Predicting Accident Severity in Seattle City

Table of Contents

1. Background 2
2. Problem 2
3. Target Audience 2
4. Data acquisition and cleaning 3
   a) Data Acquisition 3
   b) Feature Selection 3
   c) Data cleaning 4
5. Methodology 4
6. Exploratory Data Analysis 6
7. Results 11
8. Conclusion 13
9. Future direction 13

1. Background

According to a report by CDC in 2019, road traffic crashes are a leading cause of death in the United States for people aged 1–54 and the leading cause of non-natural death for healthy U.S. citizens residing or traveling abroad.

The World Health Organisation (WHO) in a 2018 report profiled the following facts concerning global road traffic injuries and deaths;

- Each year, 1.35 million people are killed on roadways around the world.
- Every day, almost 3,700 people are killed globally in road traffic crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians. More than half of those killed are pedestrians, motorcyclists, and cyclists.
- Road traffic injuries are estimated to be the eighth leading cause of death globally for all age groups and the leading cause of death for children and young people 5–29 years of age.

It is against this background that the author found it imperative and fit to suggest a measure that could help in minimising accidents and their impact in Seattle City.

2. Problem

Car crashes are a public health concern both globally and in the United States but, these injuries and deaths are preventable and the impact and consequences of these accidents can be minimised. Conventional techniques and methodologies have been used in the past to predict the severity of clashes, though these had a number of drawbacks in producing quality and accurate inferences. This project aims to predict the severity of accidents and how the impact can be minimised based on a number of factors and for the purposes of this project we will use data relating to Seattle city.

3. Target Audience

The solution seeks to provide aid to the Seattle's Department of Transportation (SDOT) in its transportation infrastructure planning, building and maintenance to ensure that it tailor makes its road networks in a manner that addresses the rise in accidents.

The solution will also help the Department of Health in planning for relevant equipment and resources required based on the predicted accident severities. This will help them in acquiring resources that appropriately address the problems and injuries on a particular accident scene.

Road users in general, that is, pedestrians and motorists will be advised accordingly if the information is publicly availed to Seattle citizens and passer-by's to take precautionary measures to reduce severity of accidents.

4. Data acquisition and cleaning

a) Data Acquisition

The accident data will be used to predict the severity of an accident given certain features. The dataset was provided within the Coursera course content.

Y = SEVERITYCODE (1 or 2)

Severity Code 1: Injury Collision

Severity Code 2: Property Damage

Total Number of features in dataset: 37

Features selected after Feature Engineering: (X)

b) Feature Selection

The table below show the features that were selected from the dataset, descriptions and basis for selection. The represent X stated above.

Feature	Description	Reason for selection
X	X coordinates for an accident scene	Gives the exact spot of an accident together with the Y coordinates
Y	Y coordinates for an accident scene	Gives the exact spot of an accident together with the X coordinates
PERSONCOUNT	The total number of people involved in the collision	Severity indicator
PEDCOUNT	The number of pedestrians involved in the collision. This is entered by the state.	Severity indicator
PEDCYLCOUNT	The number of bicycles involved in the collision. This is entered by the state.	Severity indicator
VEHCOUNT	The number of vehicles involved in the collision. This is entered by the state.	Severity indicator
SDOT_COLCODE	A code given to the collision by SDOT	Collision detail and severity indicator
ROADCOND	The condition of the road during the collision.	Accident catalyst
SEGLANEKEY	A key for the lane segment in which the collision occurred.	Accident location
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.	Accident location
month	Month in which the accident occurred.	Are number of accidents related to the month of the year
HITPARKEDCAR	Whether or not the collision involved hitting a parked car. (Y/N)	Predictor for type of severity
SPEEDING	Whether or not speeding was a factor in the collision.(Y/N)	Cause & effect determination of accidents
PEDROWNOTGRNT	Whether or not the pedestrian right of way was not granted. (Y/N)	Cause & effect determination of accidents
INATTENTIONIND	Whether or not collision was due to inattention.(Y/N)	Cause & effect determination of accidents
SDOT_COLDESC	A description of the collision corresponding to the collision code.	Collision detail and severity indicator
ADDRTYPE	Collision address type: Alley; Block; Intersection	Accident location
COLLISIONTYPE	Collision type	Severity indicator
JUNCTIONTYPE	Category of junction at which collision took place	Accident location; Cause & effect determination of accidents
LIGHTCOND	The light conditions during the collision.	Cause & effect determination of accidents
WEATHER	A description of the weather conditions during the time of the collision.	Cause & effect determination of accidents
STATUS	Matched or Unmatched	

UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.	Cause & effect determination of accidents

c) Data cleaning

The following procedures were taken into effect whilst cleaning the data;

- Replacing missing values – default values were used to replace NaN values within the dataset
- Dropping records with null values on X & Y coordinates – only records with locations that can be ascertained was used
- Dropping irrelevant features – this improves the predictive capabilities of the model
- Creating new features e.g. the month column which was computed from the date given
- Label encoding – to change categorical features into machine readable formats
- Feature scaling – to normalize the data

5. Methodology

Predictive analytics is the practice of extracting information from existing data sets in order to determine patterns and predict future outcomes and trends. It has a wide range of applications in different fields, such as finance, education, healthcare, and law (SAS, 2017). I shall likewise use previous similar data to model an algorithm that predicts the severity of an accident based on a number of factors.

For the purposes of this project, the already existing Seattle City accident csv dataset will be used to predict the severity of an accident as such no further data gathering procedures or scraping processes were performed in an attempt to collect the data. The machine learning pipeline entails that after data gathering and securing a storage facility for your data, it is imperative that the data is cleaned and that feature engineering is performed to ensure that data tidiness and quality is achieved. Below are a series of approaches that I performed in coming up with a high quality dataset that was then used to build the machine learning model;

- The first step was to perform data description checks which provided a clear understanding of the nature of data that is available.
- We then dropped useless fields or columns from the dataset which have no impact to the performance of our model as part of the data cleaning process.
- Null variables are filled/replaced or dropped in a data science project, in our case, we opted for the dropping of all NaN records on geo locations provided to ensure that every accidental incident we have has a representation of the actual location or position it occurred. This helps in building a more accurate model with minimal inaccurate inferences.

