IBM DATA SCIENCE CAPSTONE PROJECT

DATA ANALYSIS OF CAR CRASHES OCCURRED IN SEATTLE BETWEEN THE YEARS 2004 AND 2020

OCTOBER 5, 2020

DHAVAL CHAURASIA

Table of Contents

1.Introduction

With rising trend of motorization, road safety has become a major concern. According to World Health Organisation, approximately 1.35 million people die each year as a result of road traffic crashes. Not only it costs human lives, but it costs taxpayers in the form of damaged property, post accident healthcare, disability allowance etc. Road traffic crashes cost most countries 3% of their gross domestic product.

Extensive studies have been conducted to find the cause and circumstances under which the accident occurred. For example, Rautela and Sharma (2004) did an analysis on accidents in the Himalayan terrain of Uttaranchal in India using the Fatality Index (FI) and suggested mitigation strategies for accidents on hill roads. Similarly, Johnson et al (2004) studied the rate of usage of cell phones during day and night conditions along-with other demographic characteristics. They found that the most frequent distraction to a driver was the usage of cell phone. In this project we will try to find factors or combination of factor which could cause an accident.

1.1 Business Understanding

The objective of this project is to develop a machine learning model to predict the severity of an accident depending on various factors. There are many potential causes of an accident. Some of the accidents are caused by the people who defy the laws by speeding, drunk driving, distracted driving etc. which can be prevented by incorporating harsher penalties. On the other hand, there are some uncontrollable circumstances such as road conditions, visibility, weather conditions etc. We will try to find certain patterns or correlations between attributes to predict the severity of an accident.

The target audience will be the government, the emergency response services, the police and the driver. Utilizing the model created in this task, the government can plan and prepare themselves better to manage any crisis. They can likewise alert the public, the emergency services and the police to be cautious in case any circumstance that can cause an accident, emerges.

2.Data

2.1 Data source

The data was collected from *https://data-seattlecitygis.opendata.arcgis.com*. The data came with a PDF file which contains the description about the attributes. There are 37 attributes and 194673 rows of data.

2.2 Data Processing

The downloaded data had lots of missing values, duplicate columns, redundant attributes and in some columns, data was not standardized and normalized. The process of data cleaning is discussed in the following steps.

- I started with dropping the duplicate columns. Our target variable had two columns namely "SEVERITYCODE.1" and "SEVERITYCODE". Hence, I decided to drop one of them.
- Upon examining further, I found some attributes that were redundant. For example, “INCKEY” which describes a unique key for the incident, was of no use. Similarly, the “LOCATION” attributes were of no use since we already have the coordinates of the accident. A total of 12 such attributes were dropped.
- The columns “EXCEPTRSNCODE” and “EXCEPTRSNDESC” were entirely empty and the metadata had no description about them, so I went ahead and dropped them from the data frame.
- Next, I used Isnull() function to detect missing data. Missing data from the columns 'ADDRTYPE’, ‘ST_COLDESC', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'LIGHTCOND', 'WEATHER', 'ROADCOND', 'X', and 'Y' were dropped since it accounted for mere ~5% of the total data. After this step, there were 180067 rows of data and 24 attributes left for analysis.
- There were few columns were data was not in same format. For example, "UNDERINFL" had four different values such as 1, Y, 0 and N. Y and N were replaced with 1 and 0 respectively.
- Attribute "INATTENTIONIND" and “SPEEDING” had only one type of response i.e. Y. Others were left empty. It was assumed N for others.
- “INCDTTM” is in datetime64[ns] format which can’t be used for analysis. Therefore, data was extracted from it and few more attributes were created such as “Hour”, “Day”, “Day of the Week”,” Year”, “Weekend” and “Month”.
- Target variable “SEVERITYCODE” had data in 1 and 2 formats. It was converted to 0 and 1 format.
- Data imbalance was corrected which is discussed in section 4 when we start building our model.

