Contents

Predicting the severity of car accidents

(IBM Data Science Professional Certificate Capstone Project)

Ranga Rao

Executive Summary

In this project, we will be examining the factors impacting the severity of road accidents. Machine learning algorithms were used to classify road accidents. It was found that road accidents occur mostly due to poor road, lighting and weather conditions. Another factor that influences road accidents is road intersections.

Introduction

World Health Organization (WHO) provides the following statistics on Road traffic injuries

- Every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.
- 93% of the world's fatalities on the roads occur in low- and middle-income countries, even though these countries have approximately 60% of the world's vehicles.
- Road traffic injuries are the leading cause of death for children and young adults aged 5-29 years.
- Road traffic crashes cost most countries 3% of their gross domestic product.

It is imperative to study the accident data and understand the factors in depth and take necessary measure to reduce the severity and accidents using the Data science, Machine Learning techniques.

Road safety is a major concern for governments as well as major automobile corporations because there are over 38,000 people die every year in crashes on U.S. roadways out of which about 9 people die each day. There is an additional 4.4 million people who get seriously injured to require medical attention. Tech companies are also investing huge amounts of money in autonomous cars. Therefore, it is advantageous to look at car safety from a data perspective and determine what factors contribute to road accidents. We also look at the severity of accidents and build a model that predicts the severity of a road accident.

There are a number of factors that contribute to road accidents resulting in deaths and severity of injuries. These factors are:

- Weather conditions during the time of the collision.
- condition of the road during the collision
- Light conditions during the collision.
- Pedestrian right of way was not granted.
- Whether or not speeding was a factor in the collision
- Key for the lane segment in which the collision occurred.
- Collision involved hitting a parked car. (Y/N)
- Collision was due to inattention.

By understanding each of these factors and through planning, effective management and evidence-based interventions, road crashes can be predicted and prevented. Having access to accurate and updated information about the current road situation enables drivers, pedestrians and passengers to make informed road safety decisions. With the death toll of road accidents going up, governments as well as major automobile corporations are majorly concerned. For road users, be it cars or any vehicle that plies the roads, it is highly important to look at the safety of the vehicle from a data point of view and determine what factors contribute to road accidents. This project also assesses at the severity of accidents and build a model that predicts the severity of a road accident that will help commuters to decide in choosing an alternate route.

Stakeholders

The reduction in severity of accidents can be beneficial to the Public Development Authority of Seattle which works towards improving those road factors and the car drivers themselves who may take precaution to reduce the severity of accidents.

Understanding Data

This is an extensive data set from the Seattle Police Department, with over190,000 observations collected over the last 15+ years. To accurately build a model to preventfuture accidents and/or reduce their severity, we will use the following attributes

- ADDRTYPE
- WEATHER
- ROADCOND
- VEHCOUNT
- PERSONCOUNT

Data Cleaning

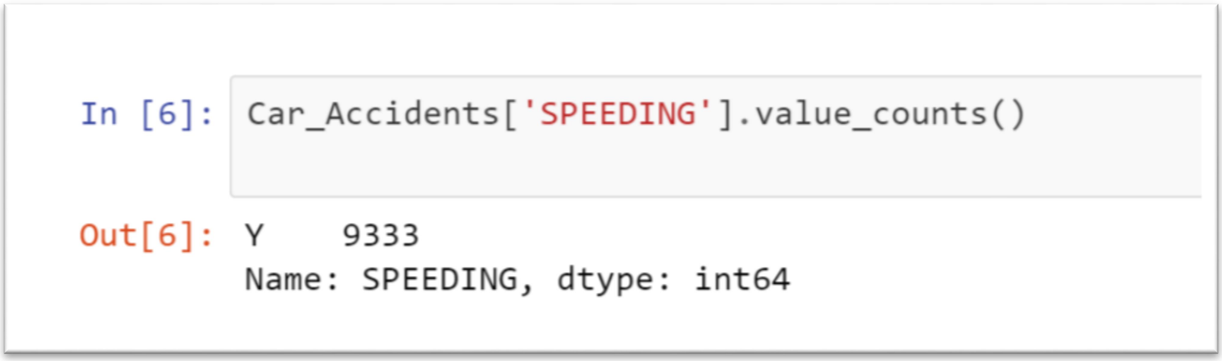
There are a lot of problems with the data set keeping in mind that this is a machine learning project which uses classification to predict a categorical variable. The dataset has total observations of 194673 with variation in number of observations for every feature. First of all, the total dataset was 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 models aim was to predict the severity of an accident, considering that, the variable of Severity Code was in the form of 1 (Property Damage Only) and 2 (Injury Collision) which were encoded to the form of 0 (Property Damage Only) and 1 (Injury Collision). Furthermore, the Y was given value of 1 whereas N and no value was given 0 for the variables Inattention, Speeding and Under the influence. For lighting condition, Light was given 0 along with Medium as 1 and Dark as 2. For Road Condition, Dry was assigned 0, Mushy was assigned 1 and Wet was given 2. As for Weather Condition, 0 is Clear, Overcast is 1, Windy is 2 and Rain and Snow was given 3. 0 was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity. Whereas, there were unique values for every variable which were either ‘Other’ or ‘Unknown’, deleting those rows entirely would have led to a lot of loss of data which is not preferred.