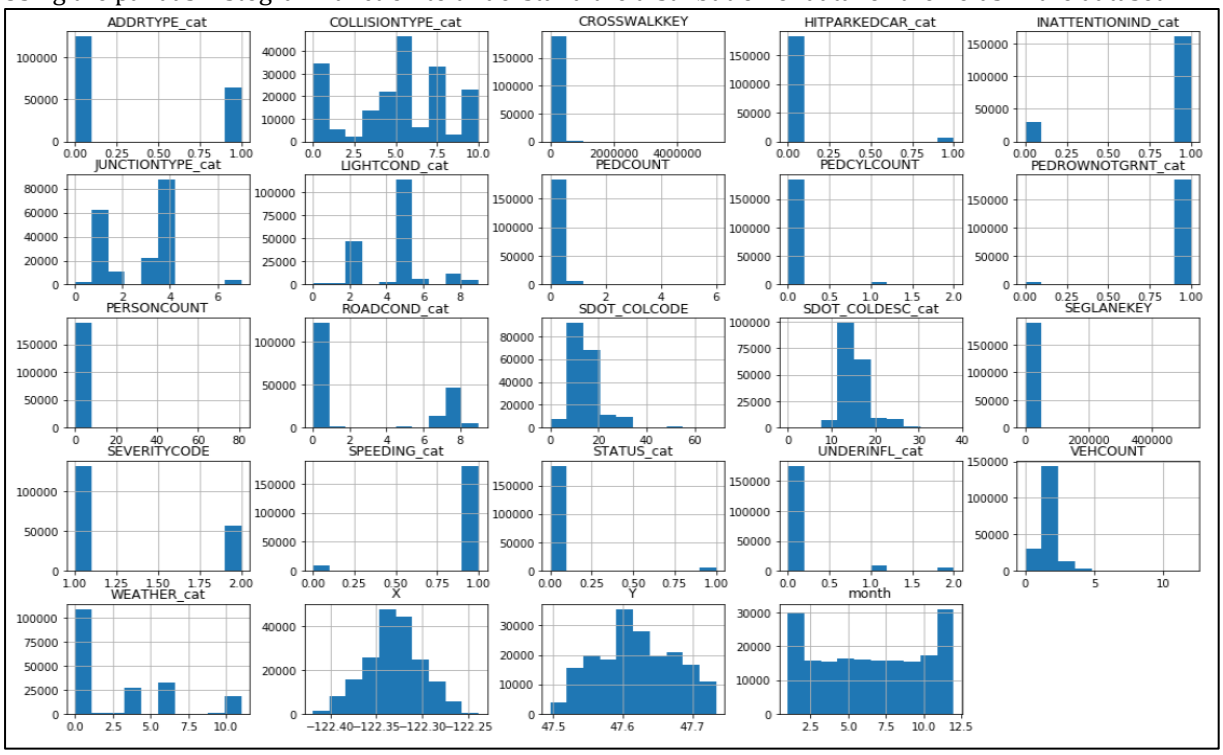
- Using label encoding techniques such as OneHot encoding and the brute force method mainly taking advantage of the Scikit-learn library, we managed to encode the categorical variables into machine readable formats.
- Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA in our case was to learn what our data can tell us. It generally starts out with a high level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modelling choices, such as by helping us decide which features to use. According to https://www.itl.nist.gov, Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to;
    - maximize insight into a data set,
    - uncover underlying structure,
    - extract important variables,
    - detect outliers and anomalies,
    - test underlying assumptions,
    - develop parsimonious models and
    - determine optimal factor settings.
- We used histograms, pie charts, scatter plots and bar graphs to have an in-depth statistical understanding of our features. We also used scatter plots to ascertain the relationships and or correlations within our available features.
- Feature Selection and Feature Engineering to ensure that we use the most relevant variables in building our model.
- Feature Scaling - Standardization was used to normalise the data for the model. Standardization (or z-score normalization) scales the values while taking into account standard deviation. If the standard deviation of features is different, their range also would differ from each other. This reduces the effect of the outliers in the features (https://towardsdatascience.com).
- The dataset is then split into training set and testing set. 80% of the dataset comprises the training set and 20% of the dataset is the testing set. After the development of the model, the model is used on another data set to test its effectiveness. Data used for such purpose is called test data or test set. The reason for using two different sets is to ensure that the model is flexible enough to be used on data sets other than the one it was built with. Otherwise, the problem of overfitting may occur, which is when a model is accurate with its original data set, but performs poorly on other data sets, because it is overly complicated A common method to avoid overfitting is to divide the input data set into training and test sets.
- We used a supervised learning approach in building our model as we tried to build a multi-classification algorithm that predicts the severity of an accident. To ensure that a higher accuracy is obtained we performed hyper parameter tuning to our model and the success of

this approach was being measured using ACCURACY to predict the test set and also using the confusion matrix to check the true/false positives and true/false negatives that were predicted.

