Predicting Seattle Road Accident Severity

Hemalatha Murugesan

October 12, 2020

1. Introduction

Road accidents have become very common nowadays. As more and people are buying automobiles, the incidences of road accidents are just increasing day by day.

As per WHO observatory data "Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally and are predicted to become the seventh leading cause of death by 2030".

Analyzing an important array of factors, including weather conditions, speed, construction work, special events, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These perceptions, could allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accident can occur as well as saving both, time and money. In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year. Other attentive private companies could be investing in technologies aiming to improve road safeness.

Stakeholders:

- Public Development Authority of Seattle
- Car Drivers

2. Data

2.1 Data Source

The data comes from Seattle car accident set. A comprehensive dataset of 194,673 accidents occurring between 2004-2020. The dataset has 38 columns describing the details of each accident including the weather conditions, collision type, date/time of accident and location.

2.2 Pre-Processing

The following features are selected for the prediction

SEVERITYCODE	A code that corresponds to the severity of the collision
INATTENTIONIND	Whether or not accident was due to inattention (Y/N)
UNDERINFL	Whether or not driver was in drug or alcohol (Y/N)
WEATHER	Weather condition
ROADCOND	Condition of the road
LIGHTCOND	Condition of the light
SPEEDING	Whether or not Speeding was a factor (Y/N)
ADDRTYPE	Area of the accident happened
COLLISIONTYPE	Type of Collision
PERSONCOUNT	The total number of people involved in the collision
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state
JUNCTIONTYPE	Category of junction at which collision took place
INCDTTM	Date and Time
Weekday	Day of the week

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).

Featured data are preprocessed to convert into continuous data. Encode accident was due to attention with 0 as No and 1 as Yes. Road conditions can be segregate into dry, mushy and wet. Dry as 0, Mushy as 1 and Wet as 2. Ice, Standing Water and oil are like wet then it encodes as 2. Snow/Slush and Sand/Mud/Dirt are encoded as 1. There are few data with 'Others', So it encodes as 'Unknown'. Weather conditions are encoded like clear as 0, Overcast and Cloudy as 1, Windy as 2 and Rain as 3. Light conditions are encoded like Light as 0, Medium as 1 and Dark as 2. Encoding under the influence No as 0 and Yes as 1.

There are 10 junction type stated in the data set. Parked Car, Angles, Rear Ended, Sideswipe, Left Turn Pedestrian, Cycles, Right Turn, Head On and Other. These junction types are converted definite values.

2.3 Cleaning

Data cleaning is the process of detecting and removing the inaccurate values from the data set. In the data frame there were many values had Null or Unknown which is outlier. To keep unknown values with replacing those values.

Here to remove Null values or Other values from the data set. The entire process of cleansing data to reduce almost 5000 records which is having redundant data.

3. Exploratory Data Analysis

Visualize the data analysis makes it easier to understand data. Relationships and patterns can be fetched out as trends. First, we plot the number of accidents happened in period like yearly, monthly, days of the week and daily as well.

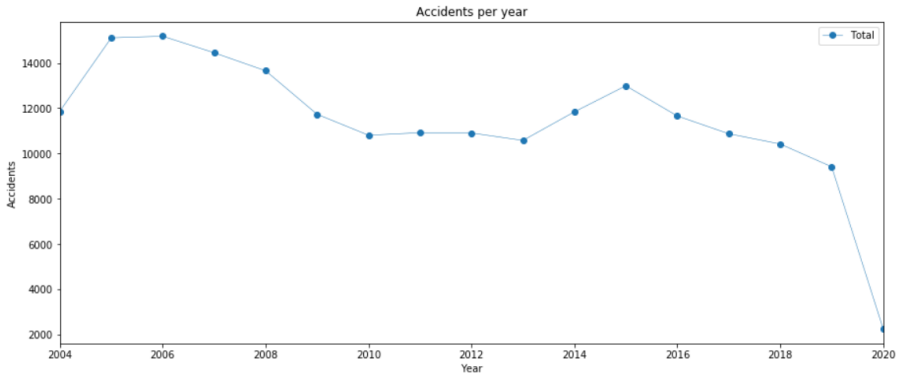

Fig1: Total Number of Accidents per Year

Figure 1 shows that the number of accidents were decreased from the years 2005 to 2019. Except 2015 was increased little bit and again reduced.

