Coursera Capstone Project

IBM Applied Data Science Capstone

Car accident severity

By: Dr Boaz SAINT-LOUIS

Oct 2020

Contents

1. Introduction /Business Problem

Traffic Collisions are one of the most common cause of human fatality in Seattle. Consequently, this increases the pressure on the public authorities to improve the quality of traffic conditions. In addition, these fatalities may lead to annual rises in insurance premium for the motorists involved in these accidents. I want to explore the Collisions dataset, from Seattle SPOT Traffic Management Division and provided by Coursera, to evaluate which characteristics or features can be used to predict accident "severity".

The target audience is road users in Seattle and those responsible for the implementation of traffic policies in transportation division. Our goal is to help this audience to:

i. Have a better transparency over the key accident “severity” drivers

ii. Incorporate these factors in their travel decision making process

iii. Act on them in order to reduce the occurrence of human fatality due to severe accidents.

2. Data Understanding

The Collisions dataset consists of data on all types of collisions (e.g. Bicycle, Car, Collisions, Pedestrian) in Seattle. All collisions data are provided by SPD and recorded by Traffic Records from 2004 to Present. The data are updated on a weekly basis. The current version contains 194,673 observations and 38 attributes or features.

Given that my goal is to determine the key factors that cause collisions and the level of severity, I will extract, for example, the following features, as predictor variables: weather conditions, road conditions, speeding, light conditions and other relevant factors.

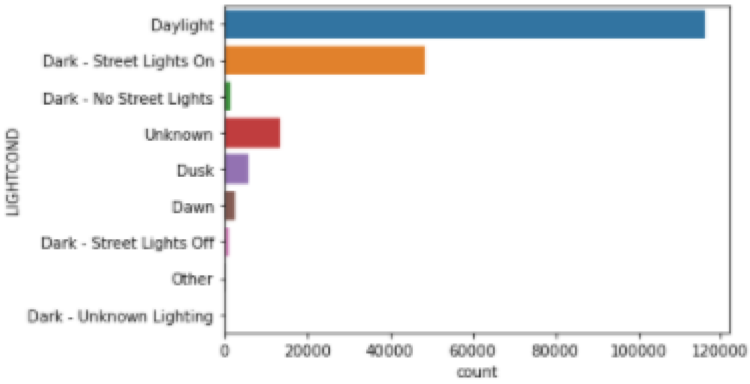
As for accident severity, in terms of human fatality, I will use the severity (SEVERITYCODE) as dependent variable or target. For example, a code of 1 is assigned to accidents which involve property damages while a code of 2 is assigned to accident with “injury”.

I use Jupyter notebook, *Pandas* and *NumPy* to load and analyse the data; Machine Learning (ML) and data science techniques to develop a model to predict car accident "severity" in terms of human fatality.

3. Methodology

The dataset also contains several missing values. I will identify these missing values and explore different methods to handle (remove or replace them) them. In addition, given the dataset size is quite substantial, there are several attributes (e.g. XCEPTRSNCODE, SDOT_COLCODE) that are irrelevant for the analysis. I will exclude these attributes from the analysis.

Moreover, as per below table, most of the variables are categorical variables. As such, it is not possible to perform regression analysis. Therefore, later this in this project, I will use *Pandas* function *'get_dummies'* to convert them into numerical values in order to use Machine Learning techniques and perform regression analysis.



The dataset is quite challenging. First of all, it is unbalanced with respect to the target variable (SEVERITYCODE). Secondly, it contains a mixture of categorical and numerical values. Thirdly, some features contain several values identified as "other" "?" and/or "Unknown" etc). Fourth, there is another category of features that are not consistent as they contain a mixture of numerical and categorical values (e.g. N, Y, 1,0). To deal with these challenges, I proceed in 4 steps:

- Step 1: I replace all values identified as "other" "?" and/or "unknown" by 'NaN'.
- Step 2: Numerical features - Only 2 numerical features have missing values X (Latitude) and Y(Longitude). I simply drop the whole rows with NaN in "X" and "Y" columns
- Step 3: Categorical variables - For these variables, it seems to be reasonable to replace the missing value ('NaN') by the column frequency which is the most common value in the columns. This is the case for ADDRTYPE, LOCATION, COLLISIONTYPE, JUNCTIONTYPE, etc.
- Step 4: Some features are not consistent as they contain a mixture of numerical and categorical values. I standardise these columns by converting all the numerical values to categorical ones. This is the case for 'UNDERINFL', where I replace 0 with 'N' and 1 with 'Y'.

