PREDICTING THE SEVERITY OF ACCIDENT

KISHAN SINGH

BUSINESS UNDERSTANDING

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides these reasons, weather, visibility, or road conditions are also the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

The target audience of the project will be local government, police and car insurance companies. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries.

Here we have been given a dataset of all the accidents occured since 2004.

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

With the help of the provided dataset we will try to predict the main reasons behind those severity and how to reduce those situations.

Data Analysis

We have used the Seattle Collision dataset.
Dataset contains several attributes such as:

1. SEVERITYCODE
2. X
3. Y
4. OBJECTID
5. INCKEY
6. COLDETKEY
7. REPORTNO
8. STATUS
9. ADDRTYPE
10. INTKEY
11. LOCATION
12. EXCEPTRSNCODE
13. EXCEPTRSNDESC
14. SEVERITYCODE.1
15. SEVERITYDESC
16. COLLISIONTYPE
17. PERSONCOUNT
18. PEDCOUNT
19. PEDCYLCOUNT
20. VEHCOUNT
21. INCDATE
22. INCDTTM
23. JUNCTIONTYPE
24. SDOT_COLCODE
25. SDOR_COLDESC
26. INATTENTIONIND
27. UNDERINFL
28. WEATHER
29. ROADCOND
30. LIGHTCOND
31. PEDROWNOTGRNT
32. SDOTCOLUMN
33. SPEEDING
34. ST_COLCODE
35. ST_COLDESC
36. SEGLANEKEY
37. CROSSWALKKEY
38. HITPARKEDCAR

Our predictor or target variable will be **'SEVERITYCODE'** because it is used to measure the severity of an accident i.e., 0/1 within the dataset.

- 0 - Property damage only.
- 1- Severe Injury

Attributes used to weigh the severity of an accident are **' WEATHER '**, **' ROADCOND '**, **' ADDRTYPE '**, **'COLLISIONTYPE'** and **'LIGHTCOND'**.

Extract Dataset & Convert

In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, when they should be numerical type.

Data Cleaning:

There were lot of imbalanced data and also some incomplete information which had to be analyzed and cleaned. Several unnecessary data has been removed so as to get only relevant data for our model.

We removed as well as replaced the null values according to our analysis .i.e., if the row contained more than 3 null values we removed that data row, else we replaced it.

Given is the relevant data frame which will be used for the model development.

	SEVERITYCODE	WEATHER	ADDRTYPE	COLLISIONTYPE	JUNCTIONTYPE	ROADCOND	LIGHTCOND
0	2	Overcast	Intersection	Angles	At Intersection (intersection related)	Wet	Daylight
1	1	Raining	Block	Sideswipe	Mid-Block (not related to intersection)	Wet	Dark - Street Lights On
2	1	Overcast	Block	Parked Car	Mid-Block (not related to intersection)	Dry	Daylight
3	1	Clear	Block	Other	Mid-Block (not related to intersection)	Dry	Daylight
4	2	Raining	Intersection	Angles	At Intersection (intersection related)	Wet	Daylight

Feature Selection:

After removing and modifying the undesired data, we selected only those features which will be required for our model building and development. The features which would be used for further stages would be:

1. SEVERITYCODE
2. WEATHER
3. ADDRTYPE
4. COLLISIONTYPE
5. JUNCTIONTYPE

6. ROADCOND
7. LIGHTCOND

We also selected our dependent as well as independent features:

SEVERITYCODE – Dependent Feature also known as predictor.

Others would be our Independent features through which we will predict the Severity of the accident.

Exploratory Data Analysis:

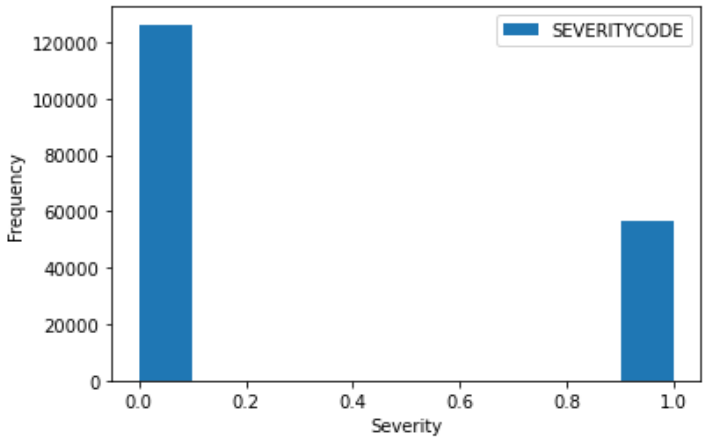
In this phase, we handle all those categorical features in our data frame. Basically, we applied two types of encoding techniques:

- **One Hot Encoding** – On ADDRTYPE as it had only 3 unique values.