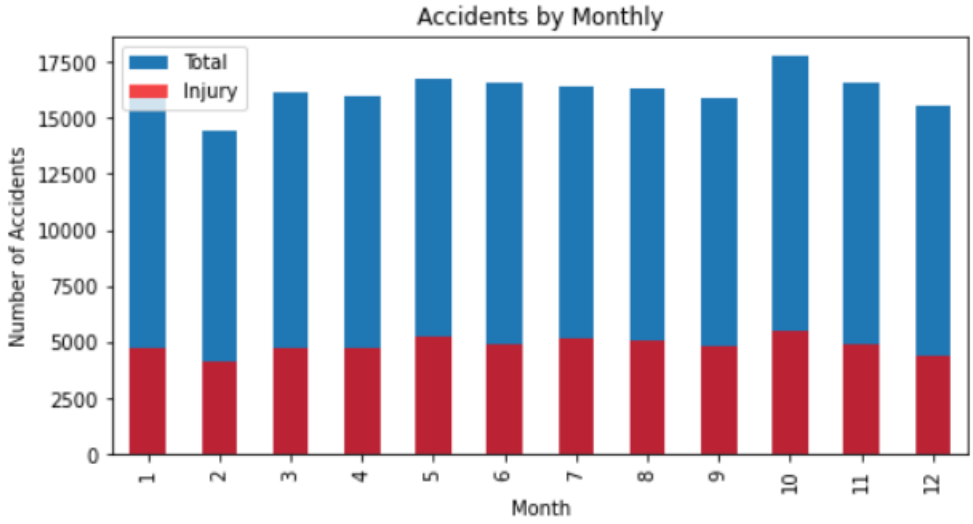

Fig: 2 Number of Accidents by Monthly

Figure 2 shows that the number of accidents were high in October and low in February. From march to September the accident counts were stable

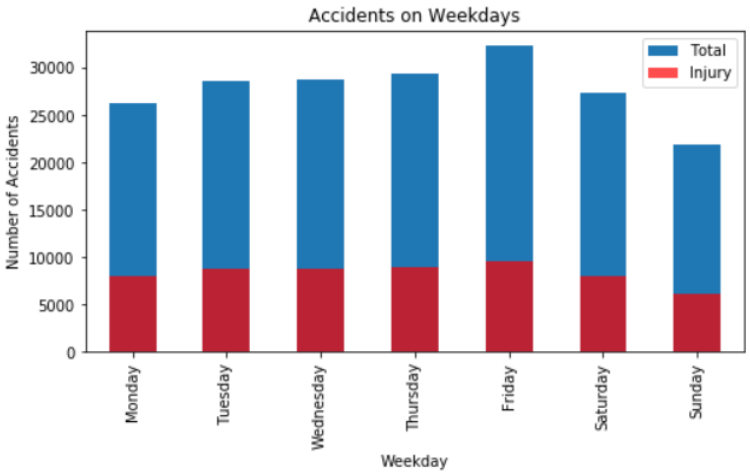

Fig 3: Number of Accidents on Weekdays

As per Figure 3 there is no significant difference between them in the Injury/Collision accidents. But the solid trend with more accidents on Friday and lesser number of accidents on Sunday of all other days.