3.1. Identification of missing values

The missing values are automatically converted to Python's default (NaN). I use Python's built-in functions, 'is null()', to identify these missing values.

3.2 Handle the missing data

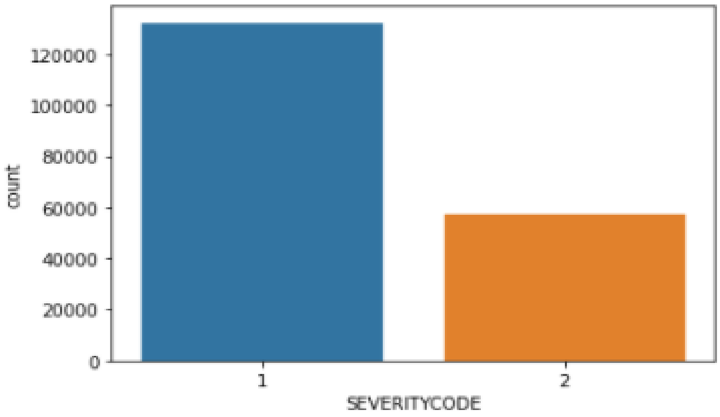
By using a *for loop* in Python, I figure out the number of missing values in each column. In addition, I calculate the percentage of missing values for each feature. Any feature which contains more than (80%) of missing values are dropped from the dataset.

As explained above, more than 80% of entries for the following features 'INATTENTIONIND'(85%), 'PEDROWNOTGRNT'(98%) and 'SPEEDING'(95%) are empty. Therefore, no meaningful analysis can be derived from these features. These 3 additional features are dropped from the dataset. This leaves 194,673 observations and 15 features which are:

FEATURES	DESCRIPTION
X	Latitude
Y	Longitude
COLLISIONTYPE	Collision type
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
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state
VEHCOUNT	The number of vehicles involved in the collision
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol
JUNCTIONTYPE	Category of junction at which collision took place
LOCATION	Description of the general location of the Collision
WEATHER	A description of the Weather conditions during the time of the collision
ADDRTYPE	Collision address type (Alley, Block, Intersection)
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)

3.3 Handle the unbalanced data

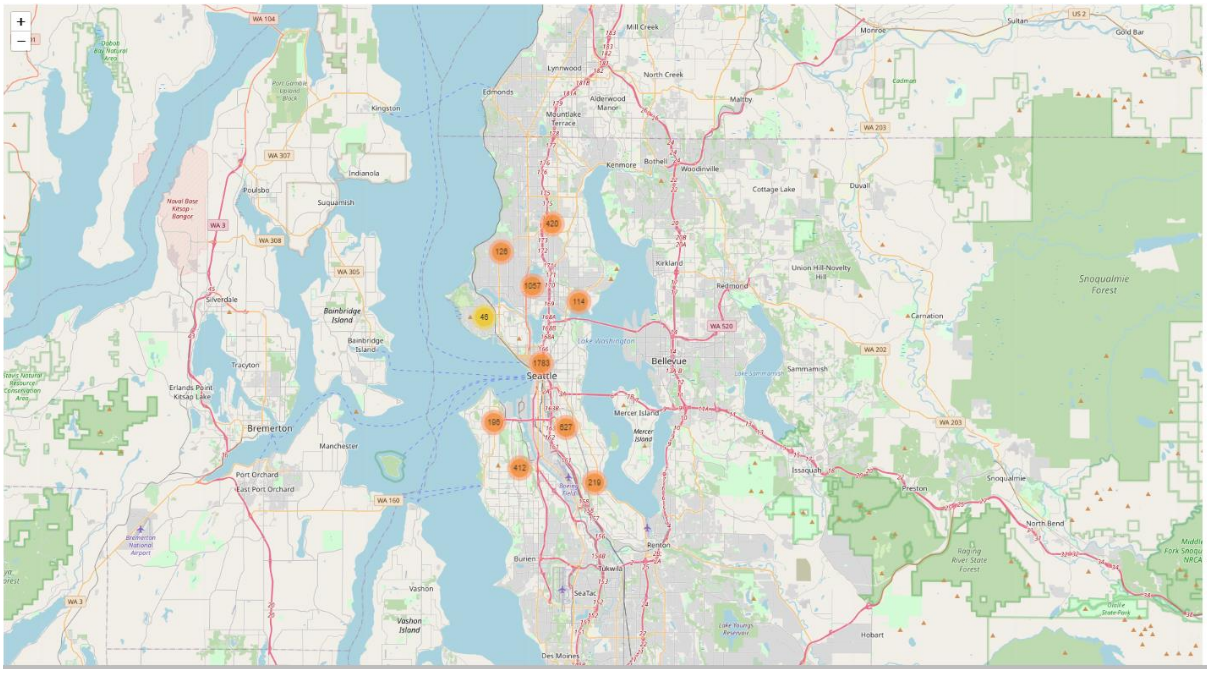
The dataset is unbalanced with respect to the target variable (SEVERITICODE): 57118 accidents identified with injury as severity (injury =2), as opposed to 132221 accidents identified with property damage (prop damage = 1). The number of class differs drastically which will lead to a biased machine learning model (ML), in the sense that the data are skewed toward the bigger class (SEVERITY CODE = 1). As such the ML won't be able to learn the minority class, in this instance SEVERITYCODE = 2. To handle this issue, I generate a balanced dataset by using '*resample*' function from sklearn. The below graph shows the unbalanced dataset with respect to the target variable.



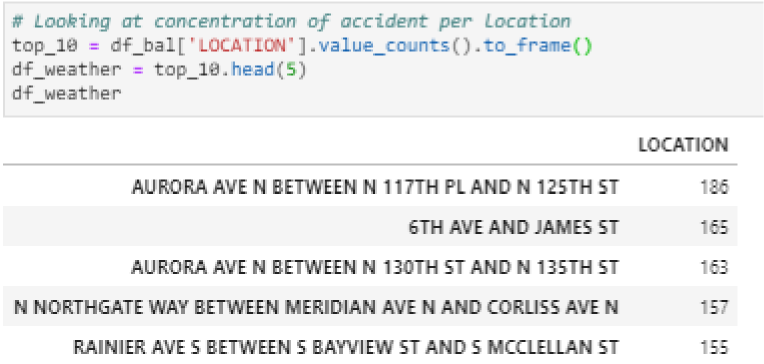
3.4 Visualisation

First, I generate a graphical representation of Seattle Choropleth map in order to determine the areas with the highest concentration of accidents.

Second, I use pandas *groupby* method to investigate the accident severity causes by looking at different features associated with the accident severity in those areas. The below map shows the higher concentration of accidents is in the city centre in the neighbourhood of the highway.



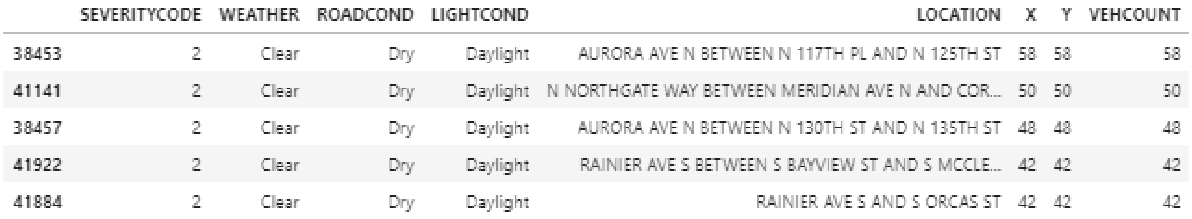
The *Seaborn* Library and data *count pivot table* show a concentration of accidents in 5 locations. The next step is to investigate the potential reasons as to why these accidents appear to be concentrated in these particular areas.



```
# Looking at concentration of accident per Location
top_10 = df_bal['LOCATION'].value_counts().to_frame()
df_weather = top_10.head(5)
df_weather
```

	LOCATION
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST	186
6TH AVE AND JAMES ST	165
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	163
N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N	157
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST	155

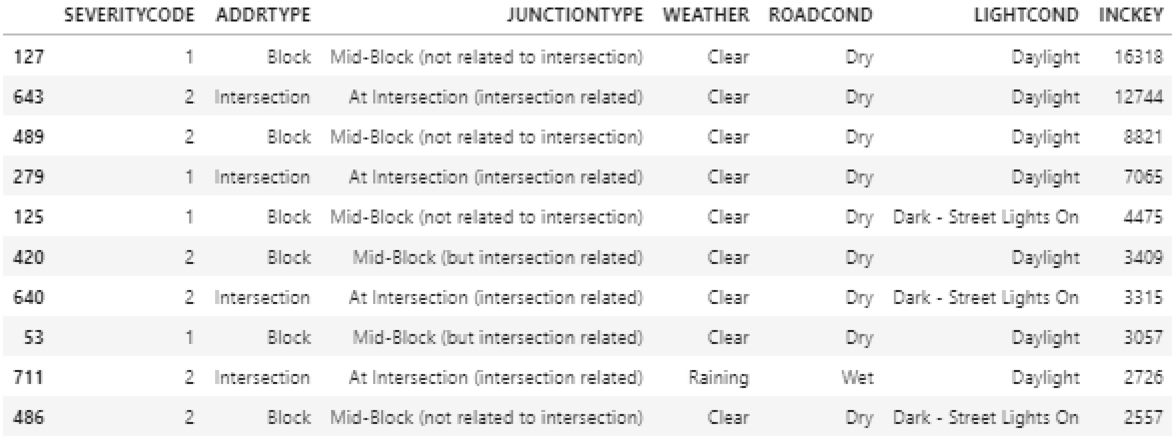
By looking at the number of vehicles involved in these accidents, I find out these accidents include a high number of vehicles. The table below shows the top 5 accidents, in terms of number of vehicles. This leads to understand these areas seem to have heavy traffic during day time.



	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	LOCATION	X	Y	VEHCOUNT
38453	2	Clear	Dry	Daylight	AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST	58	58	58
41141	2	Clear	Dry	Daylight	N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND COR...	50	50	50
38457	2	Clear	Dry	Daylight	AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	48	48	48
41922	2	Clear	Dry	Daylight	RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLE...	42	42	42
41884	2	Clear	Dry	Daylight	RAINIER AVE S AND S ORCAS ST	42	42	42

At this stage, I investigate further in order to understand the potential causes of these accidents. I would expect a close relationship between weather, light, road conditions and the occurrence of these accidents. Surprisingly bad weather, road and light conditions do not seem to be a major contributing factor to accident severity.

By using the attribute 'INCKEY' to count the number of unique incidents, there seems to be little evidence that these attributes are a major contributing factor. However, the junction category appears to be a major cause of accident severity. For example, the below table shows the top 10 accidents in terms of cumulative number of accidents. Most of them are related to junction type with intersection or block related to intersection.



	SEVERITYCODE	ADDRTYPE	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	INCKEY
127	1	Block	Mid-Block (not related to intersection)	Clear	Dry	Daylight	16318
643	2	Intersection	At Intersection (intersection related)	Clear	Dry	Daylight	12744
489	2	Block	Mid-Block (not related to intersection)	Clear	Dry	Daylight	8821
279	1	Intersection	At Intersection (intersection related)	Clear	Dry	Daylight	7065
125	1	Block	Mid-Block (not related to intersection)	Clear	Dry	Dark - Street Lights On	4475
420	2	Block	Mid-Block (but intersection related)	Clear	Dry	Daylight	3409
640	2	Intersection	At Intersection (intersection related)	Clear	Dry	Dark - Street Lights On	3315
53	1	Block	Mid-Block (but intersection related)	Clear	Dry	Daylight	3057
711	2	Intersection	At Intersection (intersection related)	Raining	Wet	Daylight	2726
486	2	Block	Mid-Block (not related to intersection)	Clear	Dry	Dark - Street Lights On	2557

Hence, after these analyses, I decide to develop a model by using the following features, as predictor variables, in order to predict accident severity: 'X', 'Y', 'INCKEY', 'VEHCOUNT', 'ADDRTYPE', 'WEATHER', 'JUNCTIONTYPE', 'ROADCOND' and 'LIGHTCOND'. The target variable, Y, is captured by using the 'SEVERITYCODE' which is a binary variable (0/1).

4. Model Selection

Three machine learning models are used for this exercise. These are Logistic Regression, Decision Tree and a Deep Neural Network. Given that I convert the Target Variable into a discrete binary variable of 0 and 1, this makes the Logistic Function a natural fit for this classification problem. Decision Tree is a tree-structured classifier with three types of nodes, typically used to solve both Regression and Classification tasks, though it is used more often for the latter. As the for the Neural Network, it is chosen for its accuracy on large datasets, in this instance 114236 observations after rebalancing the data. Note that SVM is excluded on purpose. It's performing poorly on large datasets and it's extremely slow as well.

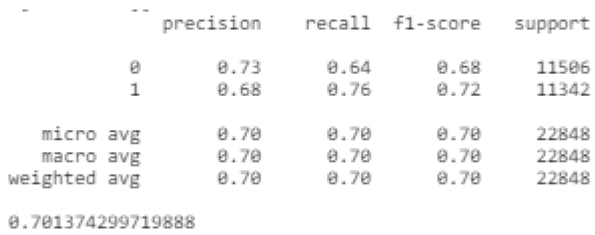
Finally, as mentioned in section 3, the dataset is made up of categorical and numerical values. As such, it is not possible to train machine learning on categorical values. To overcome this challenge, I have encoded all categorical values by using the Pandas '*get_dummies*.

5. Results

5.1. Logistic Regression

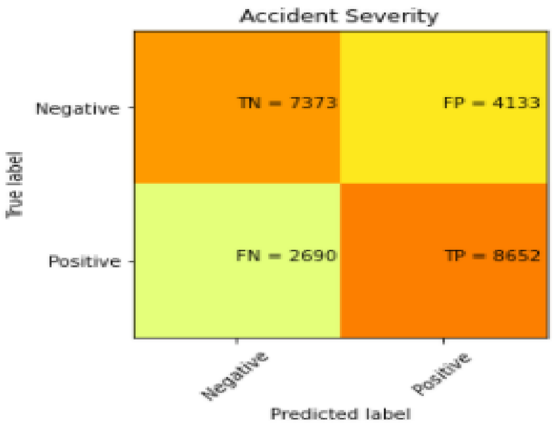
The target variable, 'SEVERITYCODE' is a discrete binary variable of 0 and 1, the logistic model appears a natural fit. Model accuracy, classification report and confusion matrix are shown below:

- **Classification Report**



```
 -          --
              precision    recall  f1-score   support

           0       0.73      0.64      0.68     11506
           1       0.68      0.76      0.72     11342

   micro avg       0.70      0.70      0.70     22848
   macro avg       0.70      0.70      0.70     22848
weighted avg       0.70      0.70      0.70     22848

0.701374299719888
```

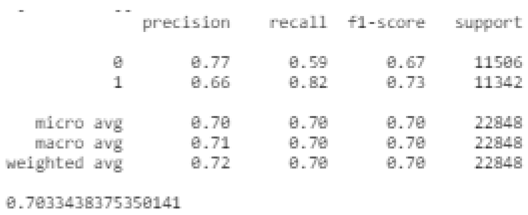
- **Confusion Matrix**



5.2. Decision Tree

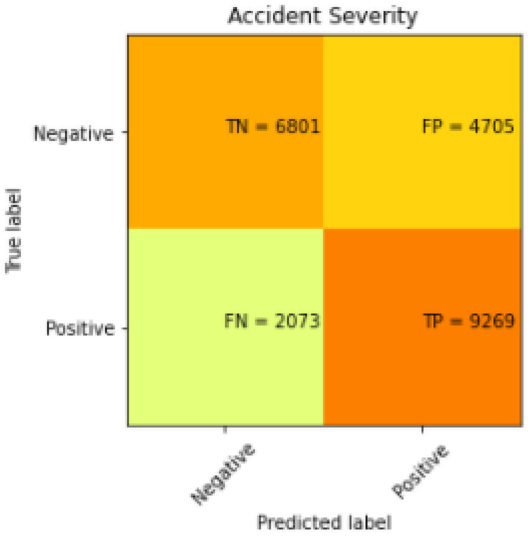
The decision tree is typically used to break down complex problems or branches. Each branch of the decision tree could be a possible outcome. Hence, one of the key advantages of the decision tree is that it provides all the possible outcomes (e.g. different junction types, different road conditions etc.), with possible probabilities attached to the final outcome. The decision tree results, confusion matrix, classification report and accuracy are:

- **Classification Report**



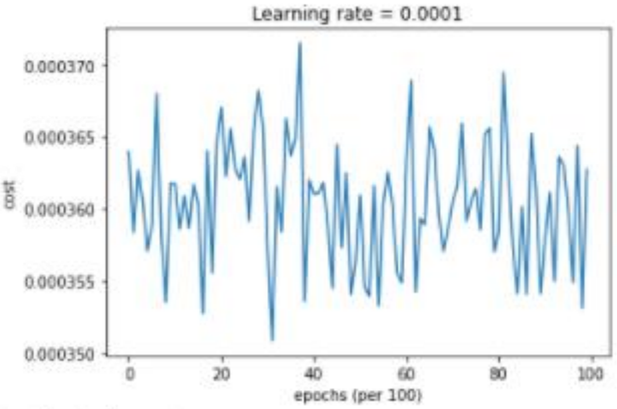
```
              precision    recall  f1-score   support

           0       0.77      0.59      0.67     11506
           1       0.66      0.82      0.73     11342

   micro avg       0.70      0.70      0.70     22848
   macro avg       0.71      0.70      0.70     22848
weighted avg       0.72      0.70      0.70     22848

0.7033438375350141
```

- **Confusion Matrix**



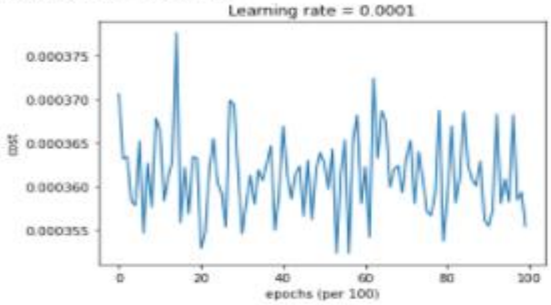
5.3. Deep Neural Network

I test a Deep Neural Network with 2 layers by using a *ReLu* activation function for the hidden layers and a *sigmoid* activation function for the output layers. I use a '*He initialization*' to initialise the weight and a mini batch gradient descent in order to improve the algorithm performance. Finally, I test the model by using different learning rates (0.0001, 0.00005, 0.00006). The results below show a train and test accuracy of about 60%. The model doesn't even fit the training data which is typically the sign of a model which is underfitting.

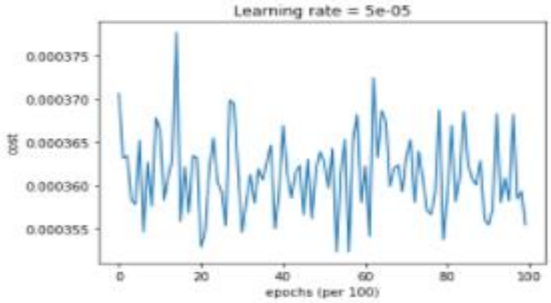


```
On the train set:
Accuracy: 0.5948702236617499
On the test set:
Accuracy: 0.5976015406162465
```

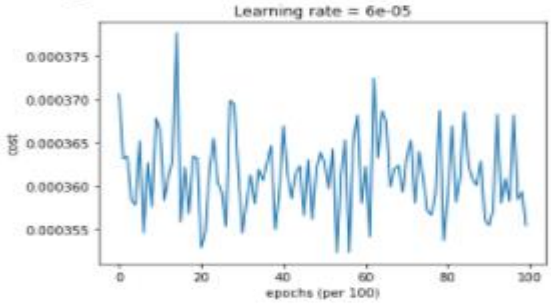
```
learning rate is: 0.0001
```



```
------------------------------------------------------------
learning rate is: 5e-05
```

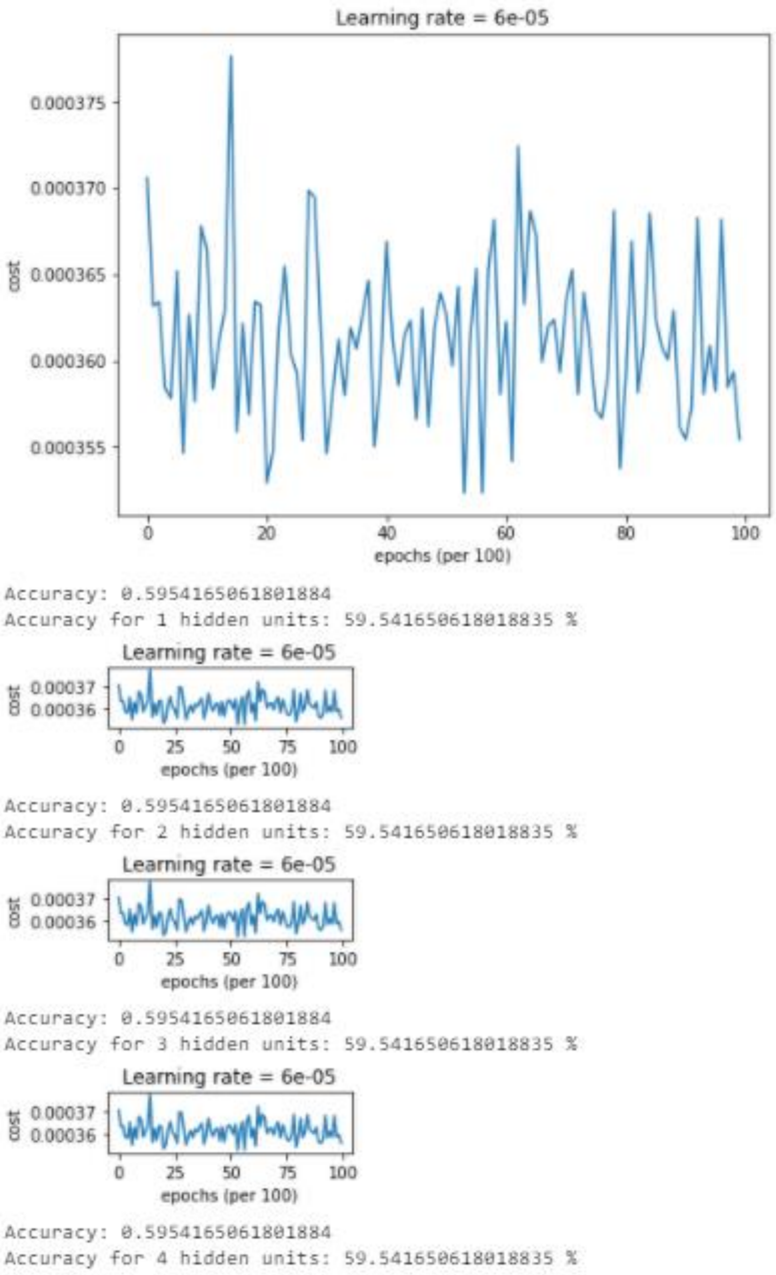


```
------------------------------------------------------------
learning rate is: 6e-05
```



```
------------------------------------------------------------
```

Even after increasing the layer size to 4 layers, the model accuracy doesn't seem to be improving.



```
Accuracy: 0.5954165061801884
Accuracy for 1 hidden units: 59.541650618018835 %
```

```
Accuracy: 0.5954165061801884
Accuracy for 2 hidden units: 59.541650618018835 %
```

```
Accuracy: 0.5954165061801884
Accuracy for 3 hidden units: 59.541650618018835 %
```

```
Accuracy: 0.5954165061801884
Accuracy for 4 hidden units: 59.541650618018835 %
```

6. Discussion

With an accuracy of about 60%, the Deep Neural Network doesn't seem to perform better than neither the Logistic Regression (70%) nor the Decision Tree (70%). In addition, the poor performance on the training data is a sign of underfitting. Therefore, the models need to be trained on more data which is surprising considering the dataset was quite substantial. This may be due to the fact that some important features (e.g. speeding) are dropped as predictor variables. Another factor could be the removal of more than 5000 rows due to missing data values given that it was not possible to draw any meaningful analysis from them.

In future, it would be good for the data collector to enhance the data quality to ensure the inclusion of potentially useful information.

7. Conclusion

When beginning the analysis of the dataset, I was optimistic there would be a strong relationship between road, weather, light conditions, even driving under the influence of drugs or alcohol, and accident severity. However, these attributes do not seem to be a major contributing factor of car accident severity. By using INCKEY to count the number of unique incidents, there seems to be little evidence that these attributes are a contributing factor. However, the junction category appears to be a major cause of accident severity. For example, in 3.4 above, among the top 10 accidents, most of them are related to junction type with intersection. Moreover, as shown by the Choropleth map, there seems to be a high concentration of accidents, with high number of vehicles, in some specific areas (e.g Aurora Ave N Between N 117th PL and N 125th ST, N Northgate Way Between Meridian Ave and Corlis Ave N) which are in the city neighbourhood. This seems to indicate these areas attract heavy traffic during day time.

As preventative measures, Seattle road users should be very careful when approaching these areas or avoid them during the daylight. As for the local authorities, they should explore the possibility of implementing preventative tools for those areas such as traffic light or better signal mechanisms or anything else that can help minimise accident severity in those particular areas.

8. Reference

https://www.coursera.org/specializations/deep-learning

http://cs229.stanford.edu/notes/cs229-notes-dt.pdf