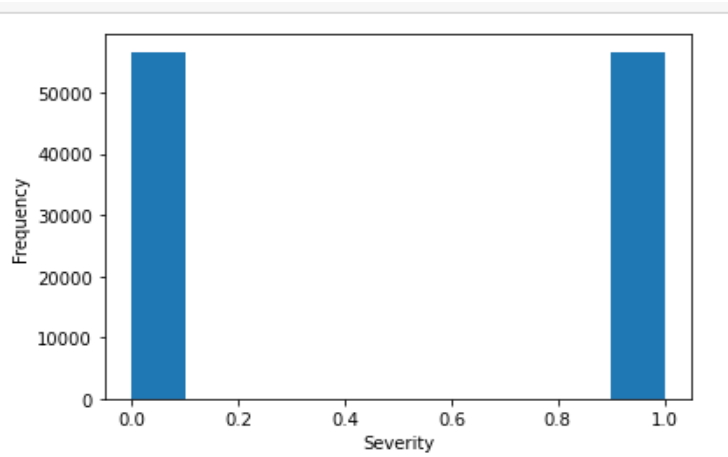
  I replace the ADDRTYPE column with 3 separate columns of those unique values so that I can get binary output that whether or not severe accidents happened in that type of area.

- **Frequency Count Encoding** – On WEATHER, COLLISIONTYPE , JUNCTIONTYPE , ROADCOND and LIGHTCOND.
  Here ,I replace the categorical feature values with the frequency ,i.e., total count of that value in that particular feature so that I can get a numerical data , which I can feed to my model in order to process further.

Here we can see that we have imbalanced predictor data , so we down-sampled our SEVERITYCODE data in order to balance the dataset.



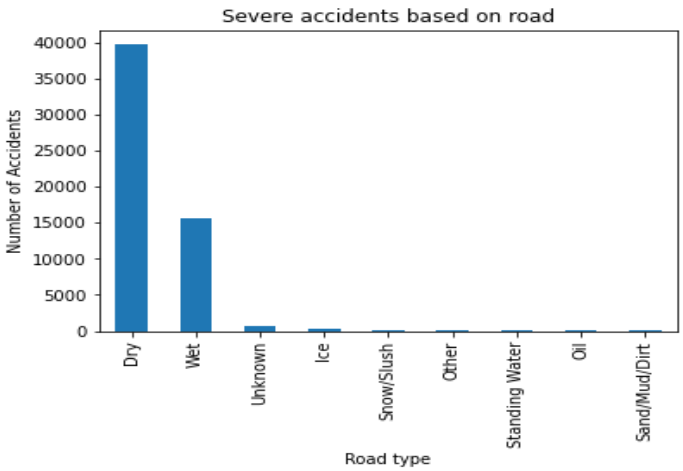
After down-sampling the majority occurred predictor:



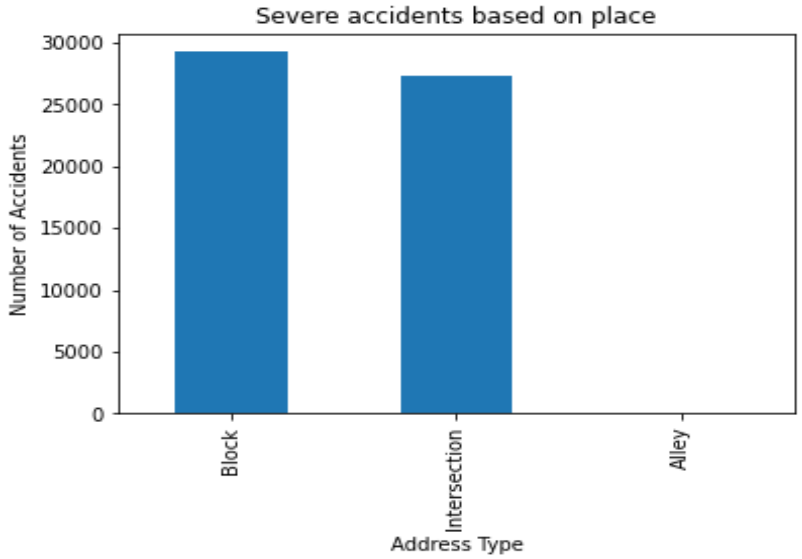
Now as we can see that both predictor values are equal.

