	Clear	1
3.	Other	2
4.	Partly Cloudy	3
5.	Overcast	4
6.	Severe Crosswinds	5
7.	Blowing Sand/Dirt	6
8.	Raining	7

9.	Fog/Smog/Smoke	8
10.	Sleet/Hail/Freezing Rain	9
11.	Snowing	10

Similarly, all the string inputs in ROADCOND and LIGHTCOND were replaced with numerical inputs for the sake of ease in analysis as given below –

Serial Number	ROADCOND original column values	ROADCOND replaced numerical values
1.	0	0
2.	Unknown	0
3.	Ice	1
4.	Sand/Mud/Dirt	2
5.	Oil	3
6.	Snow/Slush	4
7.	Standing Water	5
8.	Wet	6
9.	Other	7
10.	Dry	8

Serial Number	LIGHTCOND original column values	LIGHTCOND replaced numerical values
1.	0	0
2.	Unknown	0
3.	Dark – No Street Lights	1
4.	Dark – Street Lights Off	2
5.	Dark – Unknown Lighting	3
6.	Dark – Street Lights On	4
7.	Dusk	5
8.	Dawn	6
9.	Other	7
10.	Daylight	8

In the UNDERINFL column, the entries were in both string as well as binary format. They were all replaced with binary inputs as follows –

Serial Number	UNDERINFL original column values	UNDERINFL replaced column values
1.	Y	1
2.	N	0
3.	0	0
4.	1	1

Similarly, SPEEDING and INATTENTIONIND columns were also corrected and converted since only positives i.e. 'Y's were inputted in the dataframe as follows –

Serial Number	SPEEDING original column values	SPEEDING replaced column values
1.	Y	1
2.	Null	0

Serial Number	INATTENTIONIND original column values	INATTENTIONIND replaced column values
1.	Y	1
2.	Null	0

All the rows with null values in every column were dropped.

After cleaning and formatting, from the given dataframe "df", SEVERITYCODE, WEATHER, LIGHTCOND, ROADCOND, UNDERINFL, SPEEDING and INATTENTIONIND were extracted by dropping the remaining columns. Only these columns were selected from the large given dataframe because these were the only independent variable which could be used to predict the severity of a possible accident. In this renewed dataframe "df", the values were checked for null and appropriate corrections were made. Finally, the columns were renamed as given below –

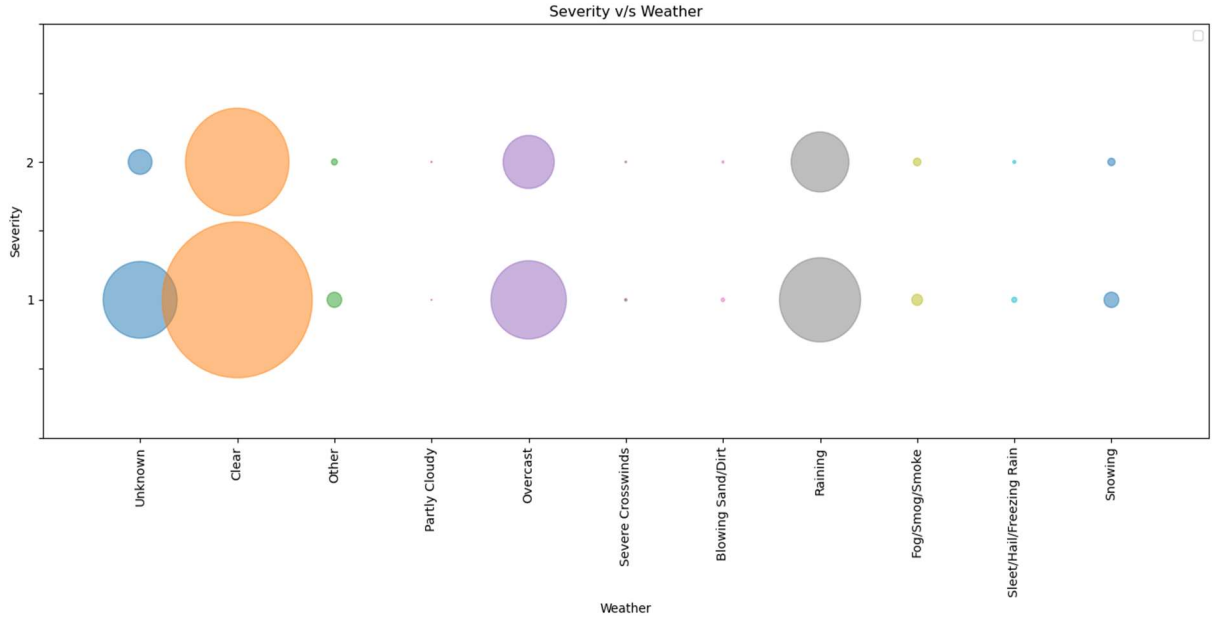
Serial Number	Original column name	Renamed column name
1.	SEVERITYCODE	severity
2.	WEATHER	weather
3.	LIGHTCOND	light
4.	ROADCOND	road
5.	UNDERINFL	influence
6.	SPEEDING	speeding
7.	INATTENTIONIND	inattention