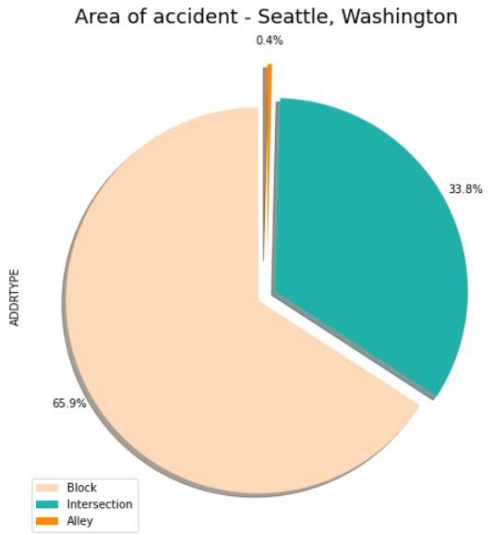

Fig 4: Percentage of accidents in areas

Figure 4 shows the most accident history recorded in the Block area also very less accidents in the Alley area

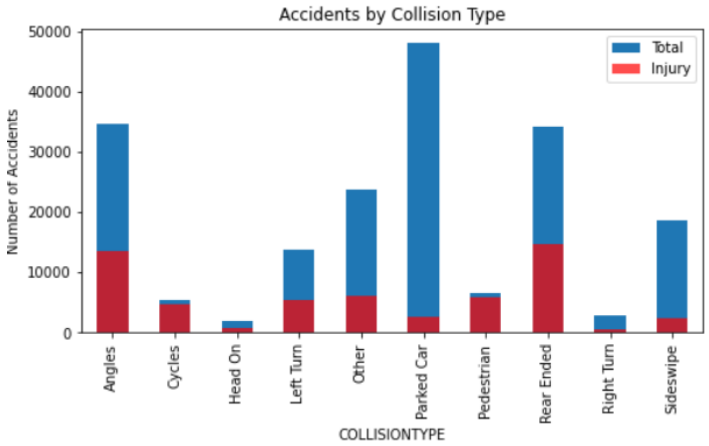

Fig 5: Number of Accidents by Collision type

There are 10 types of collision included in the data set. In figure 5 “Parked Car “ has most number of accidents and injury is only 12%. “Cycles” collision type have more than 90% accidents were injuried.

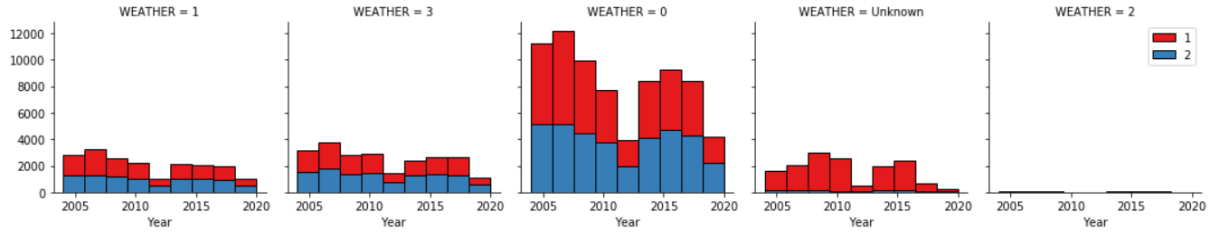

Fig 6: Number of Accidents based on the Weather condition

Figure 6 shows that number of accidents in each year by the weather condition. 0 = Clear, 1 = Overcast and Cloudy, 2 = Windy, 3 = Rain and Snow. Most of the accidents happen clear weather.

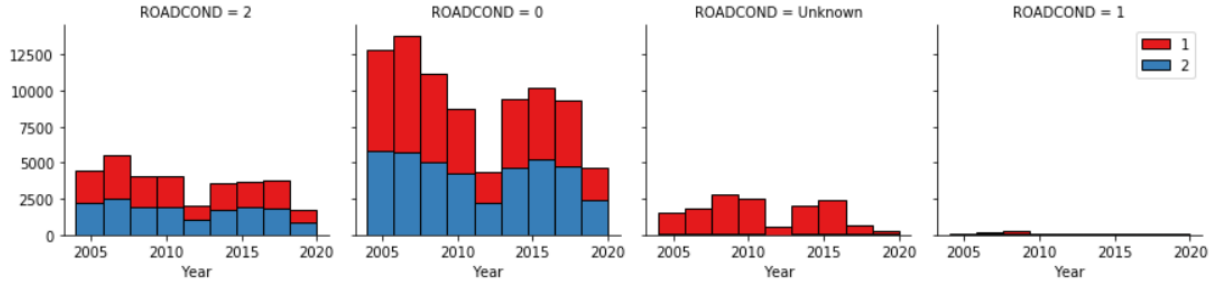

Fig 7: Number of Accidents based on the Road condition

Figure 7 shows that number of accidents in each year by the road condition. Road Conditions 0=Dry, 1=Mushy, 2=Wet. Most of the accidents happens at dry road and few accidents with Snow or Slush or Sand or Mud or Dirt.