We will check the frequency of occurring Severe accident across all independent features:

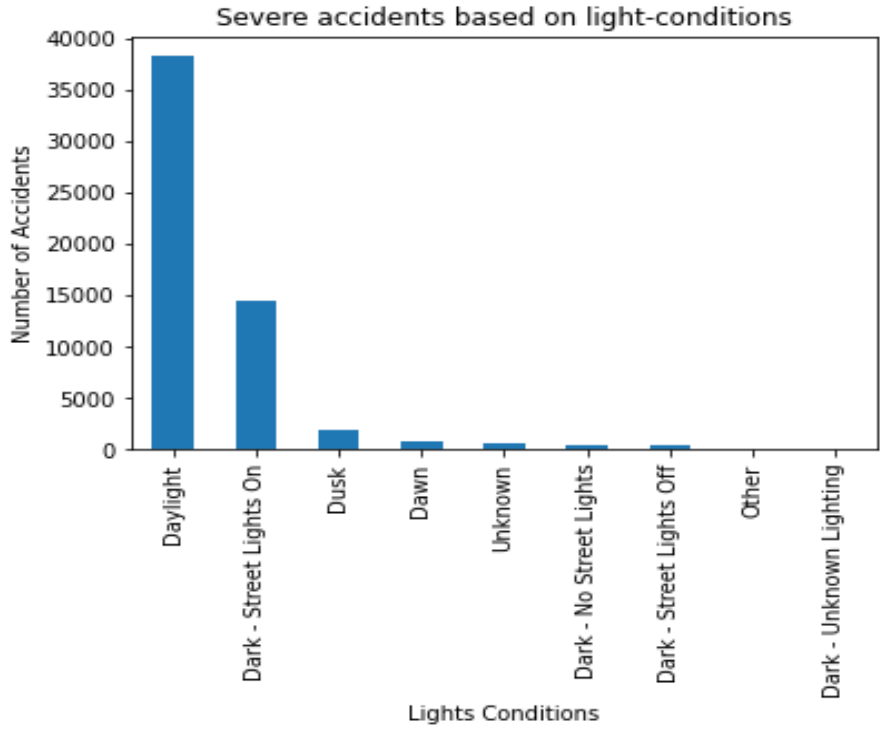
Severity based on Road condition: Here it is clearly visible that most of the severe accidents happened on dry road.



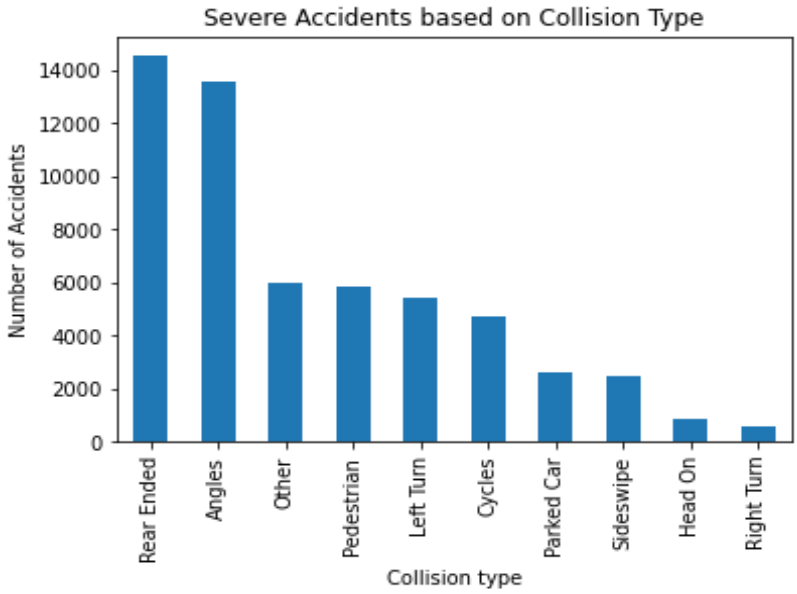
Severity based on types of address locations: Here we can see that most of the severe accidents happened in Block or Intersection. There are almost no such severe cases in Alley.