Methodology

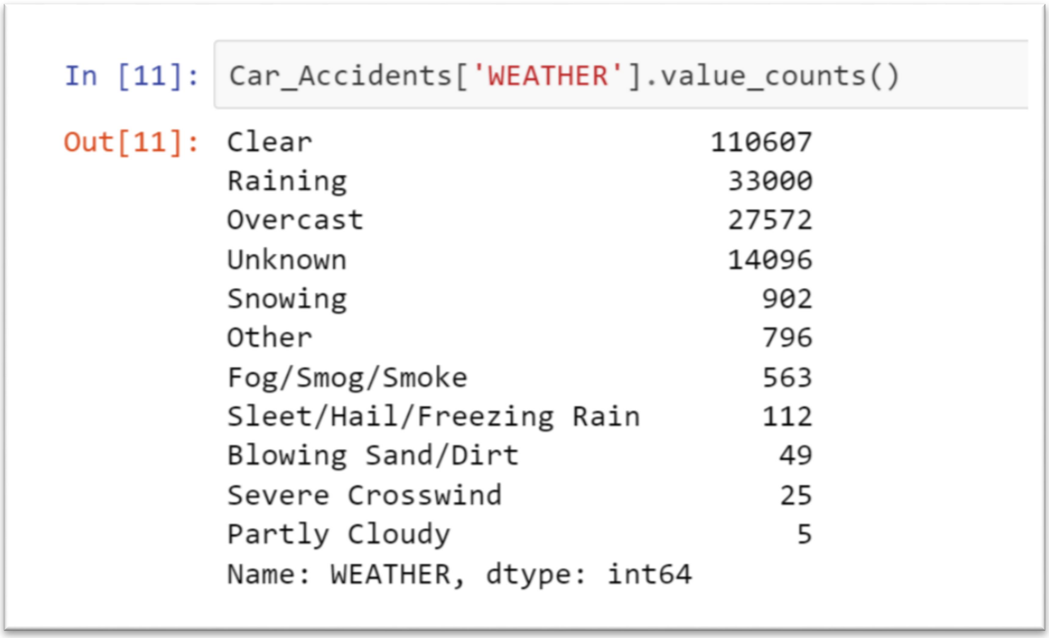
I utilized IBM cloud Jupyter Notebook to conduct that analysis and imported all the necessary Python libraries like Pandas, Numpy, Matplotlib, and Seaborn. The data was mostly categorical so I stuck to graphical representation to see correlation between various variables.

I started by importing the csv data file and to prepare the data, I dropped the columns we do not need from the dataset, i.e., columns that do not have values or where the values are unknown. Even though this is an important factor, I dropped Speeding entirely because it is missing over 180,000 values and this can hamper the results.

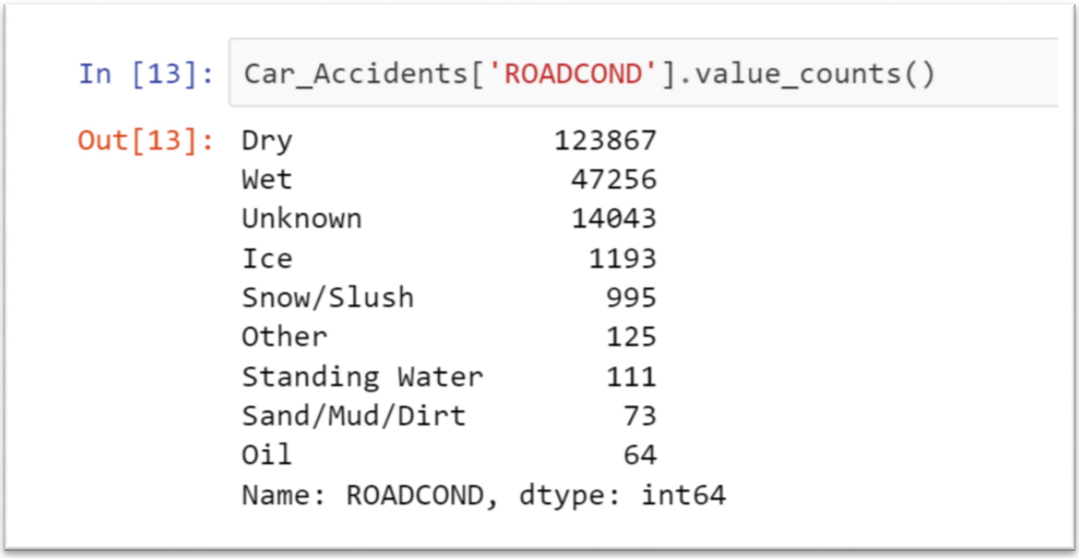


```
In [6]: Car_Accidents['SPEEDING'].value_counts()

Out[6]: Y    9333
        Name: SPEEDING, dtype: int64
```

Upon further inspection, I found out that ROADCOND and WEATHER have Unknown values. This will again hamper the analysis therefore I dropped these values where there is no information.