Ultimately, the dataframe "df" consists of 7 columns plus one index column. There are 194673 entries so the shape of the dataframe "df" is (7,194673).

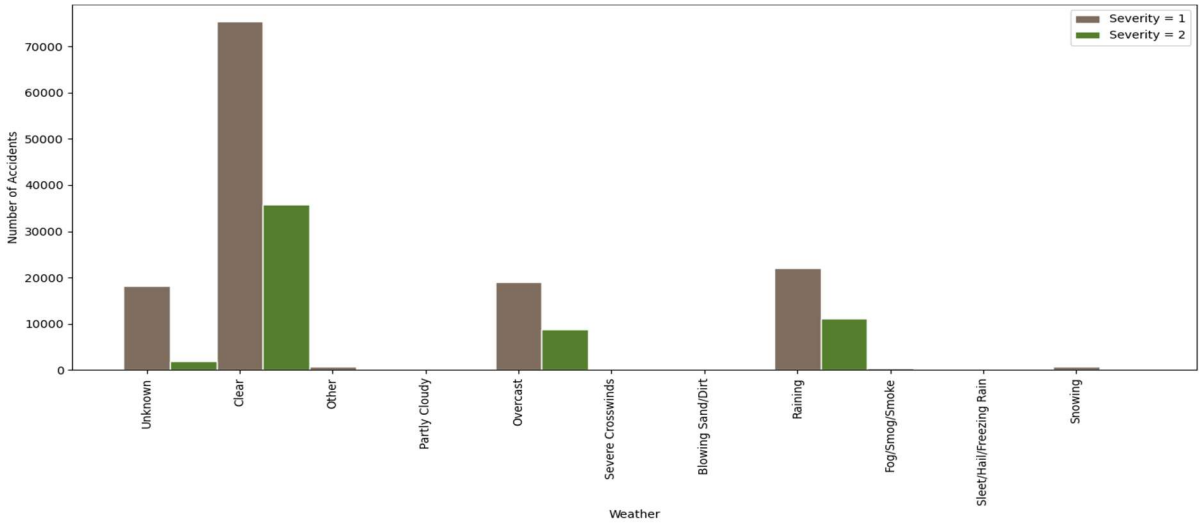
4. Exploratory Data Analysis

4.1 Relationship between Severity and Weather conditions –

It is a fact that weather plays by its own rules on the roads. A clear sky is the ideal condition to travel whereas bad weather can scrap our plans. From the given data, as it can be seen in bubble chart given below, although highest number of accidents happened during the clear sky, which is most usual, the second highest number of accidents happened during rainy weather. Snowing also contribute to the accident happenings significantly, relative to its occurrence.