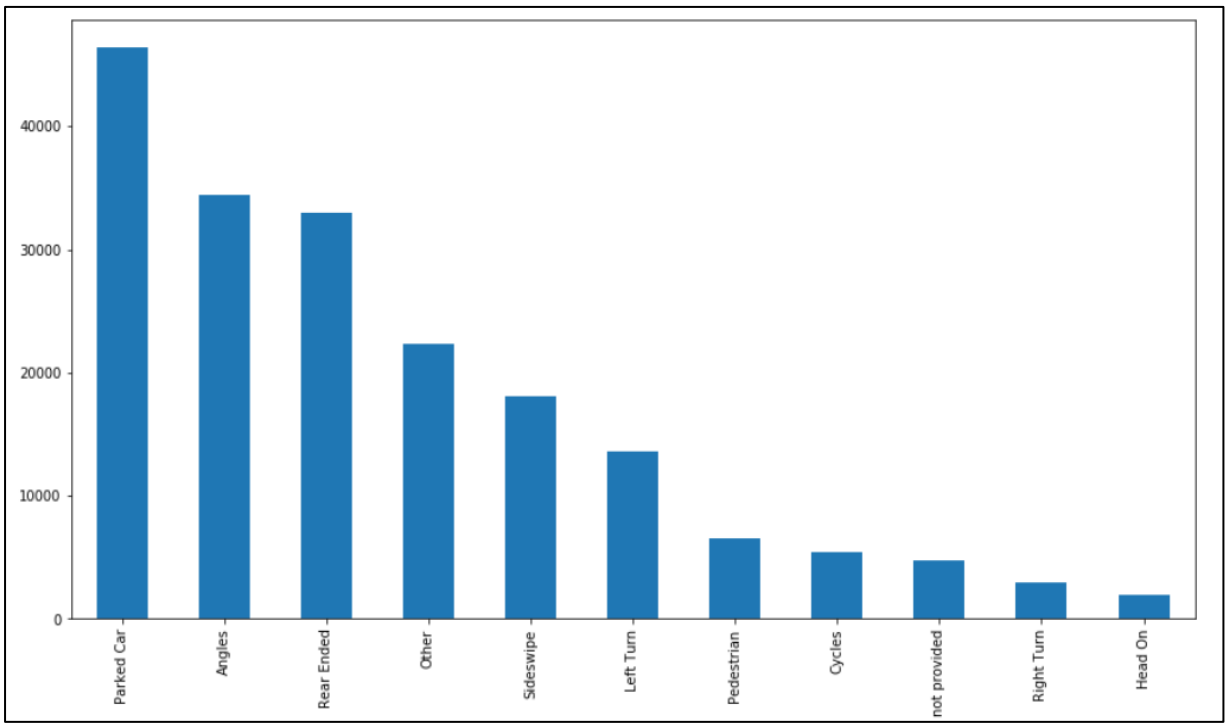
- For the purposes of this prediction model, Softmax and ReLu activation function were used. Softmax was used because it not only maps our output to a [0,1] range but also maps each output in such a way that the total sum is 1. ReLu is non-linear and was used because it has the advantage of not having any backpropagation errors unlike the sigmoid function, The output of Softmax is therefore a probability distribution and the model was trained under a log loss (cross-entropy).

6. Exploratory Data Analysis

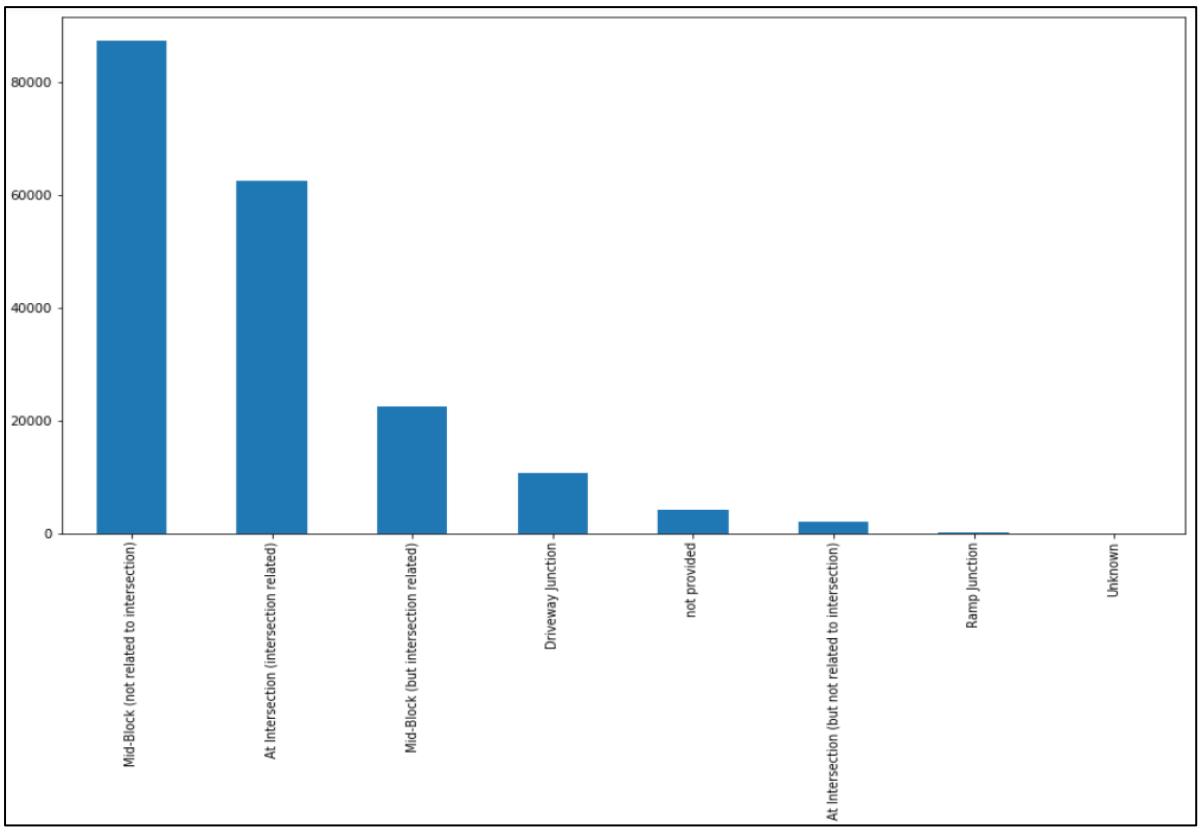
Using the pandas histogram function to understand the distribution of data for the fields in the dataset



