Applied Data Science Capstone Project

Car accident severity

Name: Shreyas Nayak

Date: 25/10/2020

Table of contents

1. Introduction/Business Problem

1.1 Background

Seattle, also known as the Emerald city, is Washington State's largest city, with home to a large tech industry with Microsoft and Amazon headquartered in its metropolitan area.

The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability. Worldwide, approximately 1.35 million people die in road crashes each year, on average 3,700 people lose their lives every day on the roads and an additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities.

1.2 Problem:

The project aims to predict how severity of accidents can be reduced based on a few factors such as weather, road, light conditions etc.

1.3 Stakeholders:

This project may be beneficial to the Public Development Authority of Seattle and the car drivers.

2. Data Understanding

Severity of accidents in Seattle city is the Data set. The Data set has 194673 rows and 38 columns. Our predictor or target variable will be 'SEVERITYCODE' because it is used to measure the severity of an accident. Attributes used to weigh the severity of an accident are 'COLLISIONTYPE', 'WEATHER', 'ROADCOND', 'LIGHTCOND','JUNCTIONTYPE', 'LOCATION', 'PERSONCOUNT' and 'VEHCOUNT'.

Description of the attributes:

1. **COLLISIONTYPE**: Describes the type of crash.
2. **WEATHER**: Describes the weather at the time of crash.
3. **ROADCOND**: Describes the condition of the road at the time of crash.
4. **LIGHTCOND**: Describes the light conditions at the time of crash.
5. **PERSONCOUNT**: The total number of people involved in the collision.
6. **VEHCOUNT**: The number of vehicles involved in the collision.
7. **JUNCTIONTYPE**: Category of junction at which collision took place.
8. **LOCATION**: Description of the general location of the collision.

3. Data Preparation

The Dataset in its original form is not fit for analysis. It has many missing values. So first we have to drop the missing values. After dropping the missing values the dataset contains 182660 rows.
Next i performed some Data Visualization. I found out that most of the collisions or accidents took place when the vehicles were parked and in clear weather, dry road and during daytime.

Fig 1: Collision types

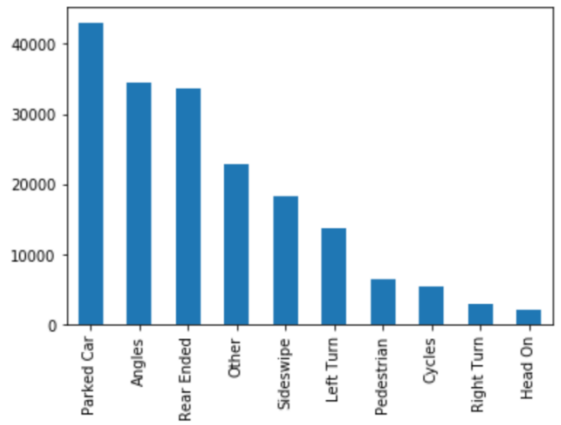
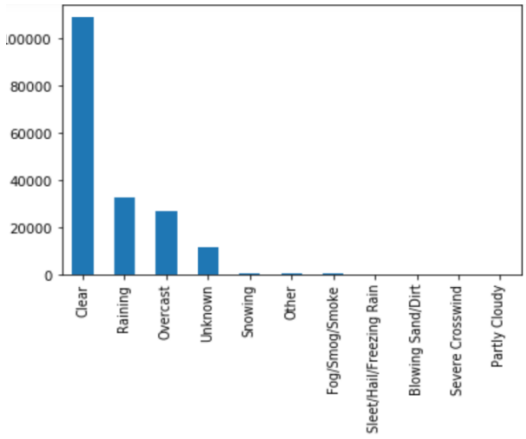
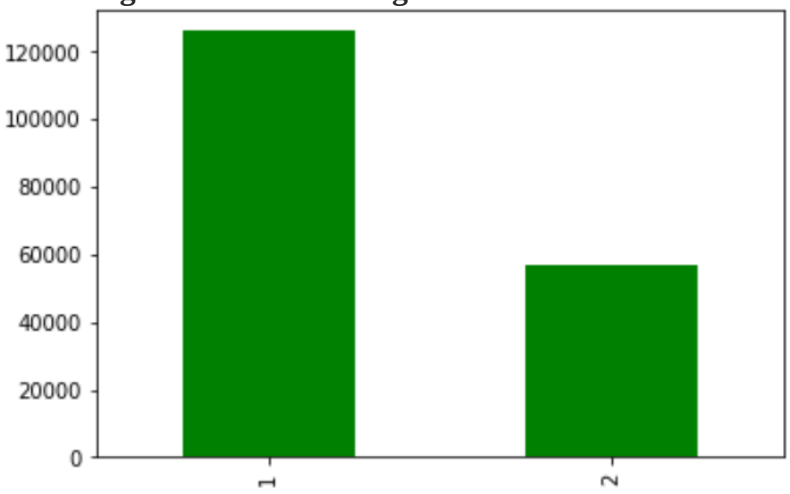