2.3 Feature selection

2.3.1 Target Variable

Our target variable is ‘SEVERITYCODE’ and it has two type of values.

0	Property Damage
1	Injury

2.3.2 Attribute selection

For the selection of attributes, I ran **Pearson's chi-square test of association.** This test is used when we have categorical data for two independent variables, and we want to see if there is any relationship between the variables. Upon running the test, following attributes were found to have some relationship with severity of an accident. In the next section we will carry out some exploratory data analysis and try to understand the relationship.

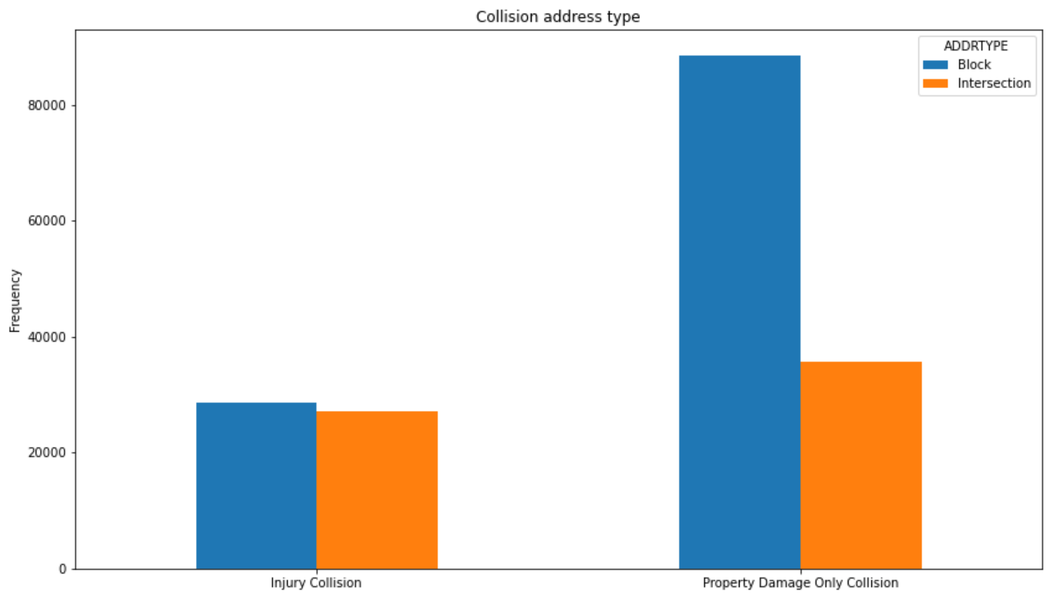
Chi square test

...

```
ADDRTYPE is IMPORTANT for Prediction
JUNCTIONTYPE is IMPORTANT for Prediction
LIGHTCOND is IMPORTANT for Prediction
WEATHER is IMPORTANT for Prediction
ROADCOND is IMPORTANT for Prediction
COLLISIONTYPE is IMPORTANT for Prediction
SDOT_COLDESC is IMPORTANT for Prediction
INATTENTIONIND is IMPORTANT for Prediction
UNDERINFL is IMPORTANT for Prediction
PEDROWNOTGRNT is IMPORTANT for Prediction
SPEEDING is IMPORTANT for Prediction
ST_COLDESC is IMPORTANT for Prediction
HITPARKEDCAR is IMPORTANT for Prediction
Day is NOT an important predictor. (Discard Day from model)
Month is IMPORTANT for Prediction
Year is IMPORTANT for Prediction
day_of_week is IMPORTANT for Prediction
weekend is IMPORTANT for Prediction
hours is IMPORTANT for Prediction
```