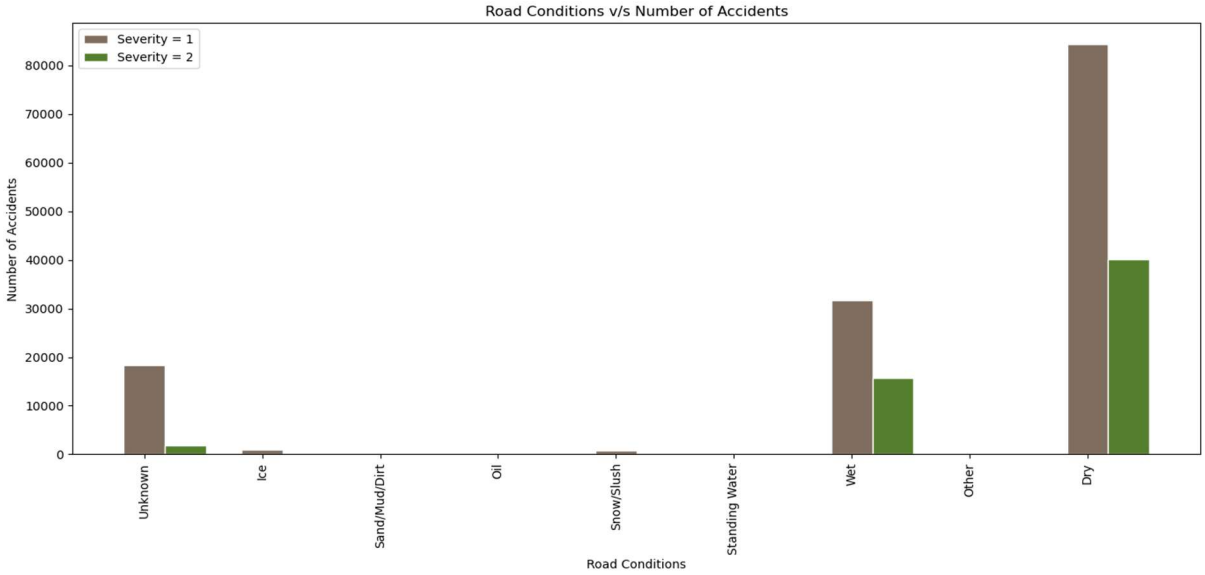


In this analysis, the peculiar point is that a significant number of accidents happened during the overcast. There should be no impact of the overcast on the driving experience but then, maximum number of accidents happened in the clear sky. Maybe they can be put into to same category.



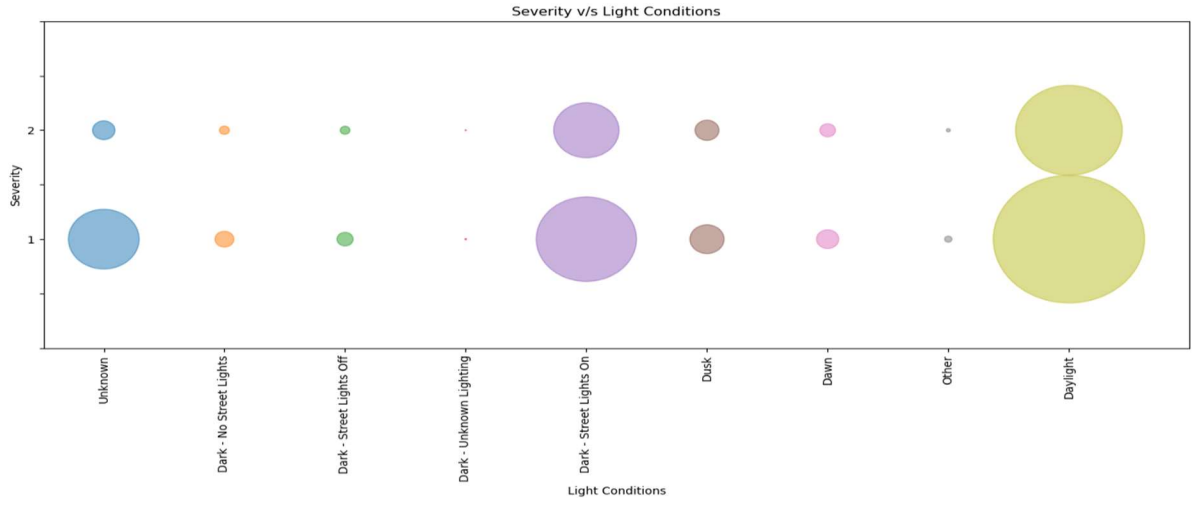
4.2 Relationship between Severity and Road Conditions –

