Applied Capstone Project
Predicting Car Accident Severity

Nabeel Sharaf

1. Introduction

1.1 Business Problem

Motorization have enhanced the lives of many individuals and societies, but the benefits have come with a price. As the number of people on roads keep increasing, the injuries and fatalities associated with it kept also increasing due to the unfortunate events happening on the road. In recent years, there have been an increasing amount of research attention given to understand the severity of the injuries caused by road accidents.

This problem requires a scientific approach to be dealt with. As more accurate and comprehensive accident data are recorded, it gives the researches a better view in analysing these incidents and come up with a solution.

Commuters will need the help of data scientist to find a solution to understand various causes of injuries happening in the roads. This way they can be prepared and decrease their chances of meeting with an accident.

1.2 The Solution

From the vast amount of data available about these incidents, this project tries to predict the severity of an accident by taking into account various dimensions like weather condition, time of the week. It can help a motorist plan his journey for a safer commute.

2. Data

The Data used in this project is SDOT Traffic Management Division, Traffic Records Group. It has the information on all types of collisions provided by SPD and recorded by Traffic Records from 2004-2020, updated weekly.

It contains a total of 37 attributes which will help in training the model and give a better prediction. Some of the entries in the data will be omitted to fine tune the model and get a better result.

Attribute Information

Attribute	Data type, length	Description
OBJECTID	ObjectID	ESRI unique identifier
SHAPE	Geometry	ESRI geometry field
INCKEY	Long	A unique key for the incident
COLDETKEY	Long	Secondary key for the incident
ADDRTYPE	Text, 12	Collision address type: • **Alley** • **Block** • **Intersection**
INTKEY	Double	Key that corresponds to the intersection associated with a collision
LOCATION	Text, 255	Description of the general location of the collision
EXCEPTRSNCODE	Text, 10	
EXCEPTRSNDESC	Text, 300	
SEVERITYCODE	Text, 100	A code that corresponds to the severity of the collision: • **3**—fatality • **2b**—serious injury • **2**—injury • **1**—prop damage • **0**—unknown
SEVERITYDESC	Text	A detailed description of the severity of the collision
COLLISIONTYPE	Text, 300	Collision type
PERSONCOUNT	Double	The total number of people involved in the collision
PEDCOUNT	Double	The number of pedestrians involved in the collision. This is entered by the state.
PEDCYLCOUNT	Double	The number of bicycles involved in the collision. This is entered by the state.
VEHCOUNT	Double	The number of vehicles involved in the collision. This is entered by the state.
INJURIES	Double	The number of total injuries in the collision. This is entered by the state.
SERIOUSINJURIES	Double	The number of serious injuries in the collision. This is entered by the state.
FATALITIES	Double	The number of fatalities in the collision. This is entered by the state.
INCDATE	Date	The date of the incident.
INCDTTM	Text, 30	The date and time of the incident.
JUNCTIONTYPE	Text, 300	Category of junction at which collision took place
SDOT_COLCODE	Text, 10	A code given to the collision by SDOT.
SDOT_COLDESC	Text, 300	A description of the collision corresponding to the collision code.
INATTENTIONIND	Text, 1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.

Attribute	Data type, length	Description
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision.
PEDROWNOTGRNT	Text, 1	Whether or not the pedestrian right of way was not granted. (Y/N)
SDOTCOLNUM	Text, 10	A number given to the collision by SDOT.
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)
ST_COLCODE	Text, 10	A code provided by the state that describes the collision. For more information about these codes, please see the State Collision Code Dictionary.
ST_COLDESC	Text, 300	A description that corresponds to the state's coding designation.
SEGLANEKEY	Long	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	Long	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Text, 1	Whether or not the collision involved hitting a parked car. (Y/N)

The attribute severity code will be the key attribute in the prediction

SEVERITYCODE: A code that corresponds to the severity of the Collision:

- 3—fatality
- 2b—serious injury
- 2—injury
- 1—prop damage
- 0—unknown

This project will be using the K-Nearest Neighbour and Decision Tree algorithms which is a type of unsupervised machine learning algorithm. The model will be evaluated using Jacob and F1- square methods.

3. Methodology

3.1 Data Analysis

In order to identify the correlation of each attributes with the severity code, the below code is used

```
Dataframe.groupby([‘attribute’])[‘SEVERITYCODE’].value_counts()
```

This will also help in eliminating columns which will hinder the accuracy of the machine learning algorithms.

3.2 Data Encoding

It is important that we encode the datasets before starting the modelling. Most machine learning models only work with numerical data.

Therefore, it is important that the other data types are converted to numerical data using the syntax:

Dataframe[‘column’].replace(to_replace=[‘A’,’B’,’C’],value=[0,1,2],inplace=True)

3.3 Feature Creation and Normalization

Now the feature set is prepared for the machine learning model to work on. Here the Feature set has 12 attributes

```
SEVERITYCODE       int64
ADDRTYPE         float64
JUNCTIONTYPE     float64
COLLISIONTYPE    float64
VEHCOUNT           int64
PEDCYLCOUNT        int64
PERSONCOUNT        int64
PEDCOUNT           int64
ROADCOND         float64
LIGHTCOND        float64
WEATHER          float64
HITPARKEDCAR       int64
dtype: object
```

All null values of the data are dropped to make the algorithm more precise.

3.4 Machine Learning Model

The data is split into Train (80 %) and Test (20%) for evaluation purposes. The data is then run on K Nearest Neighbour and Decision Tree Machine Learning Models.