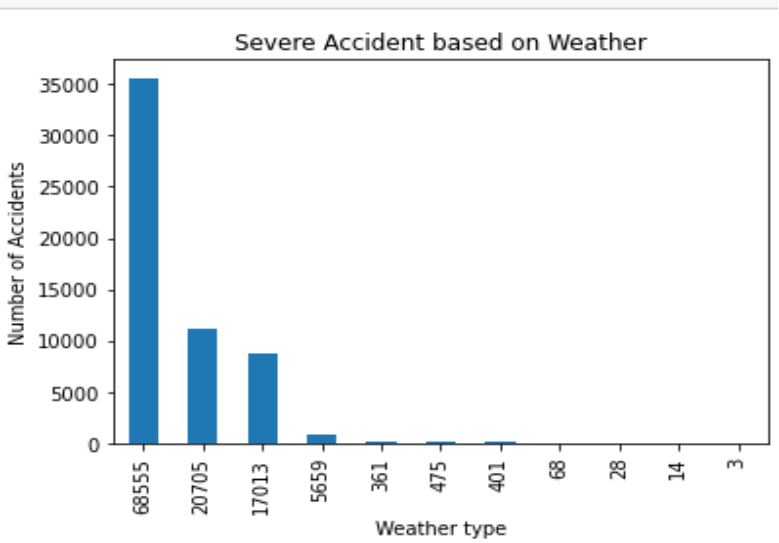
Severity based on Light Conditions: Here, we can see that most of the severe accidents happened in Daylight. Through this we can conclude that people are more careless in daytime in comparison to other time situations.



Severity based on type of Collision: Here, by this plot we can see that most of the severe accidents occurred from behind and angles.



Severity based on Weather: Here it is clearly visible that most of the severe accidents happened on dry road.



So, now we can proceed with our model development phase.

Model Development

After selecting the Dependent and Independent features, we split our data set for Training and Testing phase.

80% of the dataset would be used for training the model, while the rest 20% would be used for model evaluation phase. As, It is a classification problem i.e., whether the accident is severe or not , we developed 4 types of model to test the accuracy that which one gives the best output:

After selecting our training and testing data we normailize the training set using ***StandardScaler()*** function.

Decision Tree. –

The decision tree classifier creates the classification model by building a decision tree. Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. Each leaf represents class labels associated with the instance.

After fitting the decision tree classifier model with our training test, we predict the outcome and matched it with our test data set and got an accuracy of:

- **F1 Score: 0.691**
- **Accuracy-Score: 0.692**

K-Nearest Neighbors –

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

The KNN Algorithm

1. Load the data

2. Initialize K to your chosen number of neighbors

3. For each example in the data

3.1 Calculate the distance between the query example and the current example from the data.

3.2 Add the distance and the index of the example to an ordered collection

4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances

5. Pick the first K entries from the sorted collection

6. Get the labels of the selected K entries

7. If regression, return the mean of the K labels

8. If classification, return the mode of the K labels

After fitting the K-Nearest Neighbor classifier model with our training test, we predict the outcome and matched it with our test data set and got an accuracy of:

Accuracy-Score: 0.651

Logistic Regression

Logistic regression is a **classification** algorithm, used when the value of the target variable is categorical in nature. **Logistic regression** is most commonly used when the data in question has binary output, so when it belongs to one class or another, or is either a 0 or 1.

Here our outcome is also in the form of binary that is whether the accident is sever or not.

So after training the Logistic Regression Classifier model with our training dataset we predicted the outcome of severity with an overall accuracy of:

f1 score: 0.576
Accuracy score: 0.597

-

Support Vector Machine(SVM) –

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

So after training the SVM Classifier model with our training dataset we predicted the outcome of severity with an overall accuracy of:

F1 score: 0.705
Accuracy-score: 0.653

Classification report-

Classification Report:

	precision	recall	f1-score	support
0	0.73	0.48	0.58	11292
1	0.61	0.83	0.71	11365
accuracy			0.65	22657
macro avg	0.67	0.65	0.64	22657
weighted avg	0.67	0.65	0.64	22657

Discussion:

Through all those steps I found that, the best classification model for our given data set would be Decision Tree Classifier model which gives an overall accuracy of 70% correct predictions.

Also through all those analysis:

- It is clearly observed that people are more careless during daytime in comparison to dawn and night.
- Most of the severe accidents happens at Intersection and Blocks , so people should be more careful across intersections and blocks.
- Most of the severe accidents happens in clear weather. So, it is necessary to take measures in clear weather also.
- It is also noticed that most of severe accidents happens on dry roads.