With the growing populations and vehicles on the roads, the traffic accidents and related deaths have been ever-increasing worldwide and one the main reasons for this increasing numbers is condition of roads. A good quality road can give the rider/driver a remarkable experience whereas just the opposite of this can end his/her life. Be it due to bad weather or poor constituents manufacturing quality, there has been increasing concerns regarding the condition of roads. As it can be seen in the given below bar chart, ignoring the dry condition, which is much more common, wet road lead to second highest number of accidents. Other reasons include oil spillage, standing water, snow/slush and ice.

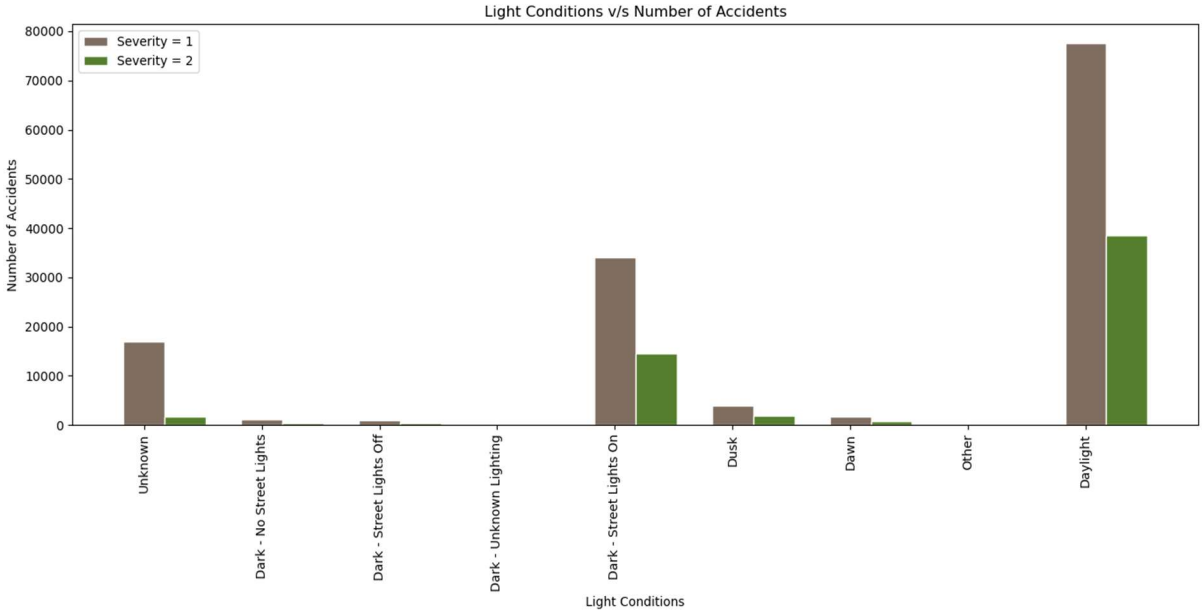


It can be seen in the graphs that for wet road conditions, approximately 33% of total accidents occurred were of severity 2. This is an alarming observation and to counter this, appropriate drainage system should be provided on the road. The wet roads have low friction coefficient that leads to severe accidents.