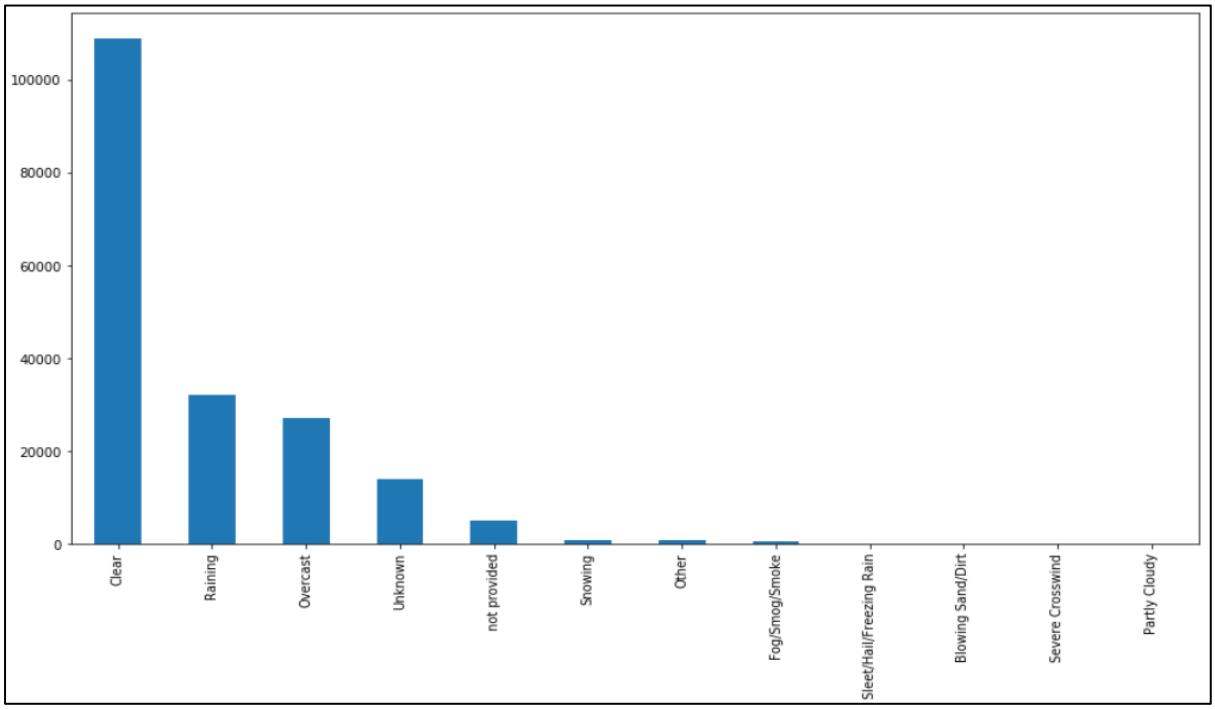
Collision type analysis: It was observed that most collisions are happening with parked cars. This might probe the administration to plan and be forceful on making parking arrangement within the city.