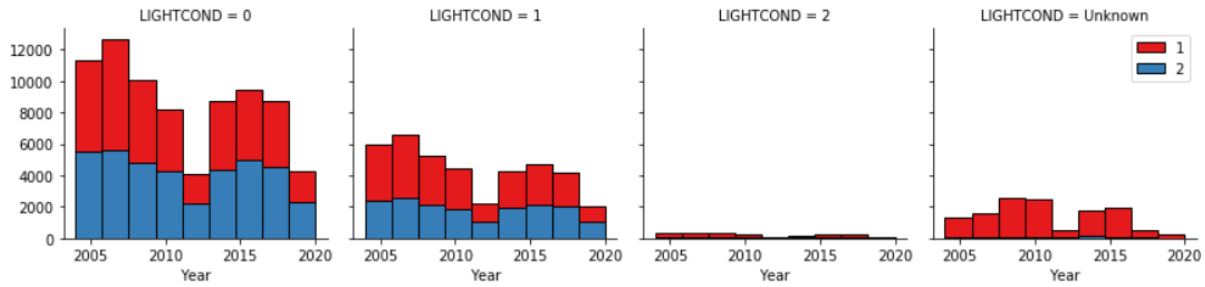

Fig 8: Number of Accidents based on the Lighting

Figure 8 shows that number of accidents in each year by the lighting. Light Conditions 0 = Light, 1=Medium, 2 = Dark. As per figure greatest accidents in bright light condition only.

As per the seasonality data hard to predict accident severity with only seasonality predictors.

4. Predictive Modeling

Different classification algorithms been tuned for prediction of the level of accident severity. These algorithms provide a supervised learning approach to predicting with certain accuracy. Accuracy of an algorithm compared to determine the best suited algorithm for this problem.

From the dataset 80/20 ratio data were split as training and test sets. There are five different kind of approaches were used in here. Similar kind of operation was performed in each model. With train and test data sets the best hyperparameters were

4.1 Logistic Regression

The linear model Logistic Regression classifier from sklearn was used to model the data. For create this model use C=0.001 and solver used as 'liblinear'. The confusion matrix was plotted for the regression model which is showed in figure 9. And the accuracy of the model is 72%.

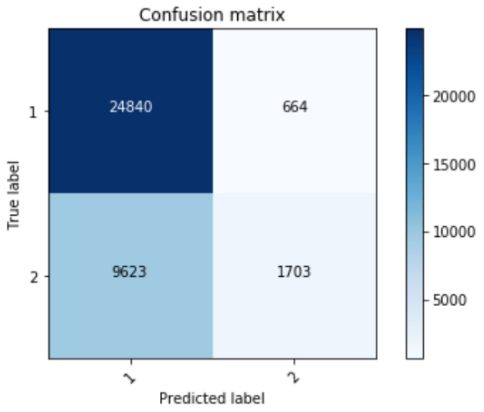

Fig 9: Logistic Regression Confusion Matrix

4.2 Decision Tree

The Decision Tree Classifier from sklearn were used to model the data. Tree depth is find based on the accuracy of the model is each depth until depth 15. Figure 10 shows the max depth as 9 which is having high accuracy. Used “entropy” criterion and tree depth as 9.