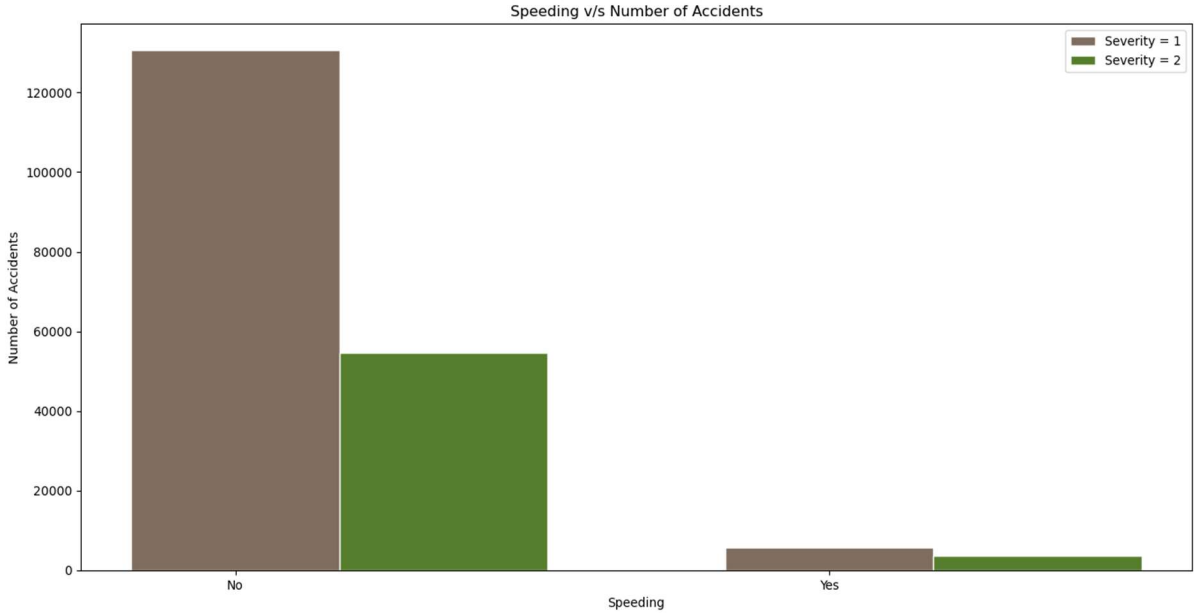
4.3 Relationship between Severity and Light Conditions –



Although major developments are being made to ramp up the vehicle headlights performance such as projector, LED, etc., the need for road lights can never be fulfilled with individual vehicle headlights. As we can see from the bubble chart given below, apart from daylight and dark – street lights on, the dawn, dusk, dark-street lights off and dark-no street lights have significant impact on the accident occurrences.



4.4 Relationship between Severity and Speeding –

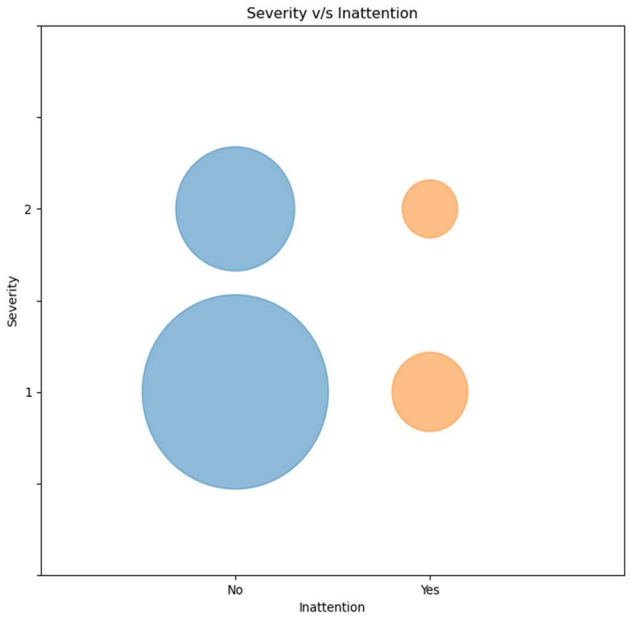


SPEEDING	SEVERITY = 1	SEVERITY = 2	SEVERITY = 2 CASES %AGE OF TOTAL CASES
YES	5802	3531	37.83%
NO	130638	54657	29.5%
YES CASES %AGE	4.25%	6.07%	

The above table gives us an alarming output that the possibility of accidents of severity = 2 increases with the speeding. Speeding causes loses not only to the property but to lives also. For vehicles with speeding charge, 37.83% of total cases were of severity = 2 which is 29.5% for cases where vehicles were not speeding.