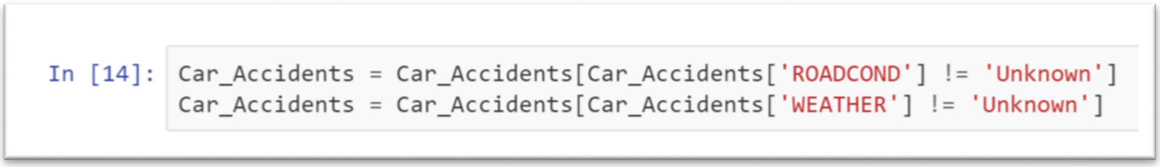
```
In [11]: Car_Accidents['WEATHER'].value_counts()

Out[11]: Clear                       110607
         Raining                      33000
         Overcast                     27572
         Unknown                      14096
         Snowing                        902
         Other                          796
         Fog/Smog/Smoke                 563
         Sleet/Hail/Freezing Rain       112
         Blowing Sand/Dirt               49
         Severe Crosswind                25
         Partly Cloudy                    5
         Name: WEATHER, dtype: int64
```



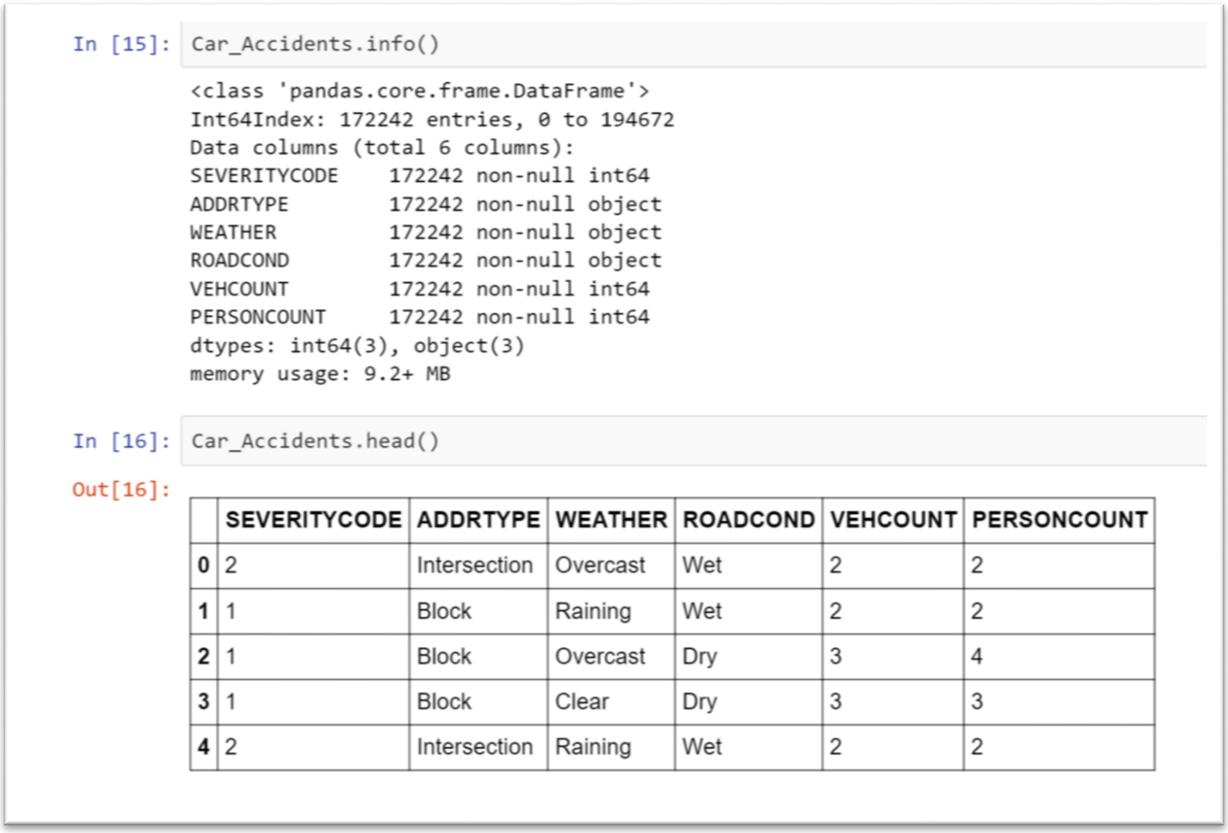
```
In [13]: Car_Accidents['ROADCOND'].value_counts()

Out[13]: Dry               123867
         Wet                47256
         Unknown            14043
         Ice                 1193
         Snow/Slush           995
         Other                125
         Standing Water       111
         Sand/Mud/Dirt         73
         Oil                   64
         Name: ROADCOND, dtype: int64
```

We notice some unknown data type in WEATHER and ROADCOND columns, so there is still some cleaning of the data to do.

