Coursera Data Science Capstone Project

Introduction

Every year, thousands of lives are lost in traffic accidents in the world and in the United States, tens of thousands of people are injured and billions of dollars in economic loss occur. In 2015, a crash occurred in Washington every 4.5 minutes. Seattle recorded the highest number of car accidents in the state that year, at 14,508 (in second place was Tacoma with just 4,756). Although the city is taking steps to make the roadways safer for citizens, vehicle collisions are still a serious danger. The economic costs of these crashes totaled $242 billion. Included in these losses are lost productivity, medical costs, legal and court costs, emergency service costs (EMS), insurance administration costs, congestion costs, property damage, and workplace losses.

In this project, some recommendations will be provided on how to prevent accidents or how to take measures to prevent accidents with less damage, depending on several factors.

Stakeholders:

- Public Development Authority of Seattle
- Car Drivers

Data

These traffic records were collected by the SPD (Seattle Police Department) from the year 2004 to 2020. The data consists of 37 attributes and 194,673 collision records. The data set used for this project can be found here!.

Then, I began choosing columns to use from the dataframe that I created.

```
1    136485
2     58188
Name: SEVERITYCODE, dtype: int64
```

The target columns is SEVERITYCODE, which assigns a crash a value of 1 (Property Damage Only) and 2 (Physical Injury) which were encoded to the form of 0 (Property Damage Only) and 1 (Physical Injury) .

Feature Variables	Description
INATTENTIONIND	Whether or not the driver was inattentive (Y/N)
UNDERINFL	Whether or not the driver was under the influence (Y/N)
WEATHER	Weather condition during time of collision (Overcast/Rain/Clear)
ROADCOND	Road condition during the collision (Wet/Dry..)
LIGHTCOND	Light conditions during the collision (Lights On/Dark with light on)
SPEEDING	Whether the car was above the speed limit at the time of collision (Y/N)

Data Cleaning

First of all, the total dataset was high variation in the lengths of almost every column of the dataset. The dataset had a lot of empty columns which could have been beneficial had the data been present there. These columns included pedestrian granted way or not, segment lane key, cross walk key and hit parked car. The project purpose is to analyze and predict the severity of an accident based on some particular features that will be chosen.

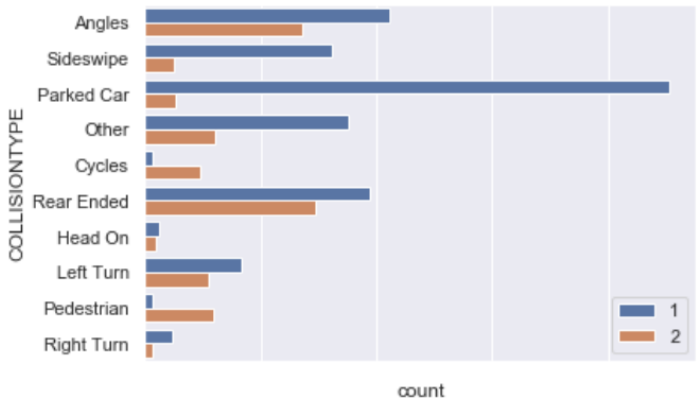
Exploratory Analysis

Let's explore the data to see if we can gather some knowledge from it and get some insights.

Most of our data columns are categorical and we need to know that how affect the severity of the accident.

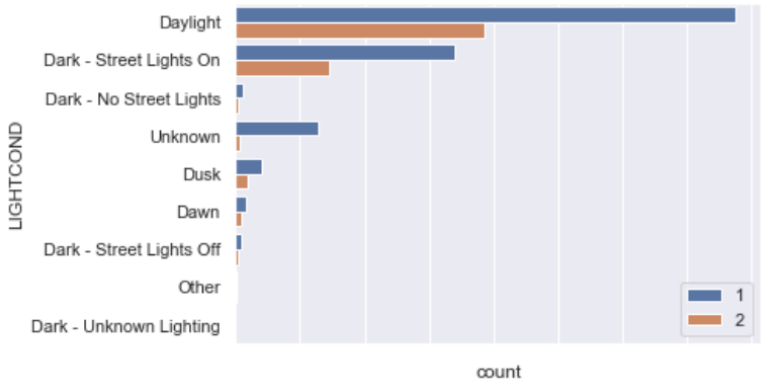
Type of Collision

Frequency of Property Damage Only Collision and Injury Collision with respect to collision type feature. This feature has different characteristics based in the area of impact, such as: angles, parked car, rear end, right turn, sideswipe, head on, left turn, pedestrian and cycles. All those variables, their frequency and the collision severity can be found in the figure below. As can be observed, the entropy of this categorical variable is pretty high (very unbalanced).

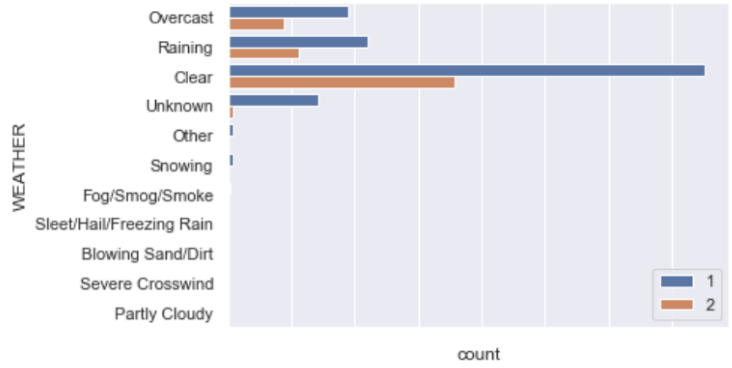


Weather, Light and Road conditions

There are more occurrences of severe collisions during daylight whereas during the night with the lights on, accidents tend to be less risky. The reason for this, may be related to a more cautious driving during the night which predispose car users to an aware state. Dusk and dawn tend to be related to more severe collisions, maybe because of the visibility reduction while facing the sun directly in the vision zone.



Considering weather data, severe accidents are slightly more frequent during rainy weather as well as with wet roads.



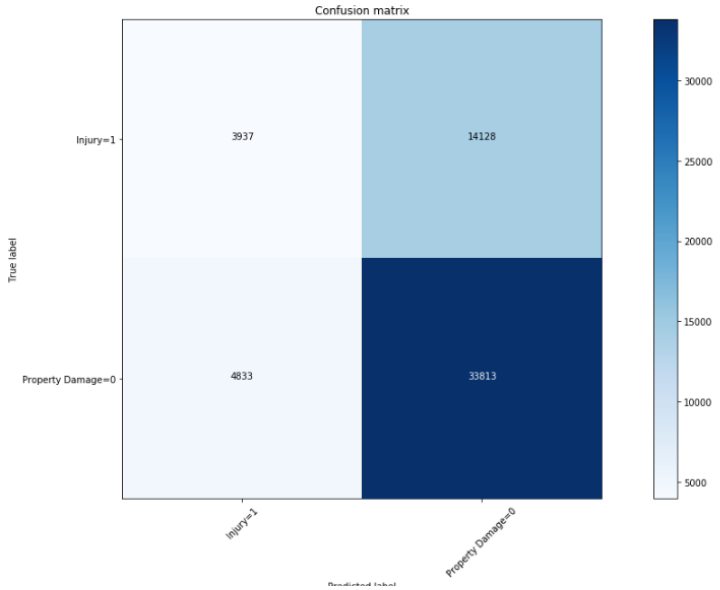
Methodology

Our data is now ready to be fed into machine learning models. We will use the following models:

K-Nearest Neighbor (KNN)