4.5 Relationship between Severity and Inattention –

Driver inattention and distraction is one of the challenging issues being faced by the Road Safety departments. It is a human error so cannot be resolved by Road Safety Departments only. Individual contribution is largely required. There could be any reason for inattention. May be a father talking to family during driving or a businessman talking on the phone while driving. There have been several attempts to keep people's attention on the road. One of them is "Musical Road". A musical road is a road, or section of a road, which when driven over causes a tactile vibration and audible rumbling that can be felt through the wheels within car body. This rumbling is heard within the car as well as the surrounding area, in the form of a musical tune.

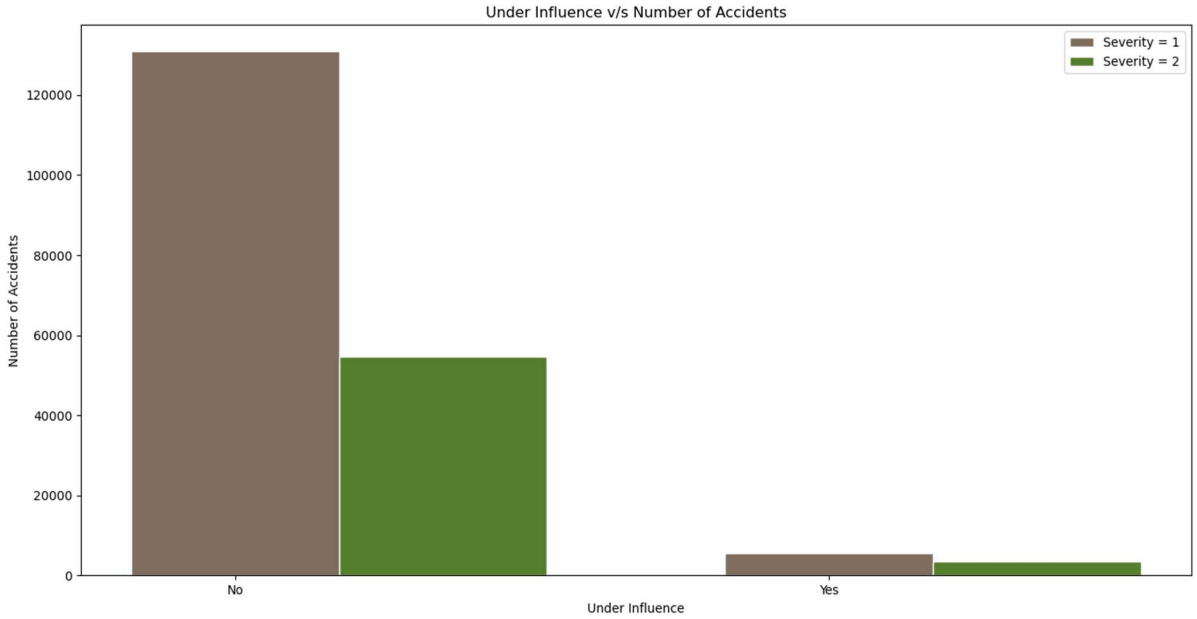


INATTENTION	SEVERITY = 1	SEVERITY = 2	SEVERITY = 2 CASES %AGE OF TOTAL CASES
YES	19408	10397	34.88%
NO	117077	47791	28.99%
YES CASES %AGE	14.22%	17.87%	

The above given table shows that percentage of severity = 2 cases in accidents where inattention was the reason is 34.88% which is way higher than 28.99%, percentage of severity = 2 cases in accidents where inattention was not the reason. This concludes that inattention increases the chances of severe accidents. It is also being supported by the numbers that

17.87% of severity = 2 cases were due to inattention whereas this stands at 14.22% for severity = 1 cases.

4.6 Relationship between Severity and Operator Under Influence –



INFLUENCE	SEVERITY = 1	SEVERITY = 2	SEVERITY = 2 CASES %AGE OF TOTAL CASES
YES	5559	3562	39.05%
NO	130926	54626	29.44%
YES CASES %AGE	4.07%	6.12%	

Don't drink and Drive! A popular road safety sign which has somehow turned into troll nowadays. The severity of accident increases with the consumption of alcohol. This perception has been truly backed by the above data. The above given table shows that percentage of severity = 2 cases in accidents where the operator was under the influence was the reason is 39.05% which is way higher than 29.44%, percentage of severity = 2 cases in accidents where this was not the reason. Also, 6.12% of severity = 2 cases were due to the reason that operator was under the influence of some drug whereas this stands at 4.07% for severity = 1 cases.