3.Exploratory Data Analysis

3.1 Relationship between collision address type and severity of an accident



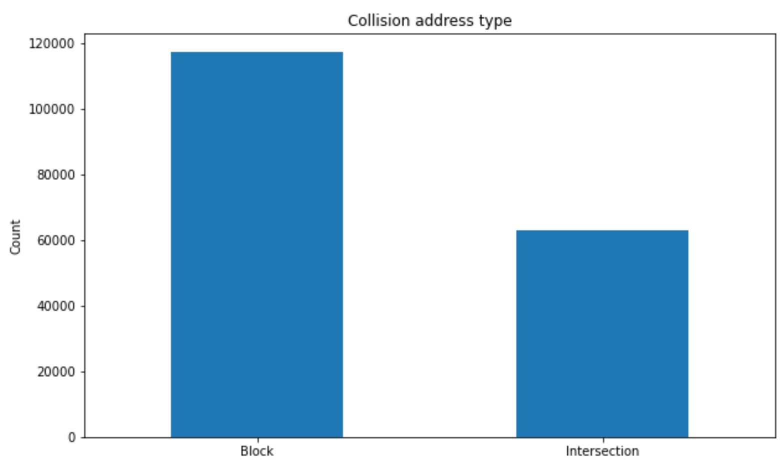


Figure1a) Relationship between address type and severity type. 1b) Bar plot of address type and accident count

As we can see in figure 1b, majority of accidents takes place around the blocks rather than at intersection. Close to 120K accidents happened at or around the block compared to about 60k at intersection. In figure 1a, we can see the relationship between severity of an accident and address of collision. A lot of accident (>80K) that resulted in property damage collision, happened around the block where number of accidents that resulted in injury collision were same for block and intersection.

3.2 Relationship between junction type and severity of an accident

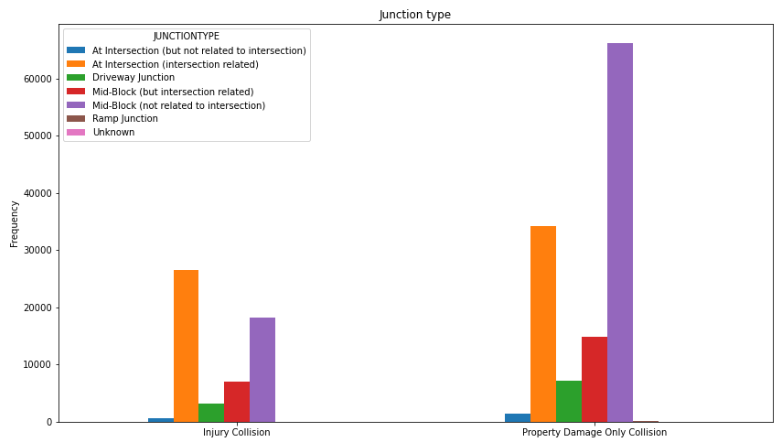


Figure 2 Junction type and severity type

Figure 2 shows a bar graph plot of frequency of car accidents based on junction type and grouped by severity type. Majority of accident (>60K) which resulted in property damage, happened at Mid-Block (not related to intersection) whereas most of the accidents (>25K) that caused injuries happen at intersection (intersection related). Accidents at intersection (intersection related) and Mid-Block (not related to intersection) resulted in a combined total of more than 140K accidents.

3.3 Relationship between collision type and severity of an accident

Figure 3 shows a plot which describes the type of collision and count of accident. Head o collision seems to be rare. Majority of accident (>40K) involved a parked car followed by angle and rear ended.