Fig 2: Weather conditions



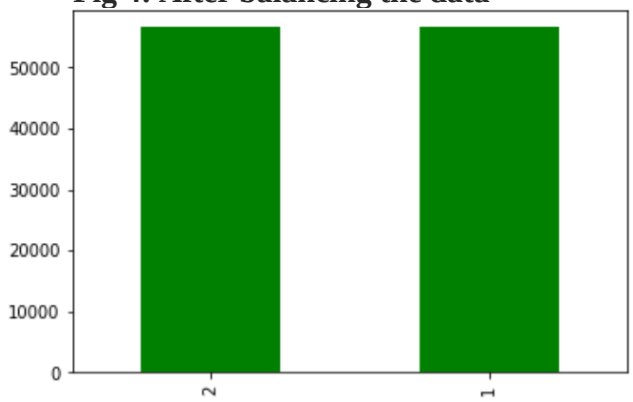
The dataset was found to be imbalanced. Data imbalance means unequal distribution of classes within a dataset. It was noted that 126064 accidents lead to property damage and 56596 accidents lead to injuries.

Fig 3: Before balancing the data



If we train a binary classification model without fixing this problem, the model will be completely biased. So I used Undersampling. Undersampling is the process where you randomly delete some of the observations from the majority class in order to match the numbers with the minority class.
After undersampling the dataset, I plotted the graph again and it showed equal number of classes as shown in fig 4 .

Fig 4: After balancing the data



Most of the columns where categorical data. So I used Label encoding. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Now our data is ready to be fed into machine learning models.

4. Modelling

The machine learning models used are SVM, Decision Tree and k-Nearest Neighbour.

4.1 SVM

An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high-dimensional feature spaces.

4.2. The Decision Tree

The Decision Tree breaks down a data set into 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.

4.3. k-Nearest Neighbour

K nearest neighbour is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance).