All the abovementioned data can be used to predict the severity of the accident given the proper inputs are given to the model. These numbers play major part in this modelling as they affect the severity of the accident and this is the reason, of all the given data, this portion of data will be used to predict the severity in this model.

5. Predictive Modelling

There are two types of models, Regression models and Classification models. Regression models are used to predict the values of the data showing continuous trend. For example, the coronavirus infection trend is continuous in nature because it makes a continuous line graph, be it linear or non-linear. Classification models are used to predict the class of a given input. For example, banks use customer data to predict whether to loan him the money he asks or not. Since we are also predicting the severity of a possible accident which is a classification, we will used classification modelling.

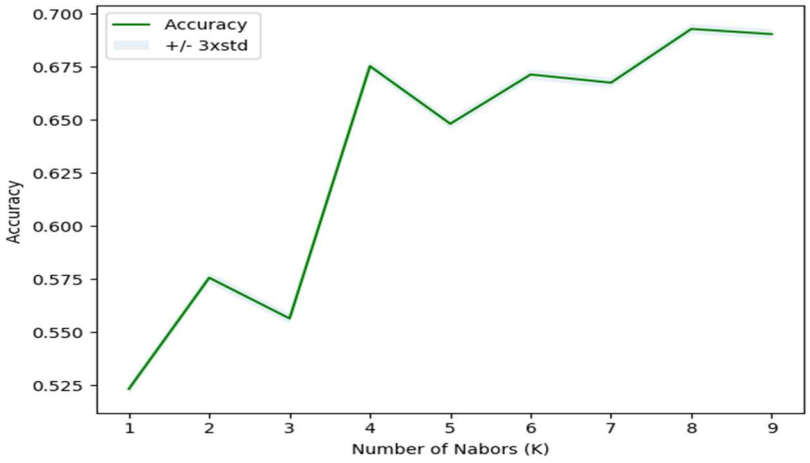
5.1 Classification Modelling

The application of classification model is very straightforward. In the data cleaning and preparation phase, we have already prepared the data. All the data has been converted into numerical form. There are many types of classification models and we are going to use, compare and select the best suited one from following four models among them –

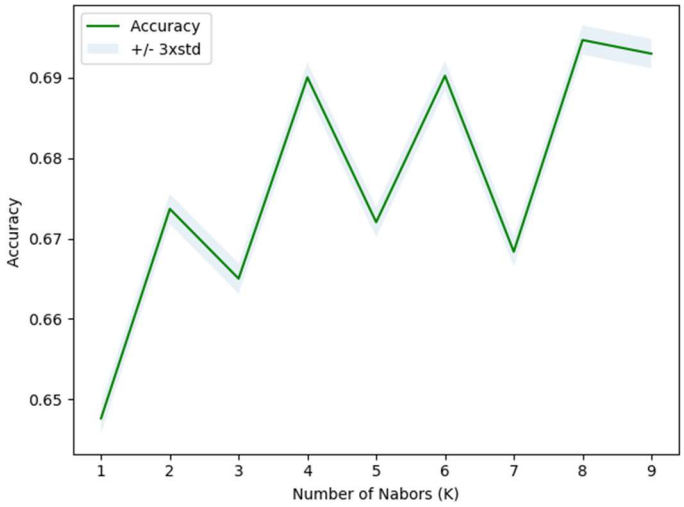
1. K-Nearest Neighbors Model
2. Decision Trees Model
3. Logistic Regression Model
4. Support Vector Machine Model

1. K-Nearest Neighbors Model –

In this classification model, the value of "k", which is the number of neighbors, was initially taken as 4 and the test dataset size was 20% of total dataset. Then the model was trained with train datasets (X_train & Y_train) and values were predicted with test dataset (X_test). The predicted values were compared and accuracy of the model was predicted which stood at 0.671 for training dataset and 0.675 for test dataset. After this the model was tested for different values of k ranging from 1 to 9 and accuracy was computed. For k = 8, the accuracy value was maximum at 0.693.