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

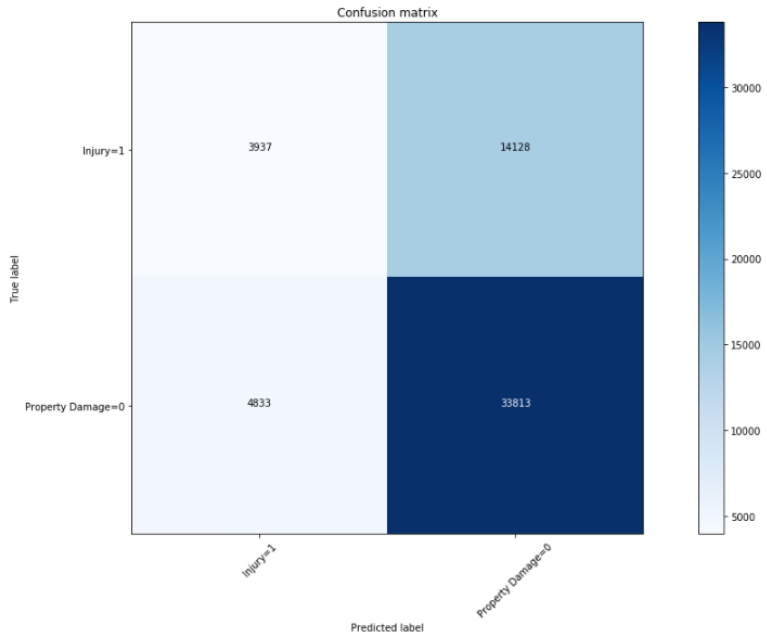
	precision	recall	f1-score	support
0	0.87	0.71	0.78	47941
1	0.22	0.45	0.29	8770
2	1.00	1.00	1.00	91
accuracy			0.67	56802
macro avg	0.70	0.72	0.69	56802
weighted avg	0.77	0.67	0.71	56802



Decision Tree

A decision tree model gives us a layout of all possible outcomes so we can fully analyze the concequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

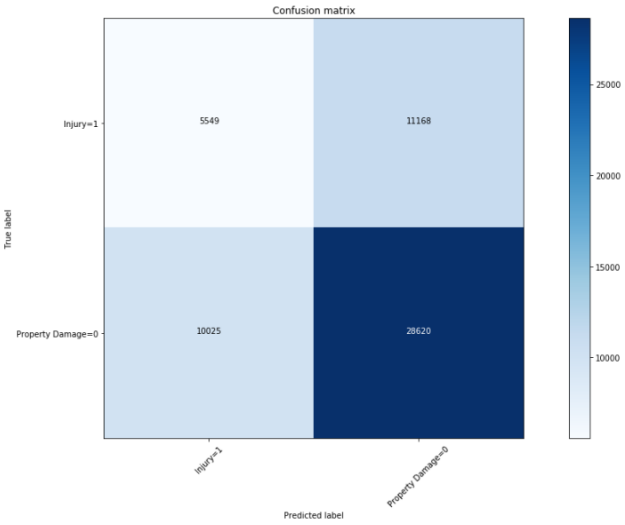
	precision	recall	f1-score	support
0	0.87	0.71	0.78	47941
1	0.22	0.45	0.29	8770
2	1.00	1.00	1.00	91
accuracy			0.67	56802
macro avg	0.70	0.72	0.69	56802
weighted avg	0.77	0.67	0.71	56802



Logistic Regression

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

	precision	recall	f1-score	support
0	0.72	0.74	0.73	38646
1	0.36	0.31	0.33	18065
2	0.06	1.00	0.12	91
accuracy			0.60	56802
macro avg	0.38	0.68	0.39	56802
weighted avg	0.60	0.60	0.60	56802



Discussions

Our data includes mostly categorical data that was of type 'object'. This is not a data type that we could have fed through an algoritim, so we created new classes that were of type int8. Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible

Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41
k-Nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

Conclusion

This project and analysis are quite helpful for the Seattle transportation department. Although this analysis has given us some good insight, there needs to be a closer inspection of certain other variables. It seems like a lot of these accidents are minor and avoidable. Before this, I thought that maybe weather, road, and light condition may cause more accidents, the results showed that it was not correct. Much of the data analyzed had revealed some important information about car accidents. Concerning the riskier ones which involve personal injuries, the focus has to be made in some important factors: intersections, rear end collisions, pedestrian and cycles. However ,caution have to be taken with rainy weather and wet roads, since after clear days and dry roads, these are the following conditions in order of importance.

Furthermore, there are some places which has more accidents during the dark time. For those places, adding lights might be a good solution to reduce the collisions. Also, when more cars involved in the accident, it seems that the level of severity will increase. They may need to be responded immediately to save more life.