In [14]:



```
In [14]: Car_Accidents = Car_Accidents[Car_Accidents['ROADCOND'] != 'Unknown']
         Car_Accidents = Car_Accidents[Car_Accidents['WEATHER'] != 'Unknown']
```



```
In [15]: Car_Accidents.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 172242 entries, 0 to 194672
Data columns (total 6 columns):
SEVERITYCODE    172242 non-null int64
ADDRTYPE        172242 non-null object
WEATHER         172242 non-null object
ROADCOND        172242 non-null object
VEHCOUNT        172242 non-null int64
PERSONCOUNT     172242 non-null int64
dtypes: int64(3), object(3)
memory usage: 9.2+ MB
```

```
In [16]: Car_Accidents.head()
```

Out[16]:

	SEVERITYCODE	ADDRTYPE	WEATHER	ROADCOND	VEHCOUNT	PERSONCOUNT
0	2	Intersection	Overcast	Wet	2	2
1	1	Block	Raining	Wet	2	2
2	1	Block	Overcast	Dry	3	4
3	1	Block	Clear	Dry	3	3
4	2	Intersection	Raining	Wet	2	2

Now the data is clean and ready to be analysed.

Criteria for selecting best model

In model classification, the difference between the actual data points and the best fit line produced by the algorithm is the model's error. In this project the accuracy and similarity score will be used to evaluate the model. In classification, accuracy classification score is a function that computes subset accuracy. This function is equal to the jaccard similarity score function. Essentially, it calculates how closely the actual labels and predicted labels are matched in the test set.

Materials and Methodology

- A typical model development workflow used in this study is as follows;
- Define the aims and objectives for model.
- Analyse current system.
- Data analysis and pre-processing.
- Develop a model for training and testing.
- Validate model against data not seen by the model.
- Deploy model.

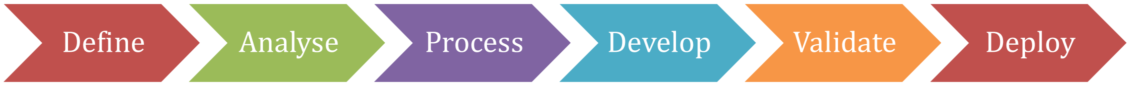


Figure 1 Sequence of a typical machine learning workflow

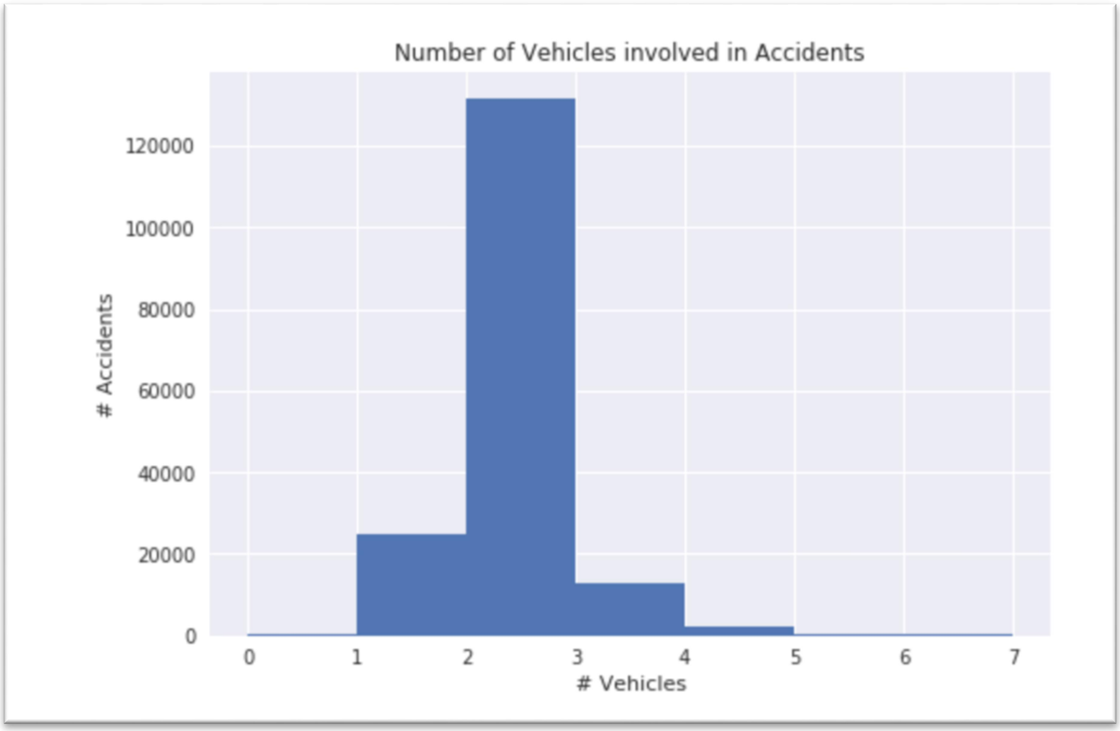
Data provided by Coursera for this capstone was used for this project. The data has lots of variables but only a few were selected for the model. The input variables used were dependent on their correlation to the target label which is the severity code. The severity code is classified as accident type: Property damage with severity code of 1 and Injury with severity code of 2

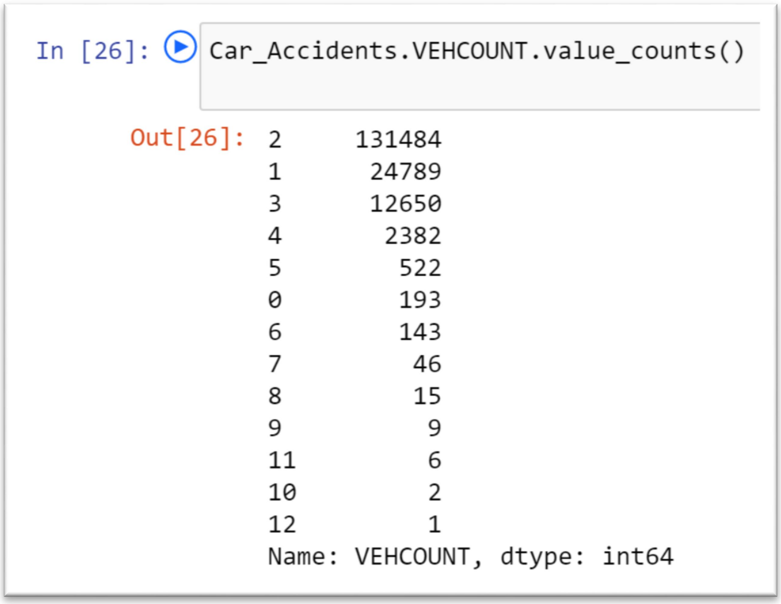
The input variables used are collision type, vehicle count, pedestrian count, weather, road condition, light condition, speeding and junction type. Table 1 gives a brief recap of the statistical description of some of the input data (data type = integer) used to train the models.

Data analysis and preprocessing

A series of data cleaning techniques were employed to change data types and identify missing data. Plots and correlations from the seaborn library were used to analyse the data. From the plots and charts below, features that have a high impact on the severity code were identified.

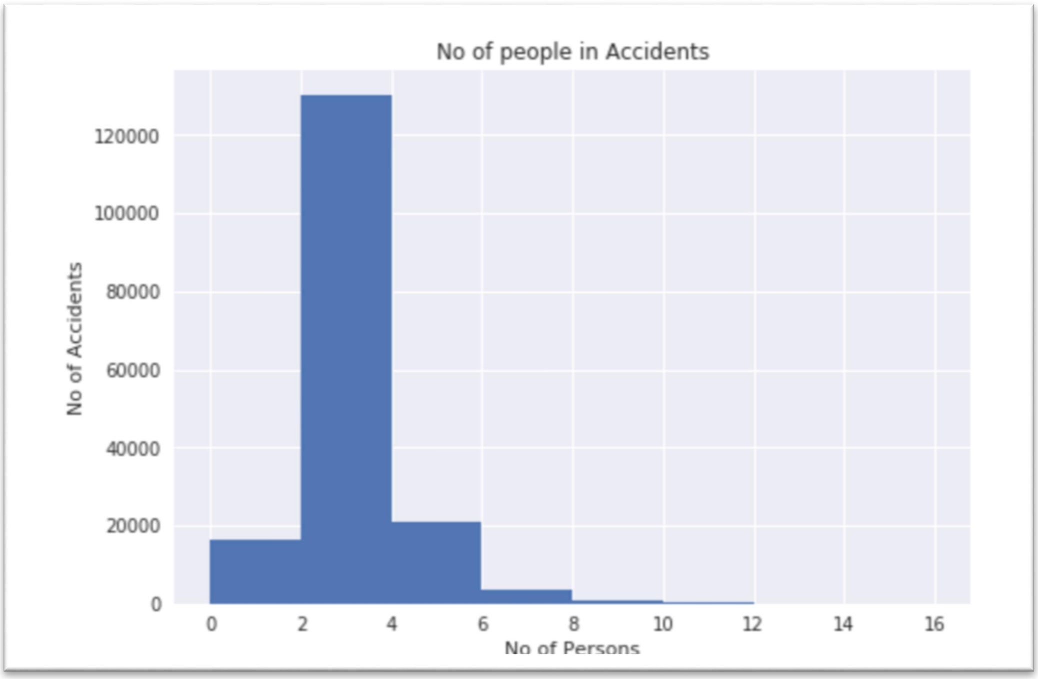
Analysis: Number of vehicles involved in Accidents