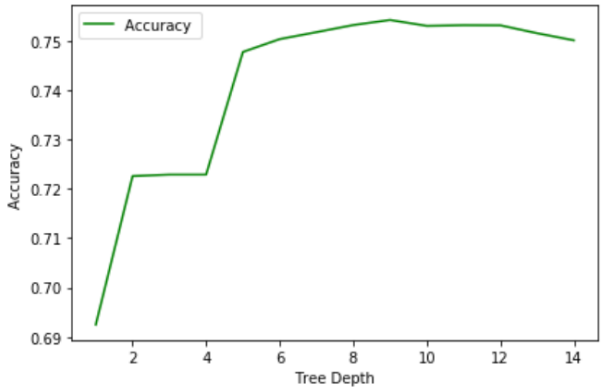

Fig 10: Accuracy of Decision Tree by increasing the max depth value

The confusion matrix was plotted for the Decision Tree model which is showed in figure 11. And the accuracy of the model is 75%.

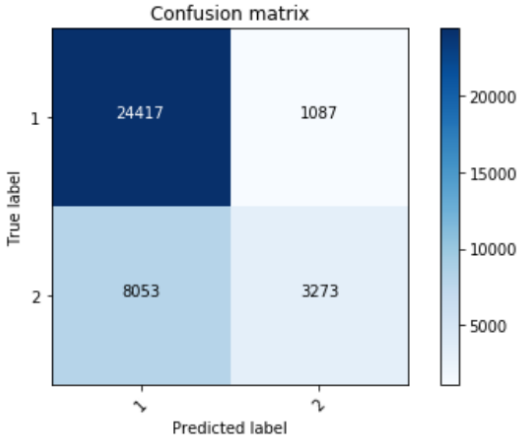

Fig 11: Decision Tree Confusion Matrix

4.3 Support Vector Machine (SVM)

The Support Vector Machine Classifier from sklearn were used to model the data. Used 'rbf' kernel to create the model. The confusion matrix was plotted for the SVM model which is showed in figure 12. And the accuracy of the model is 75%.

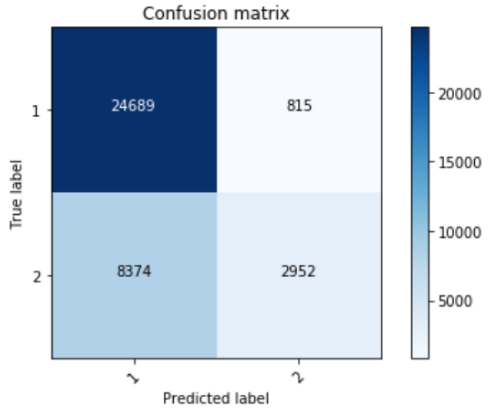

Fig 12: SVM Confusion Matrix

4.4 K Nearest Neighbor (KNN)

The K Nearest Neighbor classifier from sklearn were used to model the data. The best K is find based on the accuracy of the model by each K. Figure 13 shows the best K as 6. The Accuracy of this model is 73%.

