Report on Coursera Capstone Project- Car accident Severity in Seattle

By:

Pilli Jethin Chandra

Introduction

The main business problem is about collisions in Seattle and conditions causing them. Intention is to build predictive model which will provide reasonable prediction if under certain condition is higher probability of collisions with certain severity (certain places on exact days) and which are the mostly involved on collision.

Audience consists of people who lives in Seattle or travel through it and goal is to provide them information about current situation on a roads and possible dangers.

Audience should be interested in this problem because knowing the relationships between conditions and likelihood of collision can save their money and life.

Data

This is a city car accident dataset, which covers the city of Seattle. The data set contains the road conditions, types of collisions, weather conditions, light conditions, type of junction and the severity code.

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

Some entries were missing crucial data that were required. Some columns were filled with “Unknown “in some columns. To rectify, I dropped the entire row.

Methodology

Exploratory data science:

Examining the environmental conditions:

First, I taken on what type of weather conditions the collisions happened mostly and what types of collision happened mostly.

Then, I choose to examine further the weather and road conditions in which each accident occurred.

Weather and Road Conditions:

- The proportion of L2 severity is higher when the weather condition is bad.
- The proportion of L2 severity is higher when the road condition is bad.
  This shows that Weather and Road Conditions have a big effect on accident severity.

Predictive Modelling

Machine Learning Models

Before model building, I have transformed all the categorical features using Label encoding method and normalized the datasets.

For modelling I have spilt the dataset into 80-20 ratio using train test split method. I.e. 80% as train set and 20% as test set.

Three classification algorithms were used and evaluated to predict the accident severity. Algorithms considered for classification were KNN algorithm, Decision Tree Classifier, Logistic Regression.

The first thing came to mind is that severity is based on different decisions like Road conditions and the weather. So, I tried ensemble methods as the data is very imbalanced.

KNN Test

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.

Decision Tree

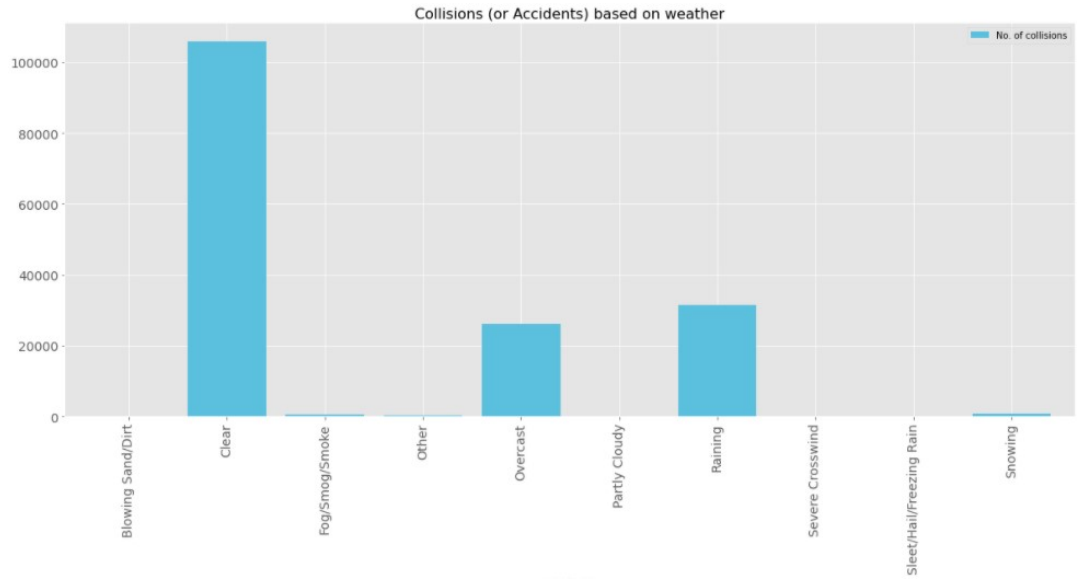
Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Logistic Regression

This model is basic and popular for solving classification problems. Logistics regression uses sigmoid function to deal with outliers. Class weight parameters sets the weights for imbalanced classes by adjusting weight inversely proportional to class frequency.

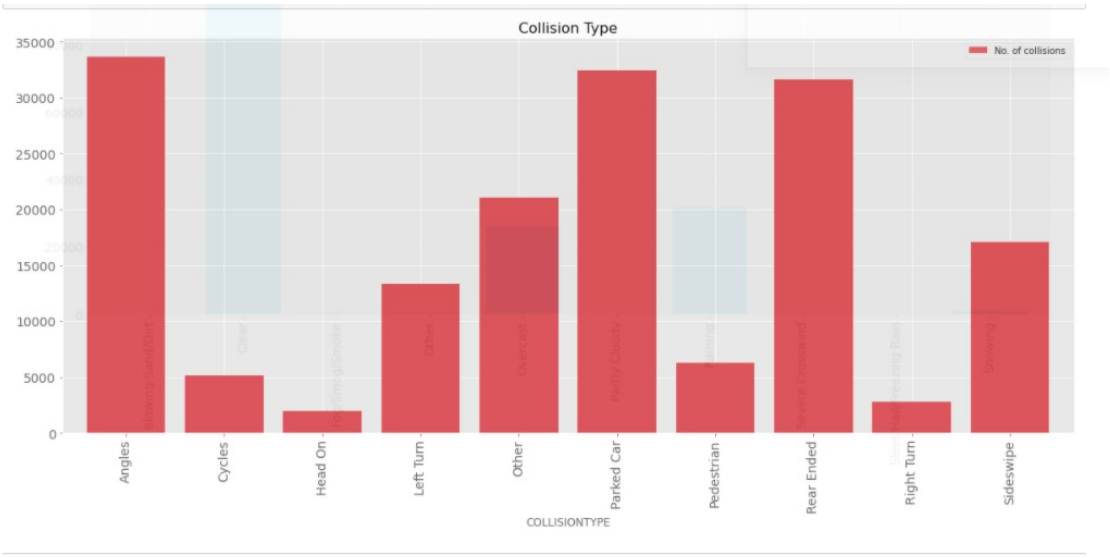
Results

The following graph shows the no. of collisions based on weather:



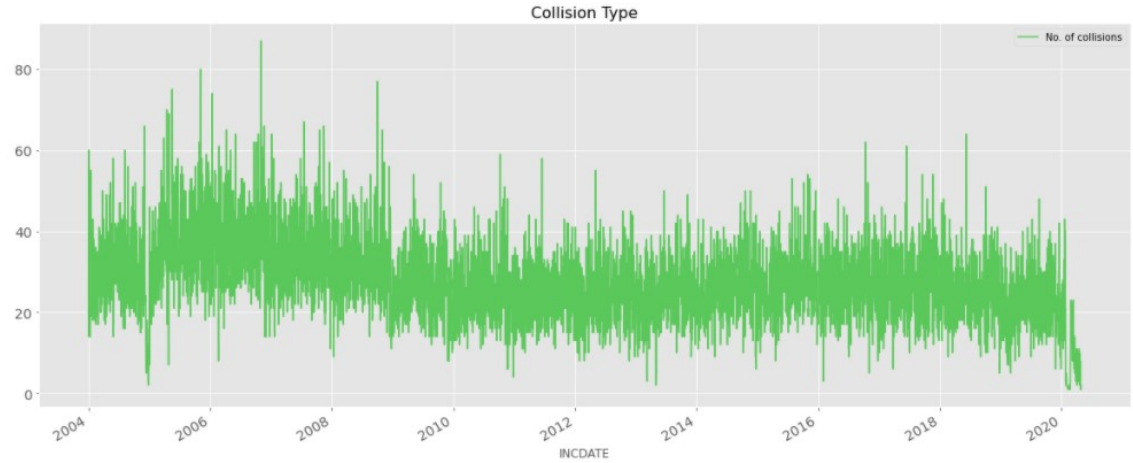
So, from the following graph we can say that the collisions happened mostly on clear weather.

The following graph shows the collision type:



From this we can say that the most no. of collisions are of angle type.

The following graph shows the no. of collisions happened form 2004 to 2020:



The accuracy and f1-score of each model is as shown below:

```
F1 score (KNN):  0.5333204189555325
Accuracy socre(KNN):  0.6666565772746534

F1 score (Tree):  0.5333204189555325
Accuracy socre(Tree):  0.6666565772746534

F1 score (log):  0.5333204189555325
Accuracy socre(log):  0.6666565772746534
```

The expectation, going on this investigation, was that bad weather and poor road conditions was dangerous, and that this would be evidenced by a pattern of higher crash rates have been supported by evidence.

Discussion

Road traffic injuries can be prevented. Governments need to take action to address road safety in a holistic manner. This requires involvement from multiple sectors such as transport, police, health, education, and actions that address the safety of roads, vehicles, and road users.

Effective interventions include designing safer infrastructure and incorporating road safety features into land-use and transport planning, improving the safety features of vehicles, improving post-crash care for victims of road crashes, setting and enforcing laws relating to key risks, and raising public awareness.

Conclusion

The objectives of the investigation were met, but there are still many areas of the data which could be investigated further, such as how junction layout or vehicle type relate to collision rates in different conditions. This model provides empirical evidence against weather and road conditions.