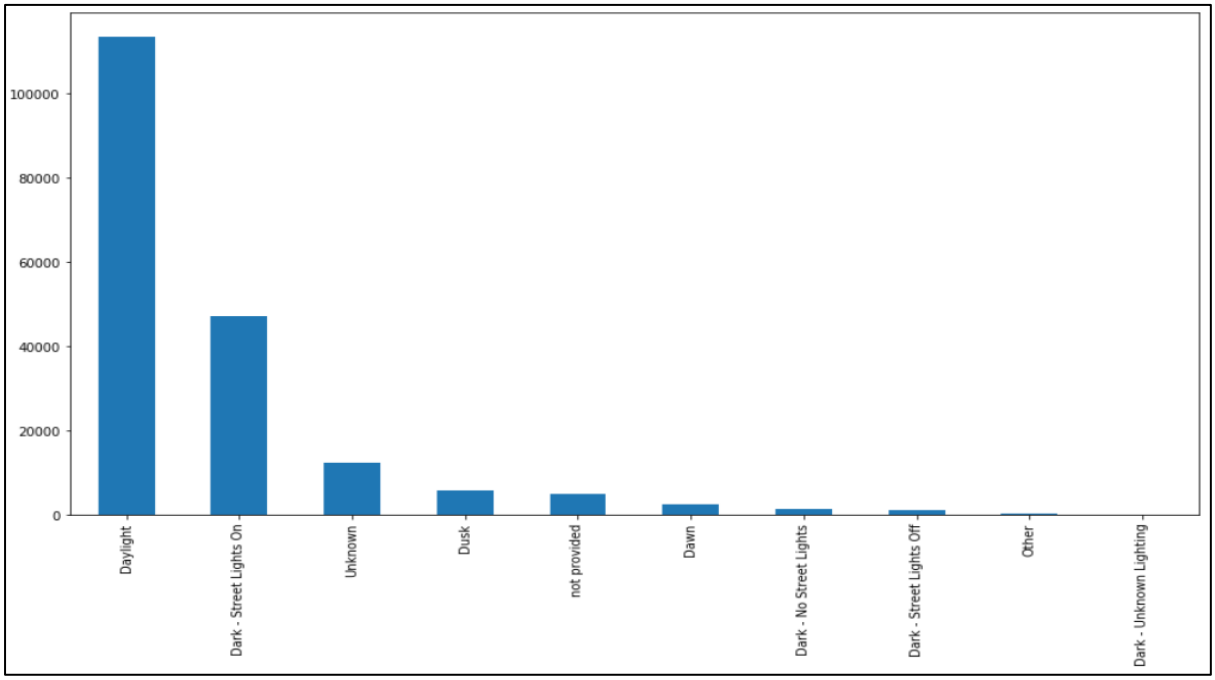
Junction Type Analysis: It was observed that mid-block junctions not related to intersection results in most accidents. There is need for motorists to take precaution when approaching such junctions. The authorities on the other should ensure that awareness and alertness of such junctions is created within motorists and the general public.



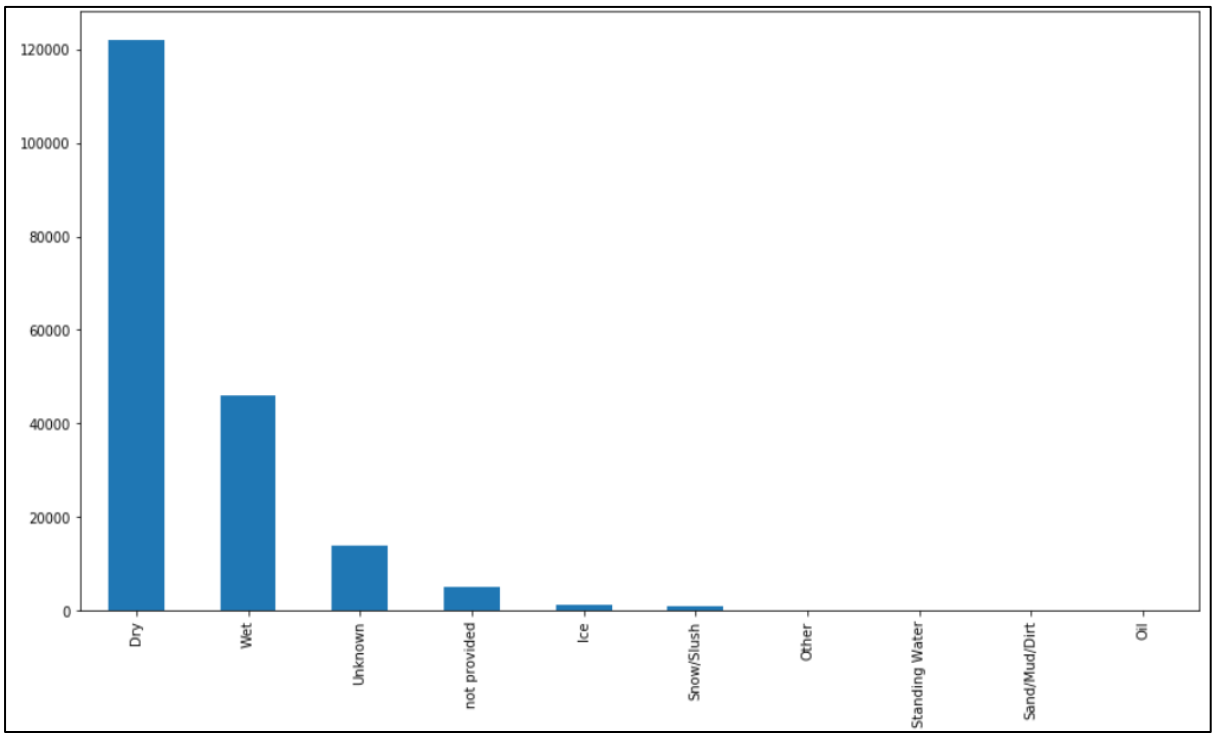
Weather Analysis: It was observed that clear weathers have the highest statistics in terms of accidents. Education and instilling a safety mind in people would be ideal particularly if this is champions by Transport and Health departments.



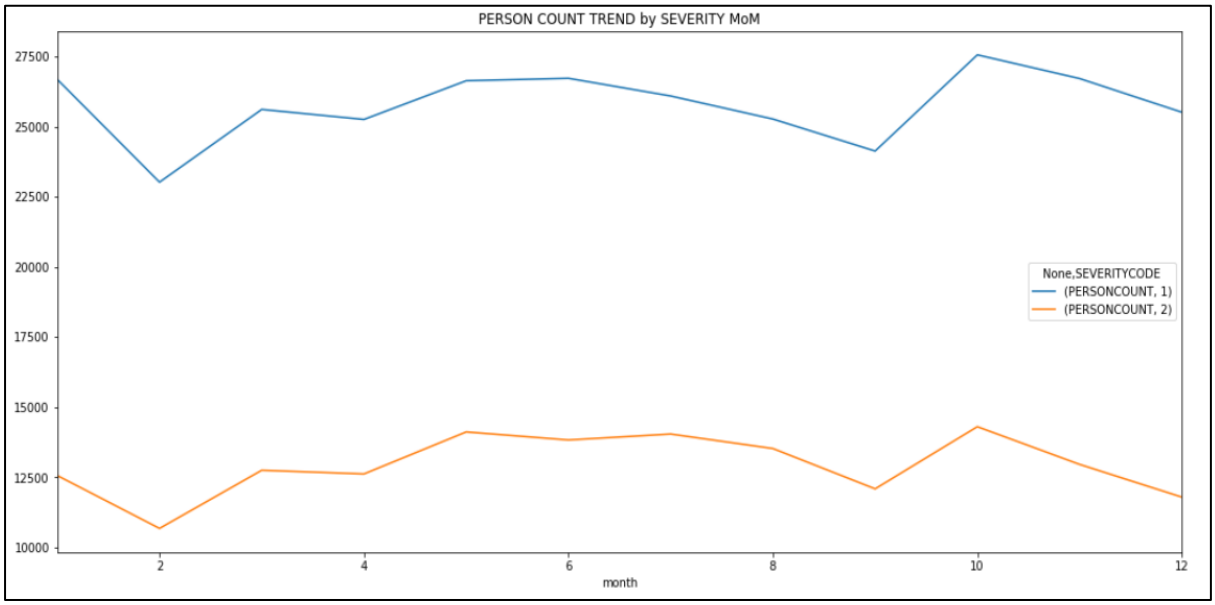
Light condition analysis: Most accidents are happening during daylight whilst the least occur whilst the light condition is dark – unknown lighting. Police spot checks should be introduced to manage traffic during the day to manage the high statistics in accidents.