X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=4)

4. Result

The data is applied with K Nearest Neighbour and Decision Tree Machine Learning Models. The Jaccard Score anf F1-score was calculated to determine the accuracy of each model. For this project both the models used delivered similar results

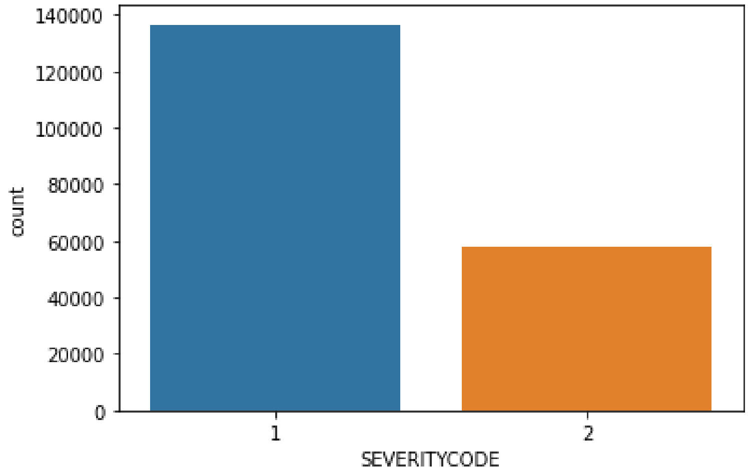
Machine Learning Algorithm	Jaccard Score	F1-Score
K Nearest Neighbour	0.74	0.71
Decision Tree	0.74	0.71

The only difference between these models used were their computing times. Decision Tree model delivered much faster results.

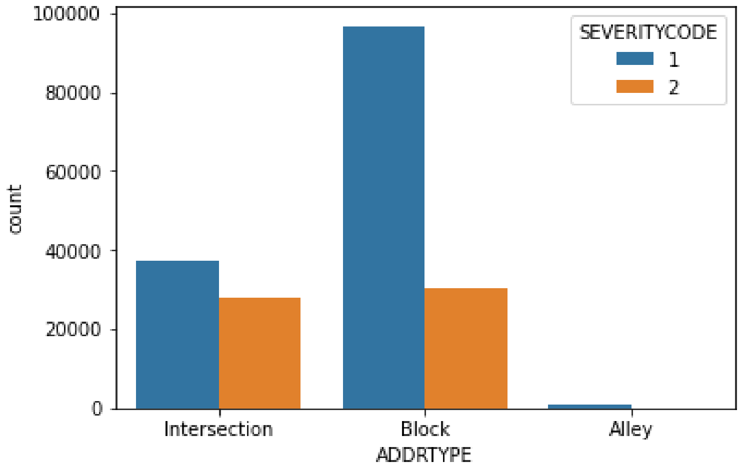
5. Discussion

In this project the accident severity were analysed and a model to predict the severity of the accident was created.

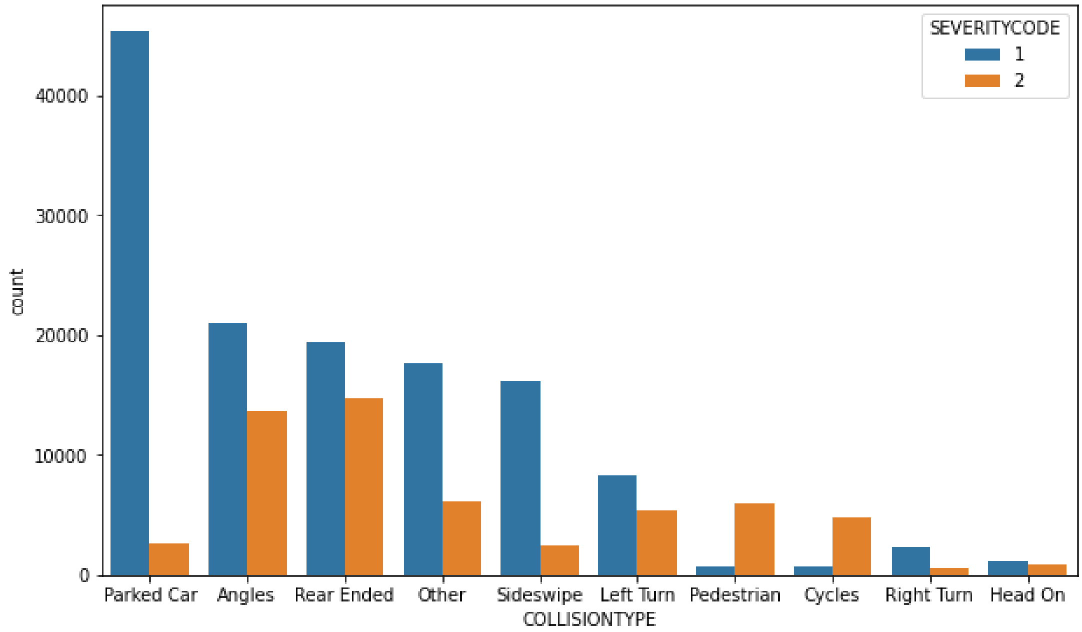
It is evident from the data that more accidents caused damage to the property than injury.



After analysing the data, it’s understood that accidents happen more at Block Addresses than intersections or alleys. So it is advised to be more attentive at the addresses.



The most common type of collision was parked car collisions for property damage collisions, but collisions at an angle caused the most number of injuries.



This project shows the advantages of using graphs and machine learning models to prevent or reduce road accident severity. A driver may look at these analytics to identify areas of possible accident causes and be prepared for such occurances.

The machine learning model may be used to predict the future outcomes and help the city management to create a plan to reduce the severity of such accidents. With more and more datas being available, machine learning model's accuracy will be improved and a safer traffic is assured.