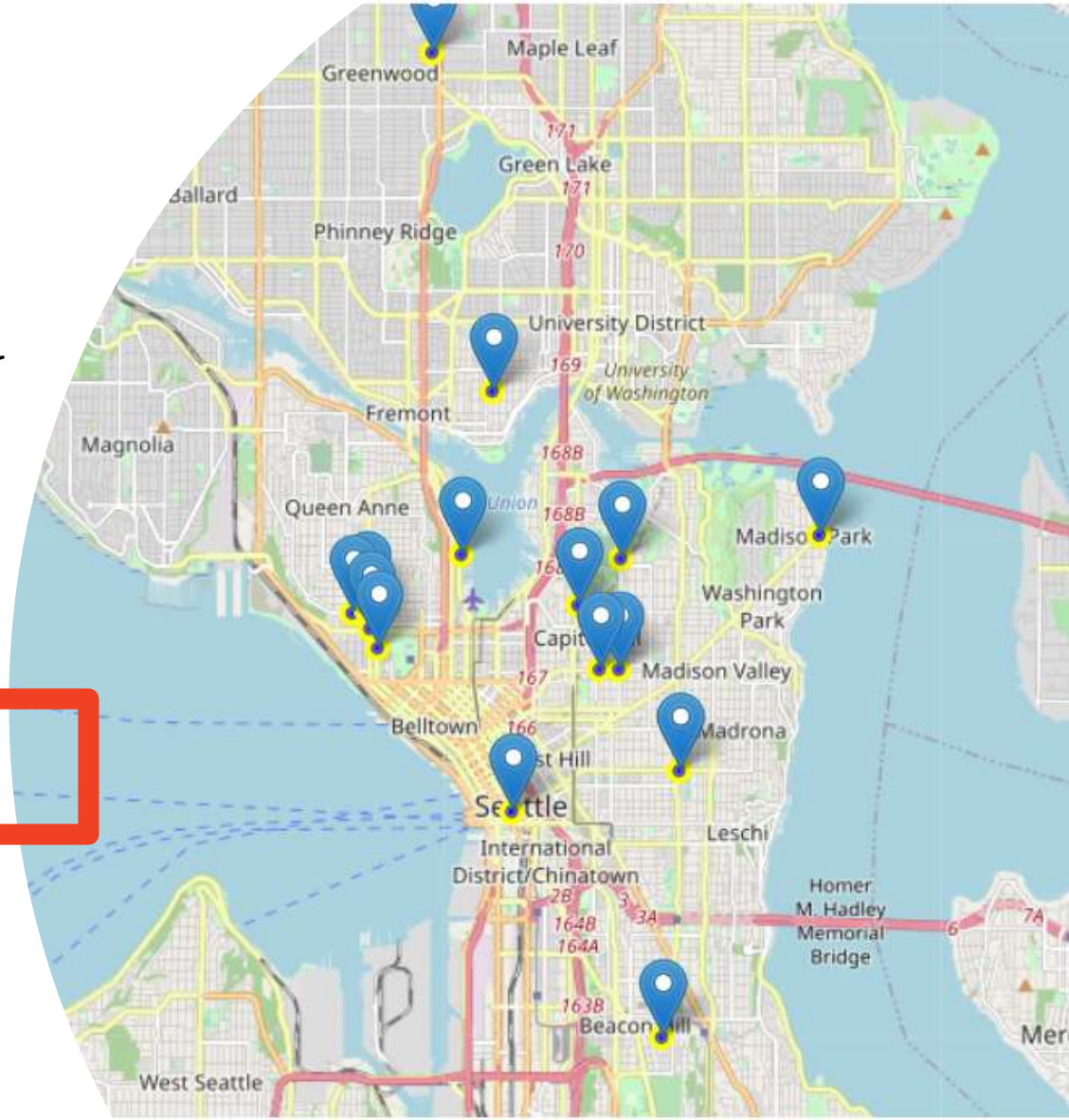
Capstone Project - Seattle, Washington Car accident severity

Chukwuemeka M. Okwuonu

October 2020



Introduction/Business Problem

Road accident has been a major human issue for ages, with the government enforcing more safety rules and standards, vehicle manufactures adding more safety feature to their vehicles, the fatality rate of car accident has not really reduced.