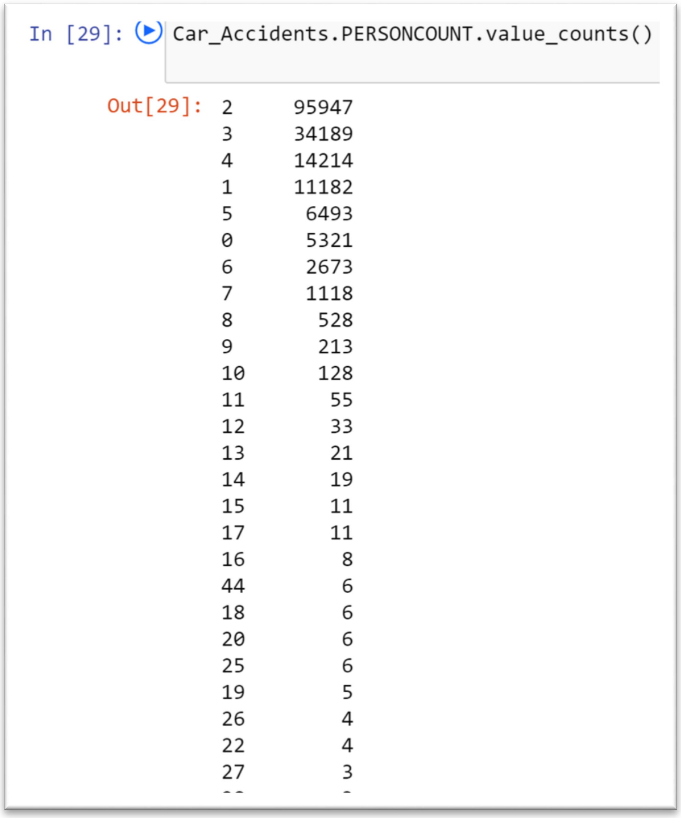




```
In [26]:   Car_Accidents.VEHCOUNT.value_counts()

Out[26]: 2      131484
         1       24789
         3       12650
         4        2382
         5         522
         0         193
         6         143
         7          46
         8          15
         9           9
         11          6
         10          2
         12          1
         Name: VEHCOUNT, dtype: int64
```

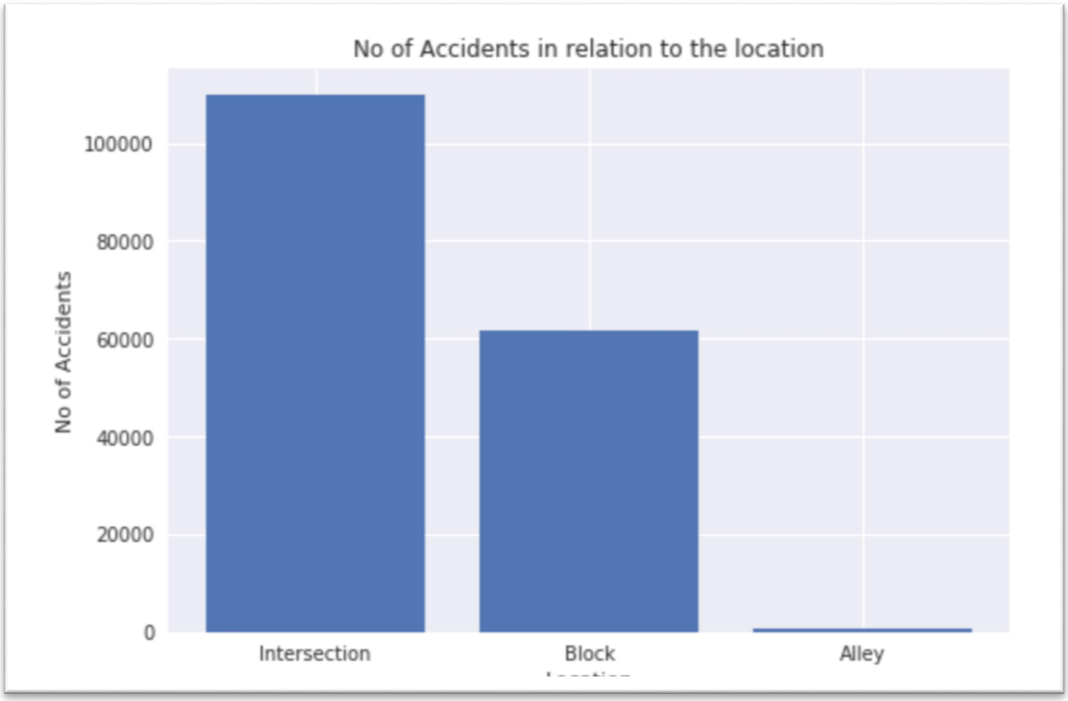
Analysis: Number of people involved in Accidents