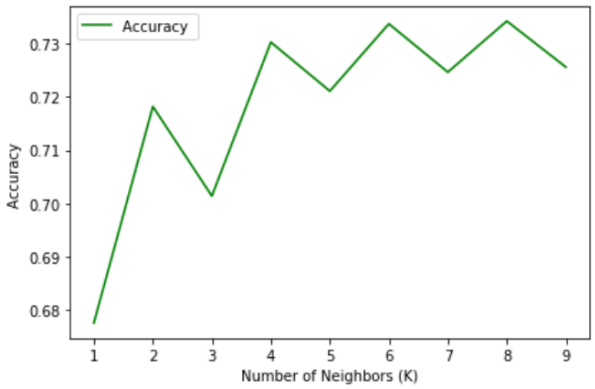

Fig 13: Accuracy of K by increase K value

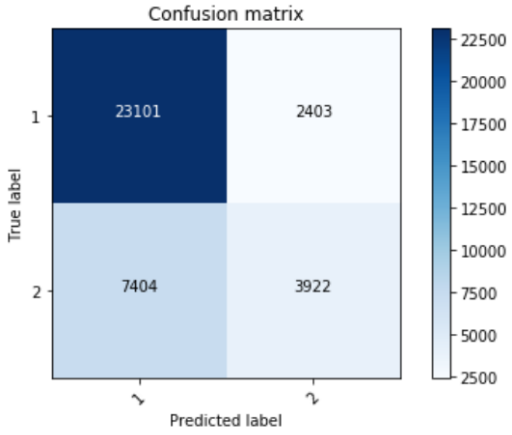

Fig 14: KNN Confusion Matrix

4.5 Gaussian Naïve Bayes

The 1.1 Gaussian Naïve Bayes classifier from sklearn were used to model the data. The confusion matrix was plotted for the Naïve Bayes model which is shown in figure 15. The accuracy of the model is 70%.

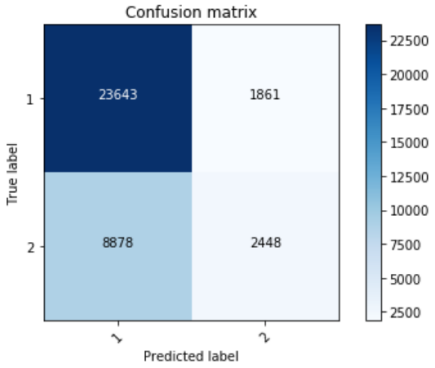

Fig 15: Naïve Bayes Confusion Matrix

5. Results

The metrics used to compare the accuracy of the models are the Jaccard Score, F1-score, Precision and Recall. This table reports the results of the evaluation of each model.

Algorithm	Test Accuracy	Jaccard	F1-score	Precision	Recall
Logistic Regression	0.72	0.72	0.65	0.72	0.72
Decision Tree	0.75	0.75	0.71	0.75	0.75
SVM	0.75	0.75	0.70	0.76	0.75
KNN	0.73	0.73	0.71	0.72	0.73
Gaussian Naïve Bayes	0.71	0.66	0.66	0.68	0.71

Table 1: Accuracy Metrics

As per the table Decision Tree and SVM have similar accuracy, however the computational time from decision tree is better than SVM. SVM take lot of time to compute the model. Also, the ROC curve shows the decision tree have higher true positive rate.

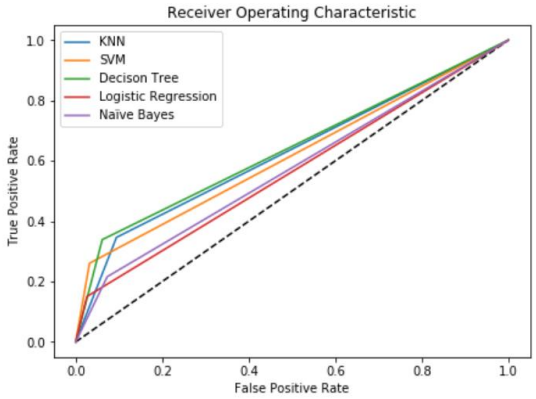

Fig 16: ROC Curve

6. Conclusion

In this project, the prediction of traffic accident severity using classification was explored. Several predictors such as collision type, address type, weather, light, road conditions, alcohol or drug influence, pedestrian, date and time to predict the severity of an accident. Different machine learning models were used to predict or classify the accident severity like Logistic Regression, Decision Tree, KNN, SVM and Naïve Bayes. The Decision Tree model performed the best with an accuracy of 75%. These models can be useful in determining the outcomes of accidents for traffic safety institutions. Also, by identifying the features that favor most of the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.

7. Observation

Although the models able to achieve well, averaging a 73% accuracy. However, there was still significant variance that could not be predicted by the models. More features like speed or uninterrupted time of travelling could be used to predict a more accurate classification.

The known problem that the severity of this classification problem was simplified to two different classes, Injury or Property Damage.

The models performance could also generally be improved by tuning the parameters with the models ang investigating with different training and test sizes. The next step on this problem could be the accident prediction model not only predicting the severity also the location where accidents can occur in advance.

8. Reference

https://www.who.int/gho/road_safety/mortality/number_text/en/

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf