Predicting Severity of Accident

Hammad Yaqoob

28/09/2020

1. Introduction

According to SDOT Traffic Management Division in 2019 there were over 9000 accidents in Seattle. 32% of those resulted in injuries. In 2020 there have been 2245 accidents. This can be avoided by using collected data to build a machine learning model which can predict whether a route is dangerous based on the current conditions.

The business problem is to analyse the collision dataset and build a model which can predict under the current condition if a route is dangerous or not based on past accidents

This model can be useful to various navigation systems and car manufacturers as a built in safety feature which could warn users of potential hazardous routes

2. Data Acquisition and Cleaning

2.1. Data

The data we will be using is provided by the SDOT Traffic Management Division and Traffic Records Group. It has various information of accidents that has occurred from 2004 to the present day.

The data has various information regarding incidents including:

- Severity of the incident
- Location of the incident
- Whether or not the accident was caused due to inattention or if drugs or alcohol was involved or if speeding was involved

- Weather at the time of the incident
- Road condition at time of incident
- Light condition at time of incident

By using the Severity of the incident as our target variable and the remainder as features we will build a classification model that will predict how dangerous a certain route is based on the given inputs

2.2. Cleaning

2.2.1. EXCEPTRSNCODE

When this value is set to NEI it states that there is not enough information regarding the accident hence I have removed rows as this will not add value in predicting severity of accidents

2.2.2. X and Y Coordinate

Location of the accident is vital in predicting the severity of the accident hence removed any entries where X and Y coordinates are not populated

2.2.3. UNDERINF, SPEEDING, INATTENTIONIND

The above values represent whether drugs, speeding or inattention were part of the accident. In all cases the data contained na values and Y were the above was true. The data was cleaned by replacing na with 0, Y with 1 and where N was populated it was replaced with 0. So the data for these columns only contained 0 and 1

2.2.4. JUNCTIONTYPE, ADDRTYPE, WEATHER, ROADCOND, LIGHTCOND

The above values represent the junction type, collision address type, weather conditions, road condition and light conditions regarding the accident. All of this is defined categorically, contained null values and also contained unknowns. The null and unknown values were removed and changed from categorical data to binary data

2.3. Feature Selection

From the data the following information was removed as it was codes assigned to the incident which would not be relevant on the severity of an accident:
'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO','STATUS', 'SDOT_COLCODE', 'SDOT_COLDESC', 'SDOTCOLNUM', 'ST_COLCODE', 'ST_COLDESC'

From the data the following information was removed as duplicate information that we can get from severity code:
'SEVERITYCODE.1', 'SEVERITYDESC'

From the data the following information was removed as duplicate information that we can get from Location information:
'EXCEPTRSNCODE','EXCEPTRSNDESC'

From the data the following information was removed as it was information pertaining to the accident after it happened which would not be helpful in predicting whether the route or conditions would cause a severe accident:
'COLLISIONTYPE','PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'PEDROWNOTGRNT', 'HITPARKEDCAR'

3. Exploratory Data Analysis

3.1. Relationship between Speeding and Severity of accident

Speeding can cause more severe accidents as the impact would be greater. Hence analysis was done to determine how many accidents involved speeding and what severity they were:

Number of accidents with speeding involved: 8681
Percentage of severity 2 accidents when speeding was involved: 38.2%

Number of accidents with speeding not involved: 176972
Percentage of severity 2 accidents when speeding was not involved: 30.2%

From the results we can see that we can conclude that the majority of accidents did not involve speeding and those that did only 38% resulted in severity 2 accidents

3.2. Relationship between Under Influence and Severity of accident

Drugs and alcohol can cause drivers' reflexes to not be as sharp as they would be under normal circumstances . Hence analysis was done to determine how many accidents involved being under the influence and what severity they were:

Number of accidents which involved being under the influence: 8855
Percentage of severity 2 accidents which involved being under the influence: 39%

Number of accidents which did not involve being under the influence: 176798
Percentage of severity 2 accidents which did not involve being under the influence: 30%

From the results we can see that we can conclude that the majority of accidents did not involve being under the influence and those that did only 39% resulted in severity 2 accidents