A study of Seattle Department of Transportation traffic accident cases from 2004 to 2020 will provide a good insight to some of the reasons behind these accident/collisions

This project will analyze the collision dataset, find patterns and determinate key factors using various analytical techniques and machine learning classification algorithms.

The target audience for this project is as follows:

- Seattle local government,
- The police
- Car insurance institutes.
- ALL DRIVERS

The Project's objective is to create a way to warn people based on factors such as weather and road conditions etc, the possibility of getting into a car accident and how severe it would be. So that the end user can drive more carefully or change travel plans if necessary

Data Section

The dataset for this project is from Seattle Police Department and Accident Traffic Records Department from 2004 to present, provided by Coursera IBM capstone project for downloading through a link.

It consists of 37 independent variables and 194,673 rows, also a dependent variable, SEVERITYCODE, that contains numbers which corresponds to different levels of severity caused by an accident/collision.

Severity codes are as follows:

- 0: Little to no Probability (Clear Conditions)
- 1: Very Low Probability — Chance or Property Damage
- 2: Low Probability — Chance of Injury
- 3: Mild Probability — Chance of Serious Injury
- 4: High Probability — Chance of Fatality

The indicator SEVERITYCODE is chosen as the dependent variable while other attributes like nature and human factors will enable us to build a model that will predict the chance of an accident and how severe it would be based on them.

Human factors are as follows INATTENTIONIND, UNDERINFL, and SPEEDING which shows the concentration of the driver's mind, drug or alcohol influence and overspeed. Nature factors are made up by WEATHER, ROADCOND and LIGHTCOND which represents weather, road and view circumstance, respectively.

Data Section Cont'd

Human factors are as follows INATTENTIONIND, UNDERINFL, and SPEEDING which shows the concentration of the driver's mind, drug or alcohol influence and overspeed.

Nature factors are made up by WEATHER, ROADCOND and LIGHTCOND which represents weather, road and view circumstance respectively

To fully understand the data, I run a value of the following features/attributes to those having most influence.

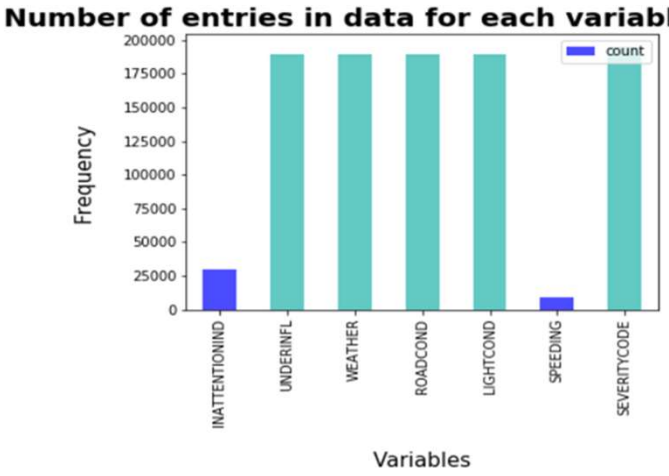
To fully understand the data, I run a value of the following features/attributes to those having most influence.

- ROADCOND: - The condition of the road during the collision.
- WEATHER: - A description of the weather conditions during the time of the collision
- LIGHTCOND: -The light conditions during the collision.
- UNDERINFL: - Whether a driver involved was under the influence of drugs or alcohol.

Methodology

Data Processing

Notebook from the IBM WATSON Studio is used to process the dataset. The data was imported, dtypes, shape and information of data was carefully studied and descriptive statistics bar plots of some variables to well understand their counts or frequency of data entry



Looking at the plot and after analyzing the dataset, the focus for an accurate result will be on only four features,

- WEATHER,
- ROADCOND,
- LIGHTCOND,
- UNDERINFL, among others with SEVERITYCODE been the dependent variable

Methodology

Data Processing

Severity Code was converted from (1/2) to (0/1) and a value counts shows huge imbalance between class 0 and 1. so we use a simple statistical technique to balance it.

```
df['SEVERITYCODE'].value_counts()
```

```
0    136485
1     58188
Name: SEVERITYCODE, dtype: int64
```

```
df_bal.SEVERITYCODE.value_counts()
```

```
1    58188
0    58188
Name: SEVERITYCODE, dtype: int64
```

Missing Data Heatmap was the technique used in the process of detecting and correcting dataset. With the removal of all unnecessary columns /null data from dataset, balancing Severity Code feature, and standardizing the input features, the data is now ready for machine learning models building.

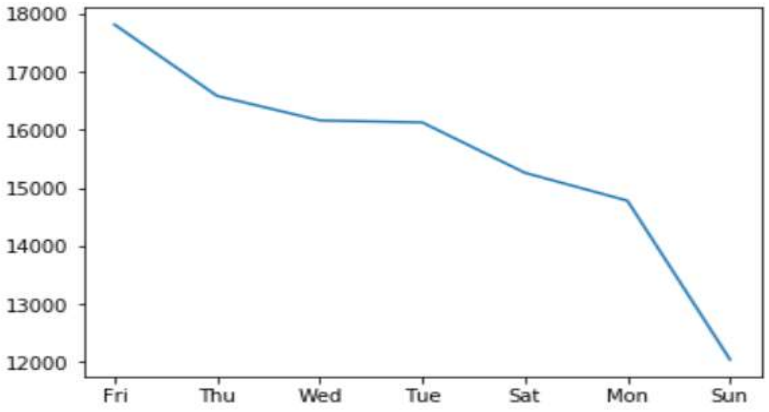
The following algorithms will be used in this project:

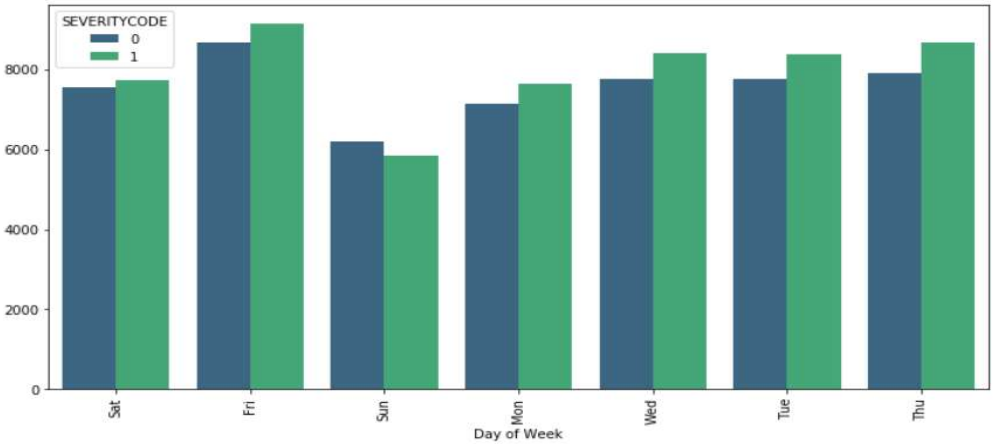
- K Nearest Neighbour - Grouping data points into categories/ groups based on similarity measures (or distance in between)
- Decision Tree - Breaking down the prediction into smaller subsets and generating a tree-like logic flow to model the prediction
- Logistic Regression - Using logistic functions to model binary output (dependent variable)

Discussion and Observation

To further understand and analysis the data, different plots were made to see the relationship between Day of the week, Month and the four main features/attributes to severity code and number of accident

Day of the Week



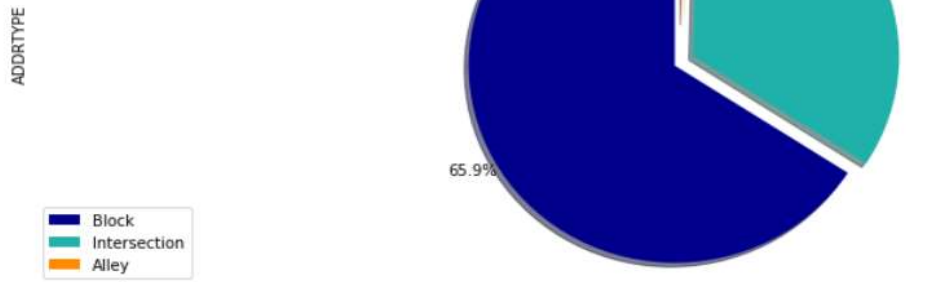


It was equally observed that more accident occur on Fridays and the trend slowly decreases to Sunday

Discussion and Observation Cont'd

Area of Accident

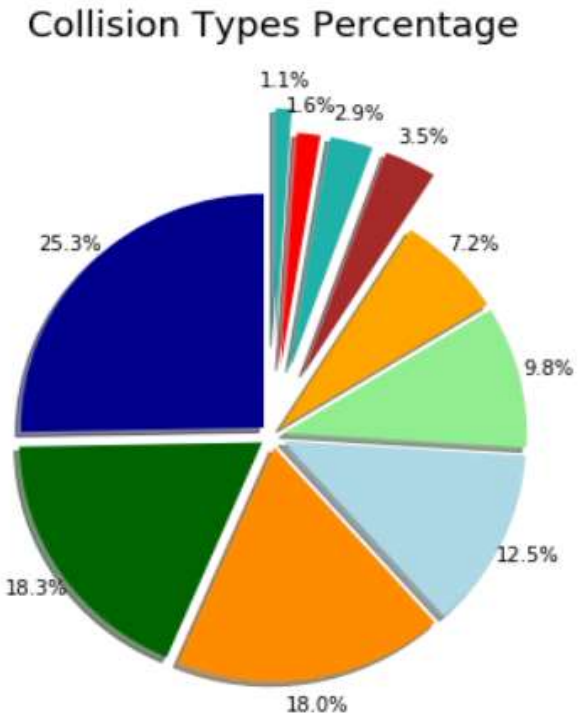
Seattle area plot indicates that the percentage of accident was high at the block followed by at intersection and then lower at the alley..



Discussion and Observation Cont'd

COLLISIONTYPE:

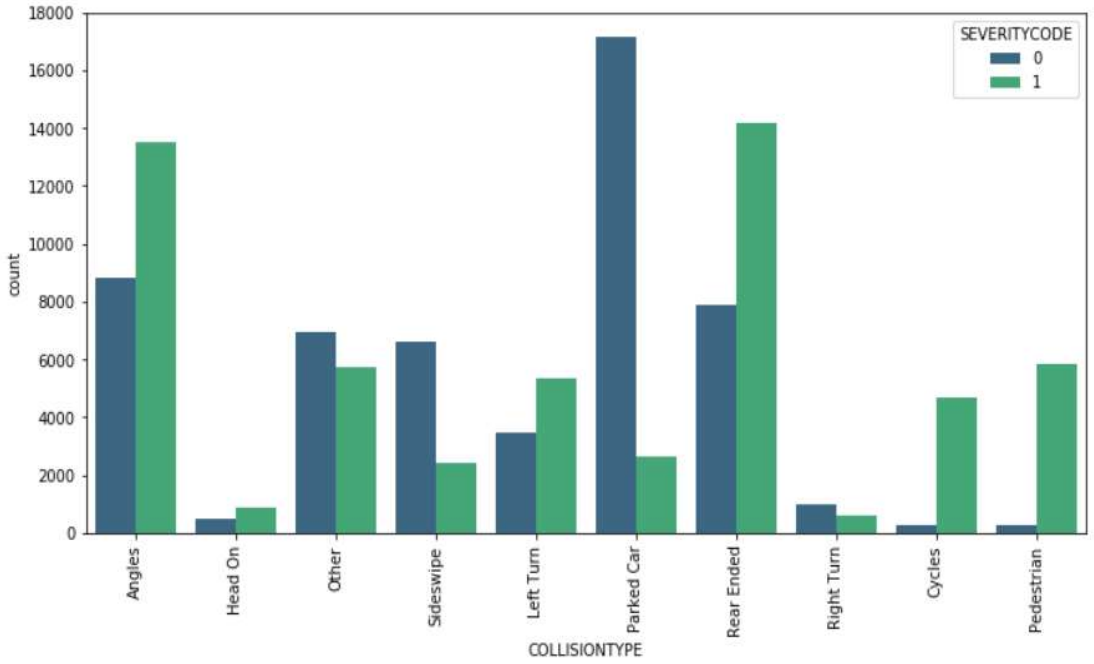
Collision types plot shows more accident occurring on parked cars with low physical injury and high property damage. Rear ended and Angles have more physical injury. high percentage at parked car followed by rear ended.



Discussion and Observation Cont'd

COLLISIONTYPE Vs Severity Code:

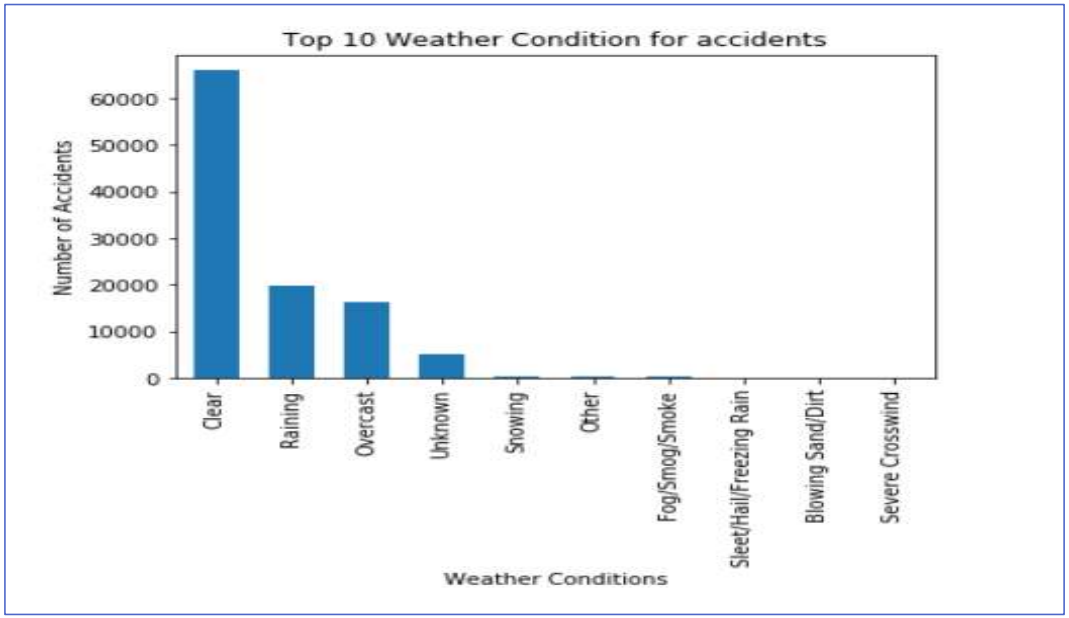
The plot shows high class 1 Severity code with Rear ended accident thus high property damage/injury and high class 0 Severity code with Parked Car .



Discussion and Observation Cont'd

Plots shows the number of accidents that occurred when the weather was clear, raining, snowing

Top 10 Weather Condition for accident



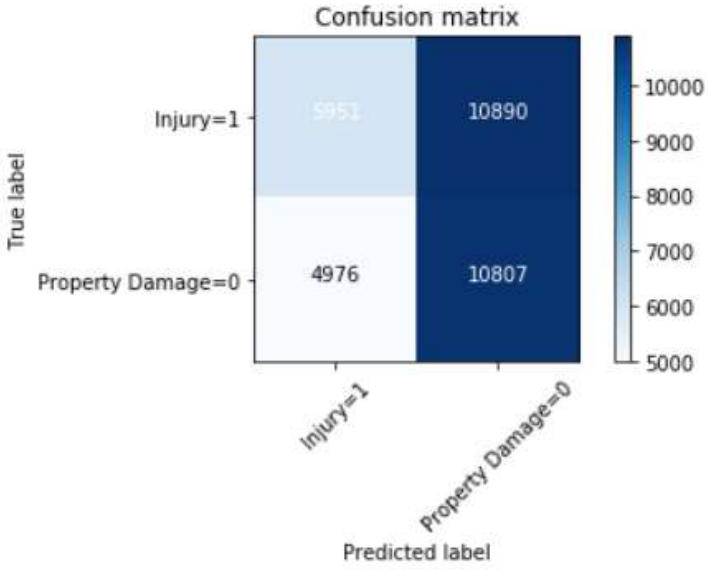
Results

K Nearest Neighbour

Train set Accuracy: 0.5115408363001012
Test set Accuracy: 0.5123528690534576
F1-score: 0.4856869229068246

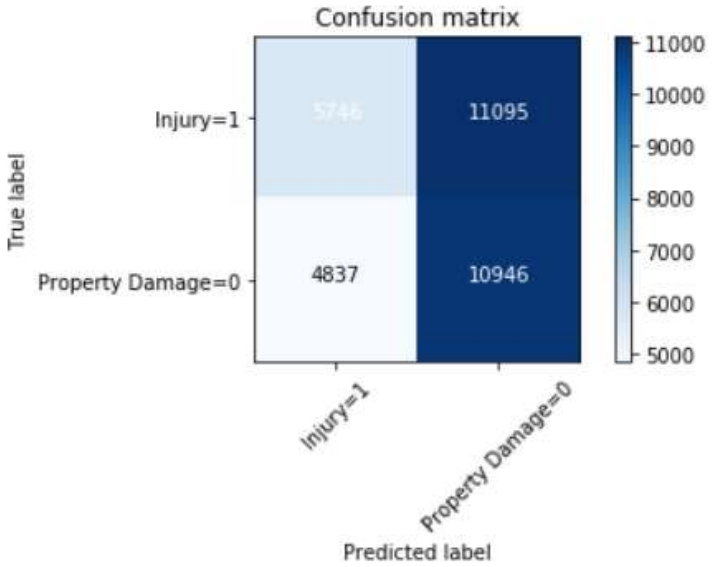
Decision Tree

	precision	recall	f1-score
0	0.68	0.50	0.58
1	0.35	0.54	0.43
accuracy			0.51
macro avg	0.52	0.52	0.50
weighted avg	0.57	0.51	0.53



Logistic Regression

	precision	recall	f1-score
0	0.50	0.69	0.58
1	0.54	0.34	0.42
accuracy			0.51
macro avg	0.52	0.52	0.50
weighted avg	0.52	0.51	0.50



Recommendations

The Seattle local government:

- Improve traffic policies, update public facilities such as streetlight, traffic signs and alert, etc.
- Warning and alert signs should include speed limits, road conditions and even weather when necessary.
- Barricades bad roads with potholes

Police:

- Enforcing the law and make sure traffic laws are followed
- Caution distracted drivers, remove ones under the influence of alcohol off the road.

Car insurance institutes:

- Check driver's license and records before issuing.

ALL DRIVERS:

- To follow/obey all traffic signs and laws, knowing that safety is very important and what their family or love ones will pass through if there was an accident.
- Assess to this information will enable them take extra precautions on the road under the given updates on light condition, road condition and weather, in order to avoid a severe accident.

Conclusion

- Weather, Road and Light Conditions have a great impact on the number of accident rate that results in property damage or injury
- Speeding equally have an impact to the accident rate but not as much as the four features/attributes mentioned above.
- k-Nearest Neighbor f1-score is low at 0.48 while Decision Tree and Logistic Regression with the average f1-score of 0.52 and 0.50 respectively are very close.
- Decision Tree and Logistic Regression models can be used side by side for the best performance and accurate predication