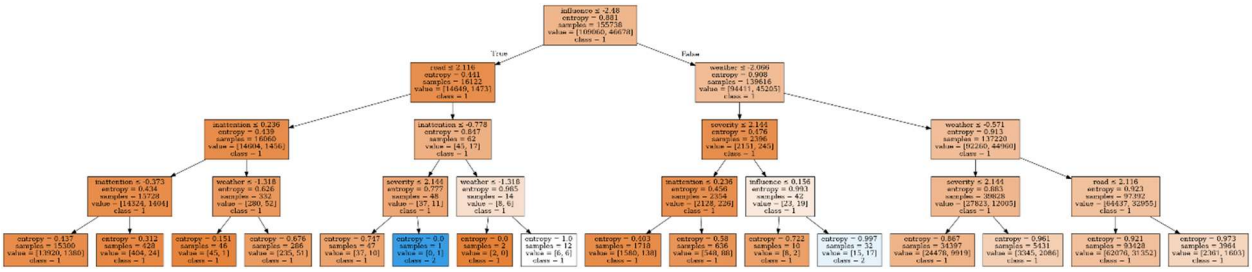


Further to this, the model was again tested in similar way for test dataset size of 33% of total dataset. In that condition also, the best value of accuracy was at k = 8.



2. **Decision Trees Model –**

For this classification model, the test dataset size was again set at 20% of total dataset. The model was built at max_depth = 4 and had the accuracy of 0.70437. Since sklearn.external.six has been removed form the sklearn 0.23.2, the decision tree has been plotted in another notebook with old sklearn library version.



3. **Logistic Regression Model –**

Using the same test and training datasets, this model was trained and tested and it turned out that this model has accuracy of 0.70445.

4. **Support Vector Machine Model –**

Similarly, the model was also used to predict the values of X_test dataset and it has the highest accuracy among the four built models at 0.70448. Running this model for higher dimension data is very time consuming and endless may be. Since I had no access to CPU with higher configuration, F1 score and Log Loss for Support Vector Machine Model could not be calculated.

6. Results

The objective of this study was to build a model which could predict the severity of possible accident on the roads of Seattle, WASHINTON. Since the output is classification, classification models have been used as explained previously. The aim was to build a model with highest accuracy given the appropriate inputs are given to the model to feed. This task has been successfully completed by building and comparing four models given below in the table with their accuracy, F1 Score and Log Loss –

SERIAL NUMBER	MODEL	ACCURACY	F1 SCORE	LOGLOSS
1.	**K – Nearest Neighbors**	0.693	0.598	1.40
2.	**Decision Tree**	0.704	0.582	0.588
3.	**Logistic Regression**	0.704	0.583	0.589
4.	**Support Vector Machine**	0.704	-	-

7. Discussion and Conclusion

All the models used in the predictive modelling have approximately equal accuracy. Decision Tree model, Logistic Regression model and Support Vector Machine model have exact equal accuracy whereas K-Nearest Neighbors model has slightly lower accuracy of 0.693. Since the value of accuracy should be higher in order to be the best model, K-Nearest Neighbors model lags in this criterion.

Additionally, F1 value should also be considered for selecting the best model. The F1 value ranges from 0 to 1, 1 being the best and 0 being the worst. K-Nearest Neighbors model has highest F1 value considering I was not able to compute the same for Support Vector Machine model due to higher CPU configuration requirements. K-Nearest Neighbors model has F1 score of 0.598 whereas for Decision Tree model and Logistic Regression model, it is 0.583 and 0.583 respectively. The K-Nearest Neighbors model leads in this criterion.

Finally, if we look at Log Loss for all the models except for Support Vector Machine model, Decision Tree model has the least value of 0.588. This is less than 1.40 of K-Nearest Neighbors and slightly less than 0.589 of Logistic Regression model. Since lower the value, best the model, Decision Tree model leads in this criterion.

Ultimately, the best predictive model for severity prediction is Decision Tree model which has highest accuracy and lowest Log Loss.