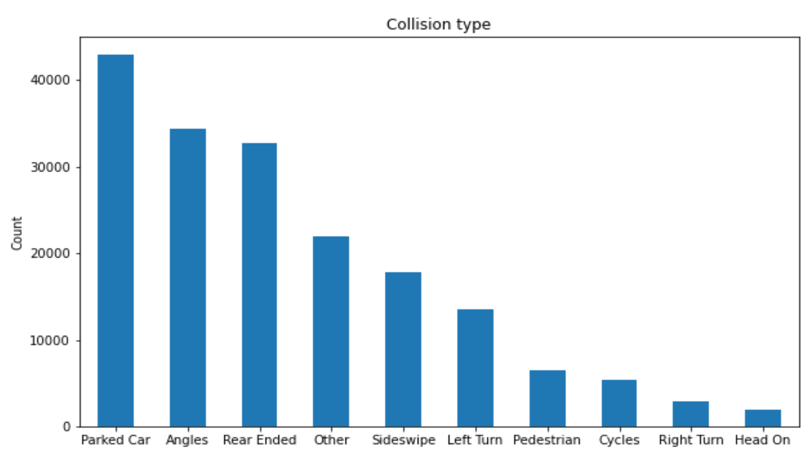


Figure 3 Type of collision and accident count

In Figure 4 shows 40K accidents that involved a parked car, resulted in property damage while less than 5K resulted in injury. Rear ended and sideswipe caused more injury than other type of collision. Apart from parked car, angles, sideswipe and rear ended caused significant property damage.