3.3. Relationship between Inattention and Severity of accident

Inattention such as looking at phones can cause severe accidents. Hence analysis was done to determine how many accidents involved inattention and what severity they were:

Number of accidents which involved inattention: 28754
Percentage of severity 2 accidents which involved inattention: 35%

Number of accidents which did not involve inattention : 156899
Percentage of severity 2 accidents which did not involve inattention: 30%

From the results we can see that a significant proportion of accidents were caused by inattention with 35% of those result in severity 2 injuries

3.4. Weather conditions

Analysis has been done on the weather conditions when accidents occurred. The table is below

	Weather Condition	Percentage of accidents
0	Clear	59.583664
1	Raining	17.543773
2	Overcast	14.871108
3	Unknown	6.706470
4	Snowing	0.484035
5	Other	0.404097
6	Fog/Smog/Smoke	0.302108
7	Sleet/Hail/Freezing Rain	0.061745
8	Blowing Sand/Dirt	0.027013
9	Severe Crosswind	0.013231
10	Partly Cloudy	0.002756

From the table we can conclude that the majority of accidents occurred when it was clear however a significant proportion occurred when it was raining or overcast

3.5. Road conditions

Analysis has been done on the weather conditions when accidents occurred. The table is below

	Road condition	Percentage of accidents
0	Dry	66.778170
1	Wet	25.236555
2	Unknown	6.622984
3	Ice	0.641467
4	Snow/Slush	0.539516
5	Other	0.063375
6	Standing Water	0.055109
7	Sand/Mud/Dirt	0.035270
8	Oil	0.027554

From the table we can see the road conditions in the majority of incidents were dry however a significant number of incidents occurred when the road was wet

3.6. Light conditions

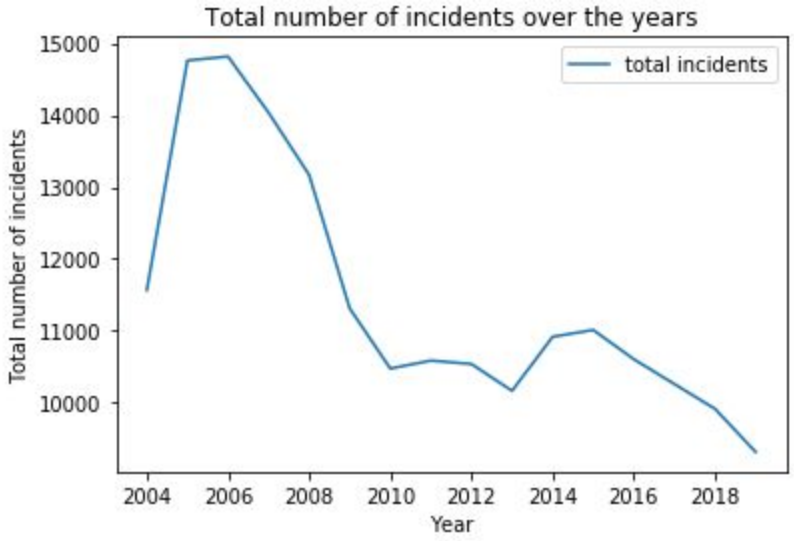
Analysis has been done on the light conditions when accidents occurred. The table is below

	Light condition	Percentage of accidents
0	Daylight	62.019501
1	Dark - Street Lights On	25.879375
2	Unknown	6.093579
3	Dusk	3.159021
4	Dawn	1.326921
5	Dark - No Street Lights	0.788101
6	Dark - Street Lights Off	0.629267
7	Other	0.098168
8	Dark - Unknown Lighting	0.006067

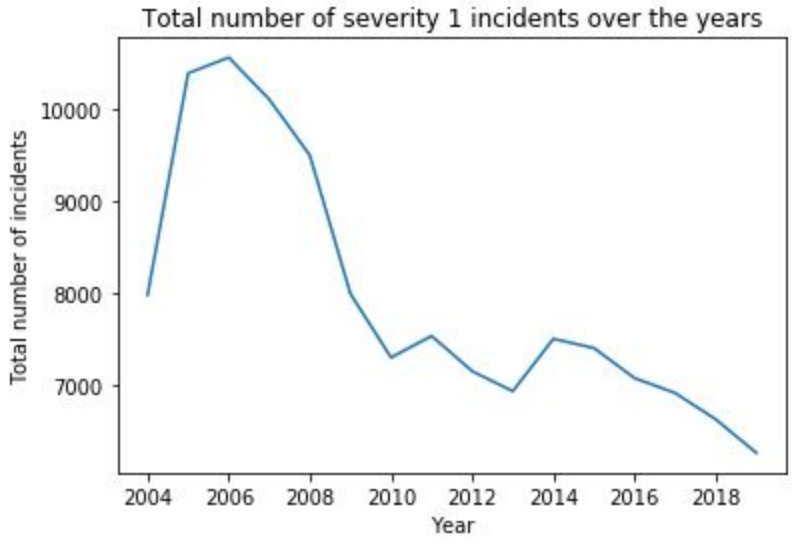
From the table we can see the light conditions in the majority of incidents were daylight however a significant number of incidents occurred it was dark with no street lights

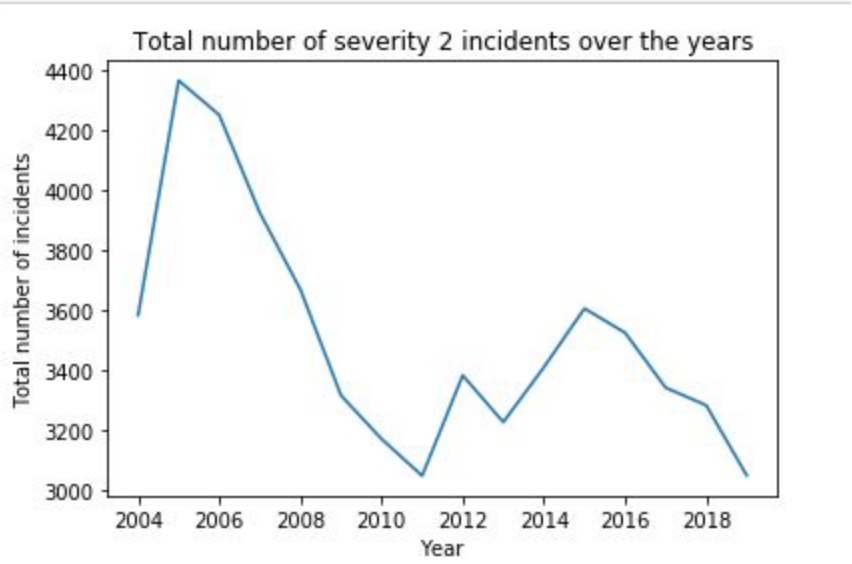
3.7. Incidents per year

Since the data contains incidents from 2004, the total number of incidents were analysed per year and the results are shown in the below graph



Also the total number of Severity 1 and 2incidents were mapped over the years. The graphs are below:





From the graphs we can see that the peak number of accidents occurred in 2006 hoever from that point there has been a decline in the number of incidents with 2019 showing the lowest numbers yet

4. Machine Learning

4.1. Machine Learning Models Utilised

For this exercise the following machine learning algorithms were used to predict the severity of an accident:

1. KNN: This is a classification algorithm which uses supervised learning. When training the model the data points and their classification are plotted. When testing the new data is plotted and depending on the nearest k points the new point is classified. This is ideal as the model needs to classify whether new points will be severity 1 or 2
2. Decision Trees: This is another classification algorithm where a tree is created based on the training data. Each leaf is a feature and 2 paths coming out of it. Depending on the value of that feature it will follow a certain path. That is how new entries are classified. Since we want to classify the severity of an accident based on the features this algorithm is ideal.
3. Logistic regression: Logistic regression is a form of linear regression, however where linear regression focuses on predicting continuous values Logistic regression is used to predict categorical values using a similar mechanism. Hence this would be good to use to predict whether the accidents is of severity 1 or 2

4. SVM works by mapping data to a high-dimensional feature space so that data points can be categorised. Hence this algorithm will be used for classification purposes

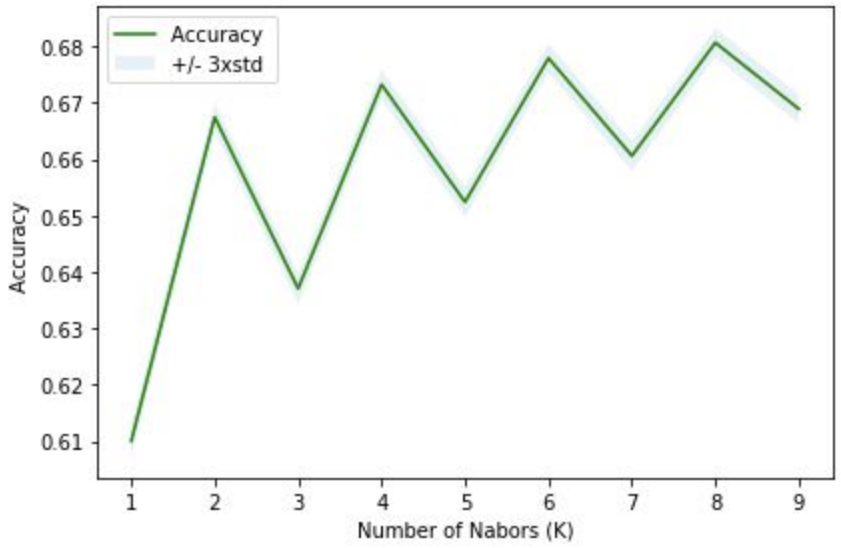
4.2. Metrics Used to Evaluate Model

The two metrics that will be used measure the performance of the models are

- Jaccard Similarity Score: This will prove a score of how many of the predicted values produced by the model match the actual values
- F1-score: This metric uses a confusion matrix to calculate the recall and precision values and then calculates the average to produce the F1-score

4.3. Best K Neighbour for KNN

For the KNN algorithm to be effective we need to find the optimum number of neighbours before a datapoint can be classified, hence we run through setting k neighbours from 1 to 10. The graph shows the accuracy of the model with different ks.



From here we have determined that the best K to be used is 8

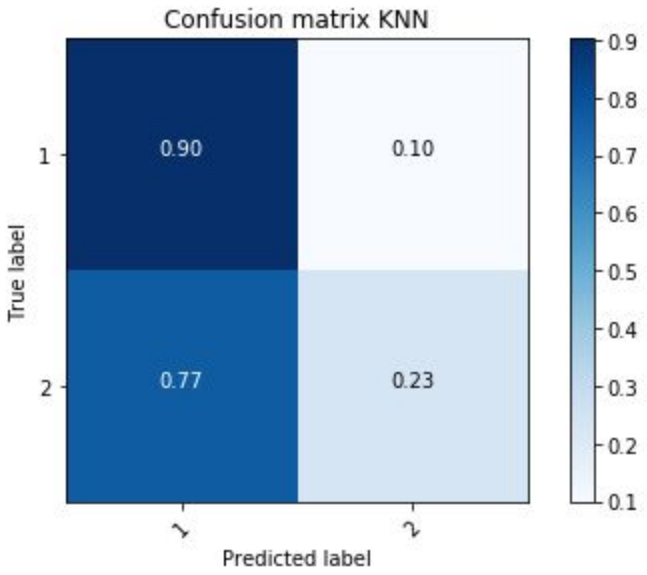
Running the model with K = 8 is:

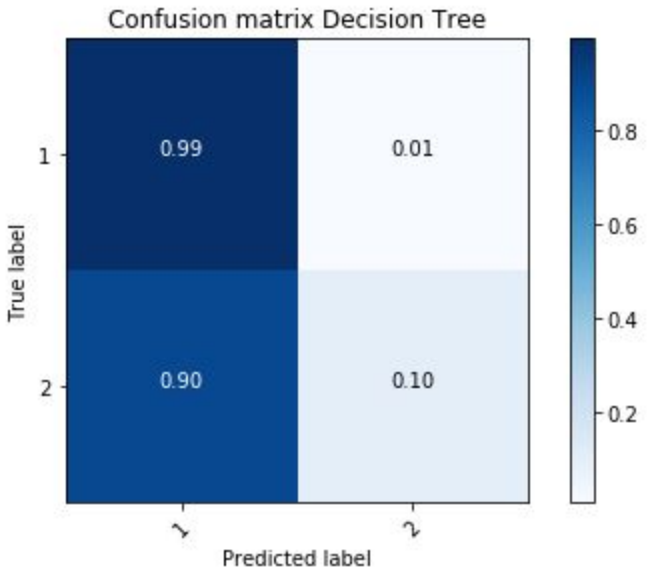
5. Machine Learning Results

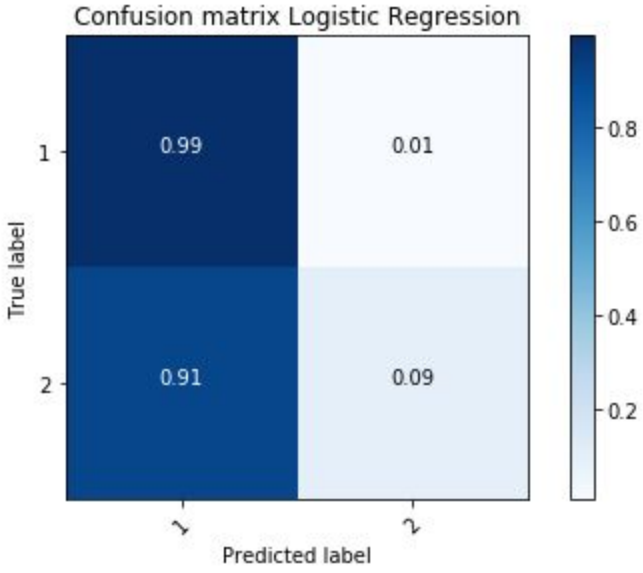
The table below shows the results of the algorithms

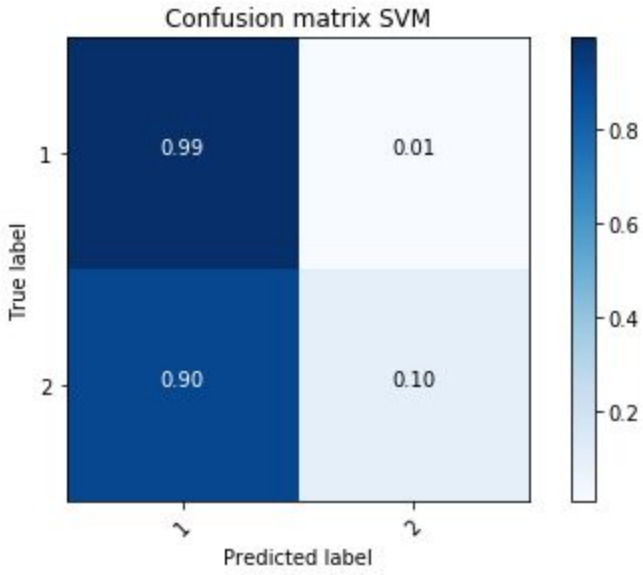
	KNN	Decision Tree	Logistic Regression	SVM
Jaccard Similarity Score	0.680	0.698	0.696	0.697
F1-Score	0.637	0.605	0.601	0.603

On closer inspection of the F1-score we have the following Confusion Matrixes for the different algorithms




Confusion matrix Decision Tree
0.99
0.01
0.90
0.10
1
2
True label
1
2
Predicted label
0.8
0.6
0.4
0.2


Confusion matrix Logistic Regression
0.99
0.01
0.91
0.09
1
2
True label
1
2
Predicted label
0.8
0.6
0.4
0.2



6. Discussion

From the results above we can see that the Decision tree model has the best Jaccard similarity score when predicting the severity of an accident where as the F1-score has the best result with the KNN algorithm

However from the confusion Matrixes we can see the none of models have high success in accurately predicting correctly when an accident is severity 2 but do have a high accuracy when predicting when the severity of an accident will be 1

7. Conclusion

During the process of this assignment we have analysed the data and during the explanatory analysis stage found the relationships between Light Condition. Road conditions, Weather conditions, Speeding, Under the influence and inattention and traffic accidents and presented the results. We also have mapped the number of incidents per year over the years to show trends

Also we have attempted to create various different classification models to predict the severity of an accident and have found the models are good at predicting when the severity of an accident will be a 1