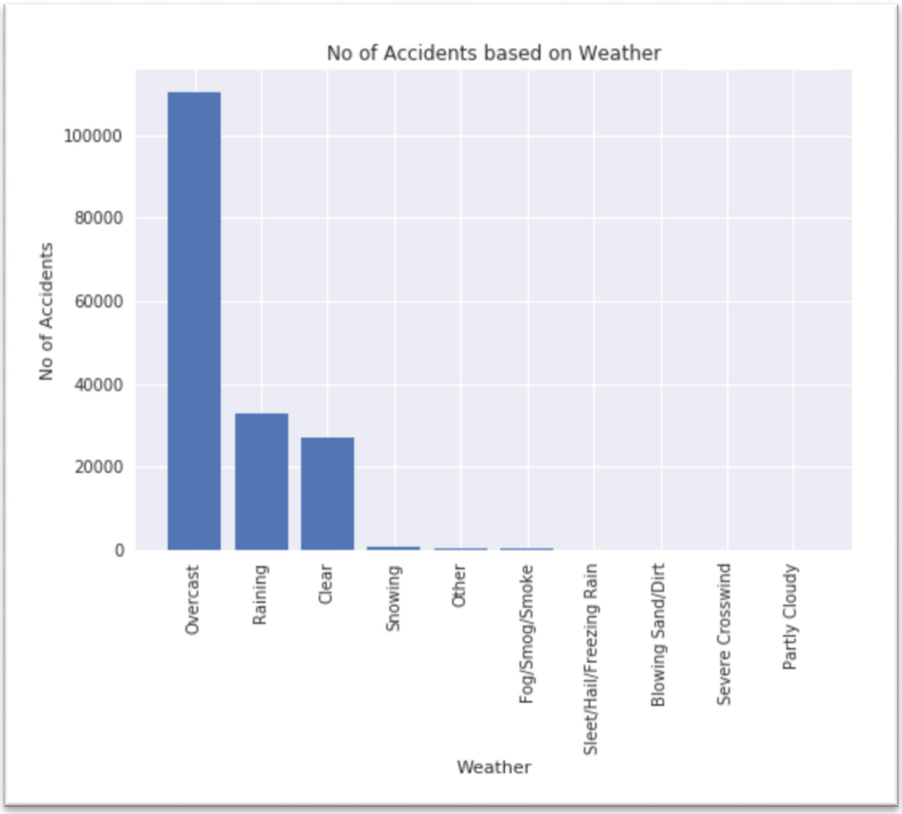


```
In [29]: Car_Accidents.PERSONCOUNT.value_counts()

Out[29]: 2     95947
         3     34189
         4     14214
         1     11182
         5      6493
         0      5321
         6      2673
         7      1118
         8       528
         9       213
         10      128
         11       55
         12       33
         13       21
         14       19
         15       11
         17       11
         16        8
         44        6
         18        6
         20        6
         25        6
         19        5
         26        4
         22        4
         27        3
```

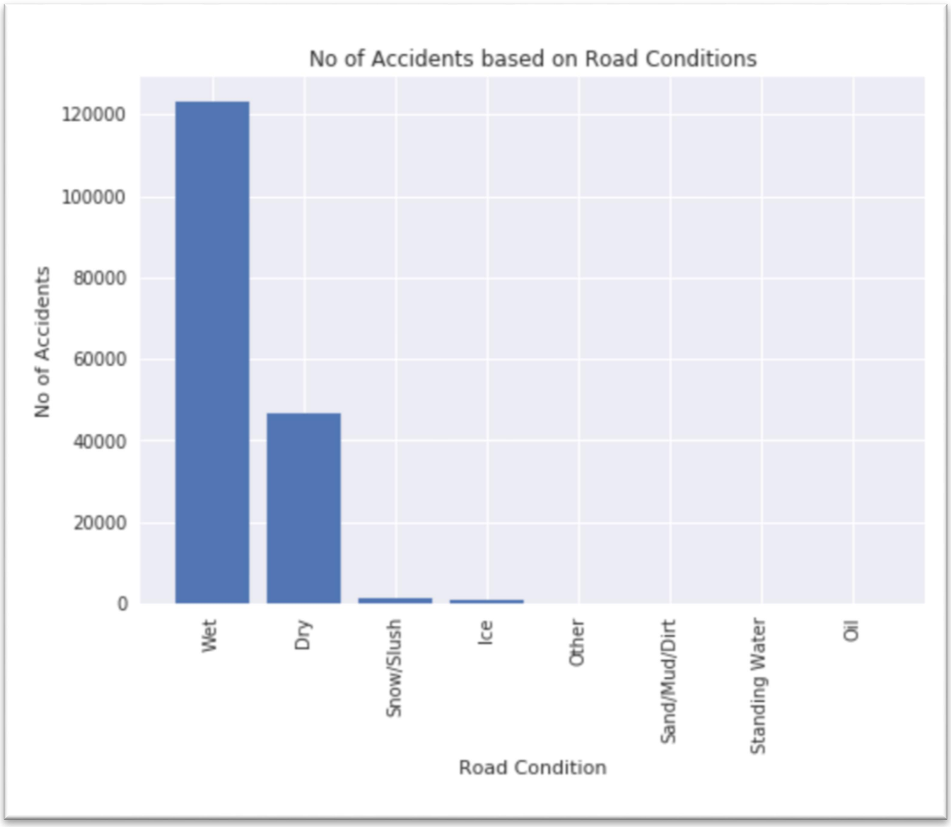
Analysis: Number of accidents in relation to the location