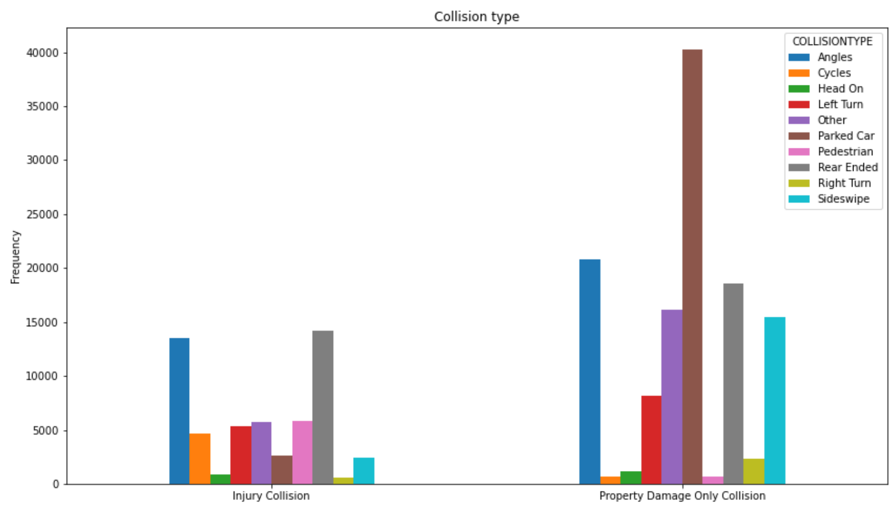


Figure 4 Collision type and severity od an accident

3.4 Relationship between weather and severity of an accident

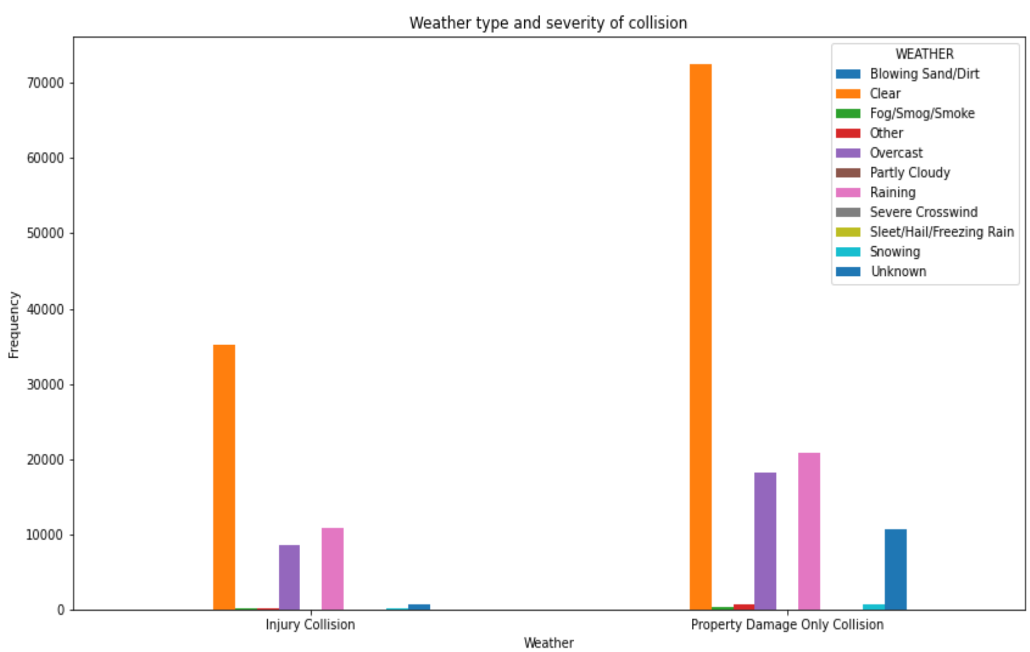


Figure 5 Weather type and severity of an accident

Most of the accident happened in clear weather condition. Close ~110k accident happened in dry conditions, of which more than 70k resulted in property damage and more than ~35k resulted in injury. Apart from this, raining and overcast conditions stood second in terms of accident count. Combined they reached an accident count of ~20k which resulted in injury and a count of ~40k which resulted in property damage.

3.5 Relationship between road condition and severity of an accident

Figure 6 shows us the relationship between accident count and road condition grouped by severity of an accident. Around 120k accident happened in dry condition of which 80k resulted in property damage and 40k resulted in injury. These was followed by wet road conditions where ~30k resulted in property damage and ~20k resulted in injury.

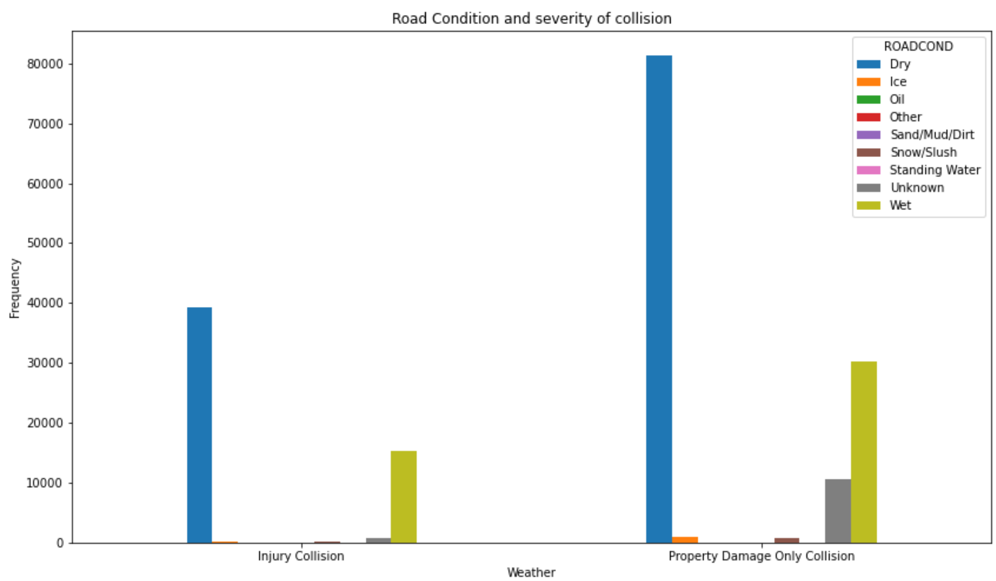


Figure 6 Road condition and severity of accident

3.6 Relationship between light condition and severity of an accident

Figure 7 shows us the relationship between accident count and light condition grouped by severity of an accident. Majority of (~110k) accident happened in a day light conditions. Close to 50k accident happened in Dark (streetlight on) conditions of which around 15k resulted in injury and 35k resulted in property damage.