5. Result

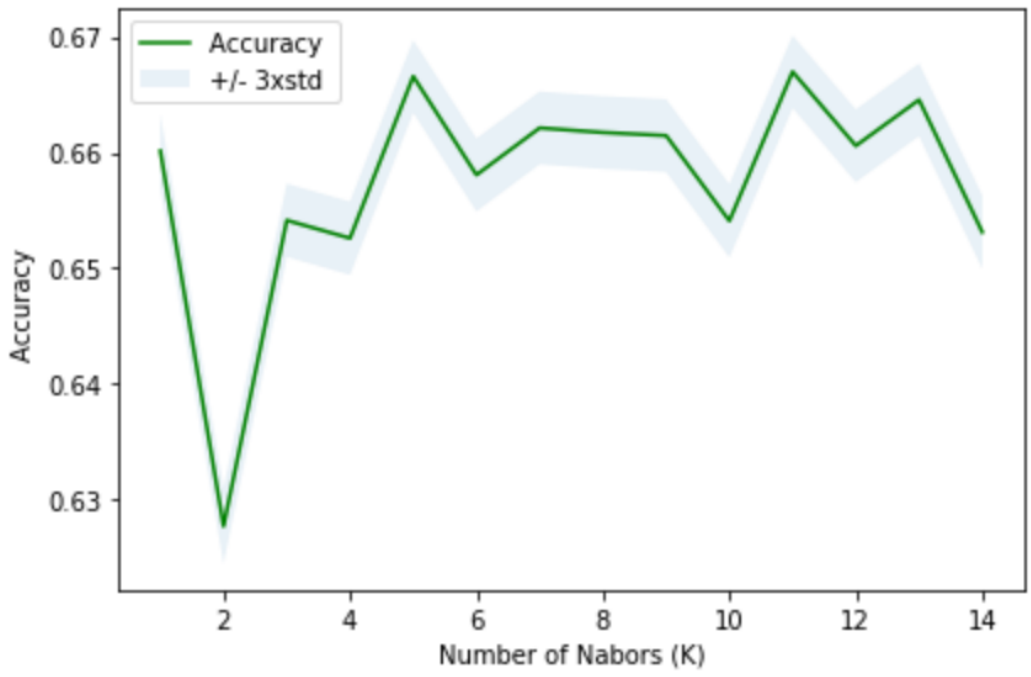
Decision Tree Classifier from the scikit-learn library was used to run the Decision Tree Classification model on the Car Accident Severity data. The criterion chosen for the classifier was ‘entropy’.

Classification report shows that the Decision Tree model has 70% accuracy.

	precision	recall	f1-score	support
1	0.75	0.59	0.66	16806
2	0.67	0.81	0.73	17152
accuracy			0.70	33958
macro avg	0.71	0.70	0.70	33958
weighted avg	0.71	0.70	0.70	33958

K-Nearest Neighbour classifier was used from the scikit-learn library to run the K-Nearest Neighbour machine learning classifier on the Car Accident Severity data. The best K, as shown below (figure 5), for the model where the highest elbow bend exists is at 5.

Fig 5: Accuracy vs K



Classification report shows that the KNN model has 67% accuracy.

	precision	recall	f1-score	support
1	0.70	0.59	0.64	11315
2	0.65	0.74	0.69	11324
accuracy			0.67	22639
macro avg	0.67	0.67	0.66	22639
weighted avg	0.67	0.67	0.66	22639

SVM classifier was used from the scikit-learn library to run the SVM machine learning classifier on the Car Accident Severity data.

	precision	recall	f1-score	support
1	0.74	0.60	0.66	16806
2	0.67	0.80	0.73	17152
accuracy			0.70	33958
macro avg	0.71	0.70	0.70	33958
weighted avg	0.71	0.70	0.70	33958

Classification report shows that the SVM model has 70% accuracy.

Average f1-score

f1-score is a measure of accuracy of the model, which is the harmonic mean of the model's precision and recall. Perfect precision and recall is shown by the f1-score as 1, which is the highest value for the f1-score, whereas the lowest possible value is 0 which means that either precision or recall is 0.

Our result shows that the f1-score of Decision tree and SVM are almost the same (0.66) where as that of KNN is 0.63. From these results we can assume that all the three f1-scores fairly good in predicting Property Damage and Injury.

Precision

Precision refers to the percentage of results which are relevant, in simpler terms it can be seen as how many of the selected items from the model are relevant.

The precision level to predict the "Property damage Collision" of Decision tree is highest (0.75) and that of KNN is lowest (0.70). However, the precision level to predict the "Injury Collision" of all the 3 models is almost same. From these results Decision tree is good in predicting Property Damage and Injury.

6. Discussion

A close examination of the dataset used in this project shows many missing values. If, we had large number of observations with consistency in values then the machine learning models would have given better accuracy and results. So it is a limitation of this analysis.

7. Conclusion

In this project, I analysed some of the factors which may lead to accidents. I built 3 machine learning models to predict the severity of accidents. This project may be beneficial to the Public Development Authority of Seattle and the car drivers as it may be useful in preventing future accidents in the city.