Analysis: Number of Accidents vs Weather conditions

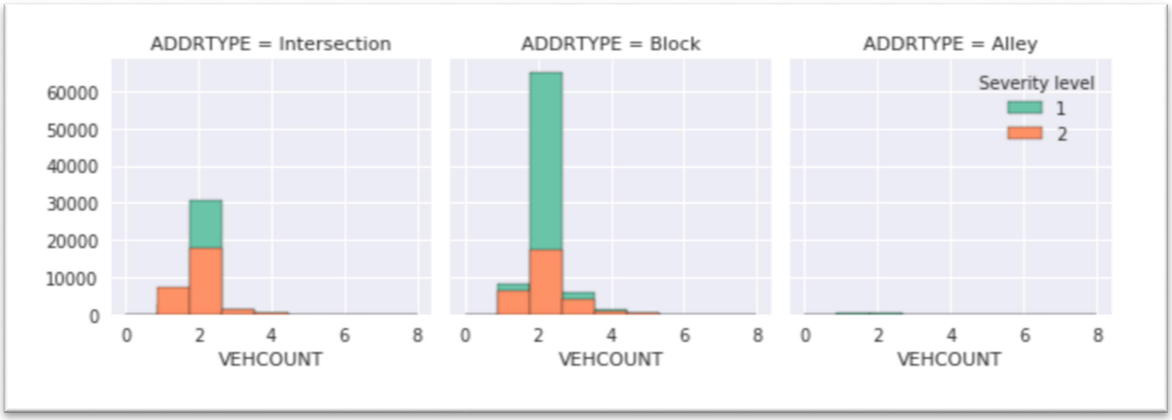


Analysis: Number of Accidents vs Road conditions

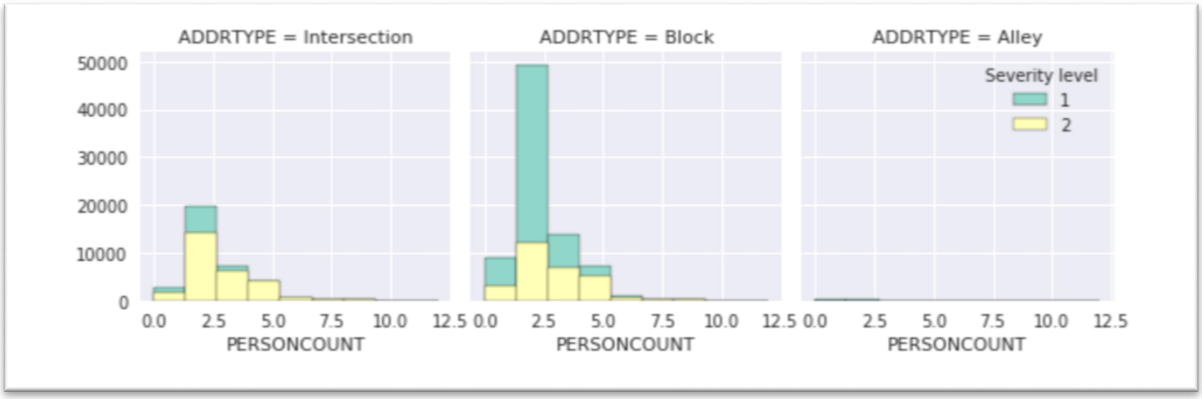


Analysis: Number of Accidents vs Severity vs counts

Next, I moved on to understand the severity of accidents based on our chose variables. I noticed the severity of accidents is higher (level 2 — injury) on an intersection whereas most non-severe accidents (level 1 — property damage) occur on blocks. I also found that most severe accidents occur at intersections and involve 2-3 people.



Analysis: Number of Accidents vs Road conditions



Analysis: K Nearest Neighbor, Decision Tree & Logistic Regression Model

Model	**Accuracy**	
K Nearest Neighbor	0.7372543983562347	
Decision Tree	0.7530756388853217	
Logistic Regression Model	0.7552844484397072	

Discussion

At the start of our analysis, I was trying to figure out the severity and frequency of road accidents based on weather conditions, road conditions, and other factors. Even though our data was a good size, there were a number of missing elements and we needed to clean the data in order to get a good result. We had to drop 'SPEED' because there were too many missing elements butI think that is an importantfactor that should be considered. From the analysis, it is clear that most accidents involve solo drivers, on wet roads, bad weather, at intersections, and are minorin nature. This could be helpful to the police department in understanding where to install more stop signs, or maybe adding camerasto intersections to compel people to slow down. Wealso live in a technologically friendly world so maybe we can develop some in built technology in our cars that warn us when the road and weather conditions are bad, or the car is approaching a stopsign.

Conclusion

I have used three different machine learning methods to predict the car accident severity, namely k nearest neighbour, decision tree and logistics regression. Based on the accuracy evaluations for each of these methods, logistic regression appears to be the method with highest accuracy. Therefore I would recommend the use of logistic regression to predict car accident severity in this scenario.

There are some data issues which may need to be addressed, for improved accuracy of the model. Firstly, there are many blank fields in the dataset, which suggest that the dataset may be incomplete. Further actions may need to be taken to ensure data integrtiy. Also, the sample size is relatively small. The accuracy of the model may be compromised if there is bias in the dataset. Conclusion Based on the data provided, we can develop a logistic regression model to predict car accident severity in Seattle, for the consumption of insurance companies. However, some data issues may need to be addressed to ensure the robustness of the model.

Summary and Conclusion

The project helps identify factors that contribute significantly to road accidents using a data-driven decision-making process. It can be concluded that poor road, weather and lighting conditions lead to most accidents whiles road users at intersections also tend to contribute to road accidents. Speeding may have had an inconclusive determinant on road accidents but thi is due to missing data in the dataset used. Data such as alcohol tests was not part of the dataset and would have been interesting to see its effect on road accidents. Based on this dataset the KNN established itself as a better machine learning model in road accident classification.