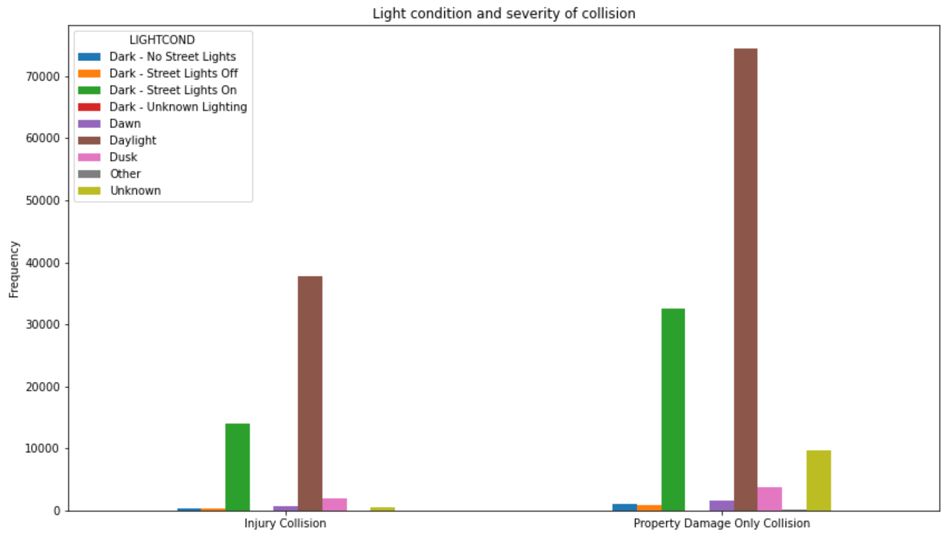


Figure 7 Light condition and severity of accident

3.6 Relationship between day, time and severity of an accident

Figure 8 shows us that most of the accident happened at mid night (>25k). This was followed by accident occurred during 14hrs to 17hrs. This is where office hours end for most of the people. Close to 45k accidents occurred during this time. Accidents were relatively on lower side throughout the night until 6hrs in the morning. The number of accidents then keeps on increasing as the day progresses. As expected, weekdays account for a greater number of accidents (figure 9).

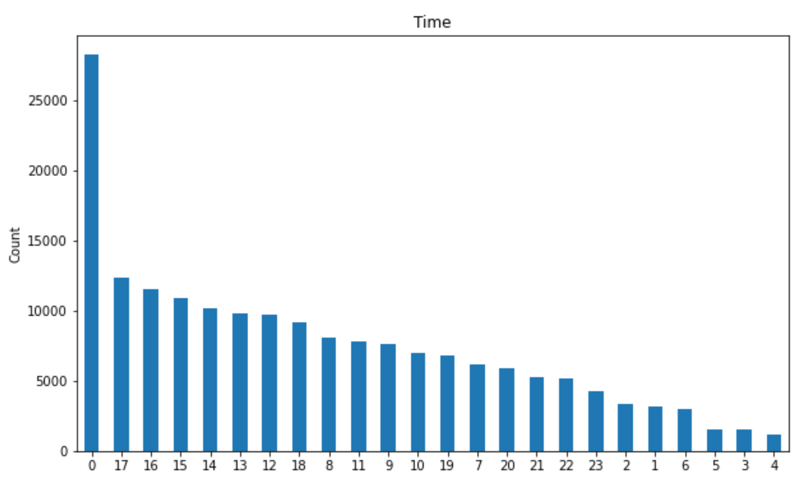


Figure 8 Time of accident and accident count

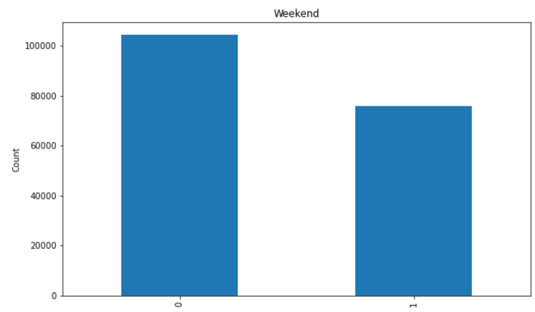


Figure 9 Accident count on weekdays and weekends

3.7 Scatter plot of latitude and longitude of the accident

Figure 10 shows a scatter plot of the longitude and latitude of the accidents that occurred in the city of Seattle. The dense blue area in the centre is downtown Seattle. Since downtown of the city mainly comprises of the offices, it was expected that most of the accidents occurred here. I tried generating the heat map of city using Folium, but because of the size of the data, it failed. This plot shows the relationship between accident count and location.

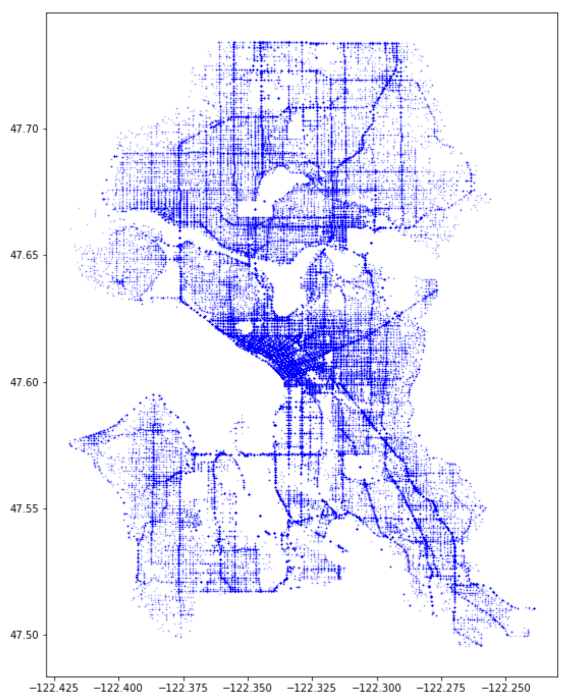


Figure 10 Scatter plot of latitude and longitude

4. Machine learning Model

Before we start to build our machine learning model, we must check for data imbalance. It can be seen from the picture below that our target variable is highly imbalanced. This may result in overfitting of data by model to the class which is represented more which is property damage in our case. There are various ways to deal with imbalanced data such as oversampling minority data or under sampling majority data.

```
[62]: df_ml.SEVERITYCODE.value_counts()

[62]: 1    124258
      2     55809
      Name: SEVERITYCODE, dtype: int64
```

For our problem I have decided to up sample minority data. This is done as shown below.

```
[43]: # upsample minority
      pos_upsampled = resample(positive,
       replace=True, # sample with replacement
       n_samples=len(negative), # match number in majority class
       random_state=3) # reproducible results

[44]: upsampled = pd.concat([negative, pos_upsampled])

[45]: upsampled.SEVERITYCODE.value_counts()

[45]: 2    93239
      1    93239
      Name: SEVERITYCODE, dtype: int64
```

Now that we have balanced dataset, we can build a machine learning model which will help us predict the severity of an accident. Since it is a classification-based problem, I have decided to use Decision tree and Logistic regression machine learning algorithms.

4.1 Decision tree

I imported the decision tree classifier from 'sklearn' package. After building the model I ran a couple of tests to know the accuracy of the model. For decision tree, I ran Accuracy score, f1 score, Jaccard similarity score, recall and precision score. Model achieved an accuracy score and Jaccard similarity score of 0.73. Recall, Precision and f1 score is shown below.

	precision	recall	f1-score	support
0	0.80	0.81	0.80	31019
1	0.57	0.56	0.56	13998
micro avg	0.73	0.73	0.73	45017
macro avg	0.68	0.68	0.68	45017
weighted avg	0.73	0.73	0.73	45017

4.2 Logistic regression

I imported the logistic regression classifier from 'sklearn' package. After building the model I ran a couple of tests to know the accuracy of the model. For logistic regression I ran log loss, f1 score, Jaccard similarity score, recall and precision score. I have also tried changing the solver in order to achieve better score. I have only included the best result achieved here. Model achieved a Jaccard score of 0.68 and log loss of 0.58. Recall, Precision and f1 score is shown below.

	precision	recall	f1-score	support
0	0.81	0.71	0.76	31019
1	0.49	0.63	0.55	13998
micro avg	0.68	0.68	0.68	45017
macro avg	0.65	0.67	0.65	45017
weighted avg	0.71	0.68	0.69	45017

4.3 ROC

To evaluate the performance of the models, I decided to plot ROC (Receiver Operating Characteristic Curve). It is the plot between the True Positive Rate (y-axis) and False Positive Rate (x-axis). It is a performance measurement for classification problem at various threshold value. AUC (Area under curve) represents the degree of separation. An excellent model has an AUC close to 1 which means it has good measure of separability.

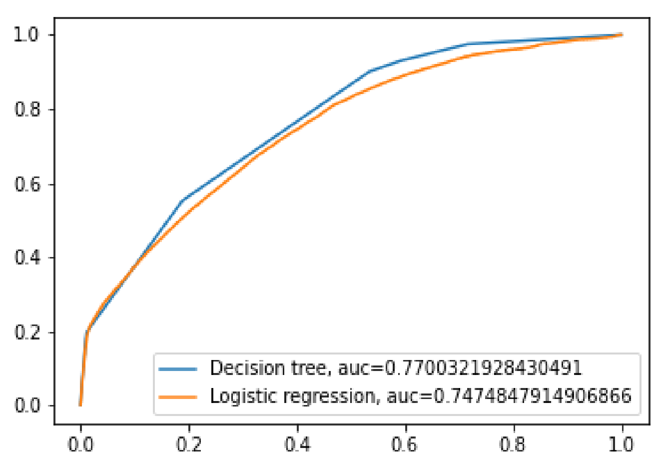


Figure 11 ROC and AUC

At the lowest point, i.e. at (0, 0) the threshold is set at 1.0. This means our model classifies all accidents result in property damage. At the highest point i.e. at (1, 1), the threshold is set at 0.0. This means model classifies all accidents result in injury. The rest of the curve is the values of FPR and TPR for the threshold values between 0 and 1. At some threshold value, we observe that for FPR close to 0, we are achieving a TPR of close to 1. This is when the model will predict whether accident will result in property damage or an injury.

4.4 Results

Based on the evaluation I found that the performance of decision tree model with AUC as 0.77 and higher Jaccard score, Precision score, Recall score and F1 score was better than that of logistic regression model which had the AUC of 0.74. Even though Decision tree had slightly better performance, Logistic regression can also be used side by side for prediction since it too had a decent performance.

5. Observation

In this project, I analyzed the relationship between severity of an accident and various circumstances which resulted in that accident. There are various factors that may cause an accident to occur. In my study I found that Junction type, place where collision occurred, type of collision, time of collision etc. also plays an important role other than the obvious choice of weather, light condition and road condition. For example, based on coordinates of various accidents, I found that density of the accidents occurred in downtown Seattle were way more than other areas. Similarly, I found that accident that involved a parked car are more likely to result in property damage. I also found that more accidents occurred during mid-night when compared to daytime.

6. Conclusion

I was successfully able to build a model that predicts the severity of accident to an accuracy of 73 percent. I believe the accuracy can be significantly improved by adding few more attributes such as gender of driver, age of driver, engine capacity, type of vehicle, number of casualties etc. With the help of this model, improving road infrastructure and awareness programs about road safety and rules, Government can reduce the number of cases that result in loss of life and property.