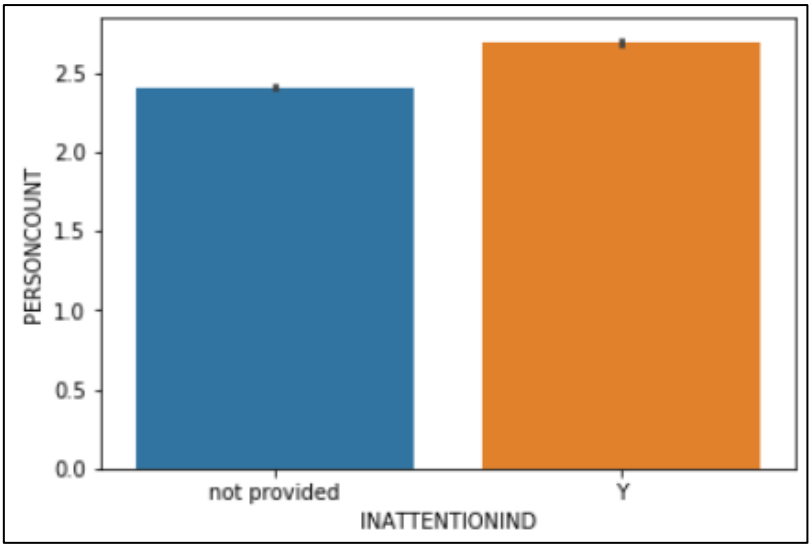
Road condition analysis: It was observed that dry roads have the highest number of accidents as compared to the common wet and oily roads.

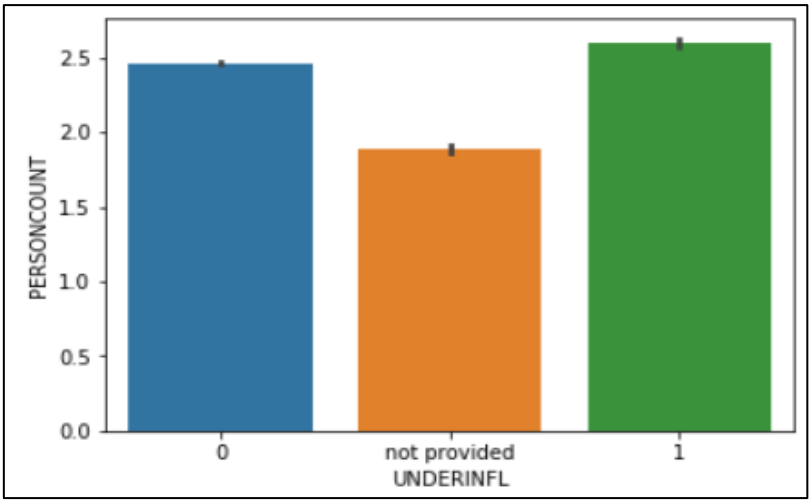


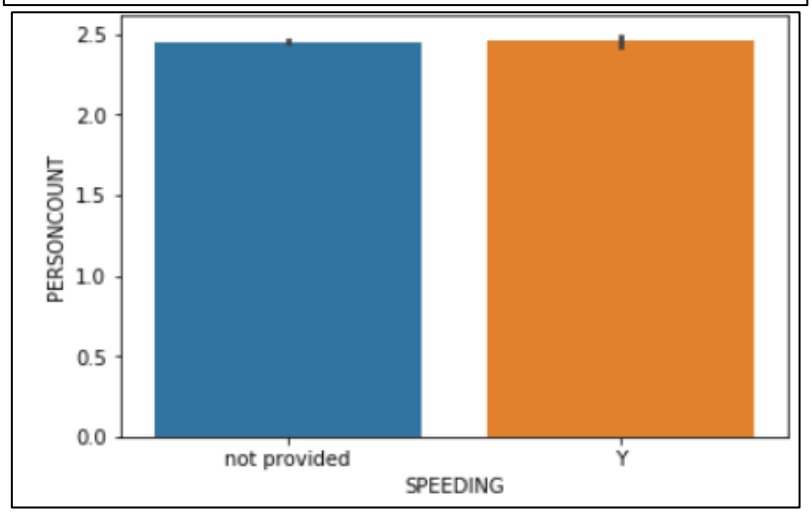
Person Count Trend by Severity Month on Month: It was observed that people involved in collisions in both injury and property damage collisions do spike on the 10th month of the year. Also observed was that property damage accidents have a larger probability of involving more people than injury collisions.



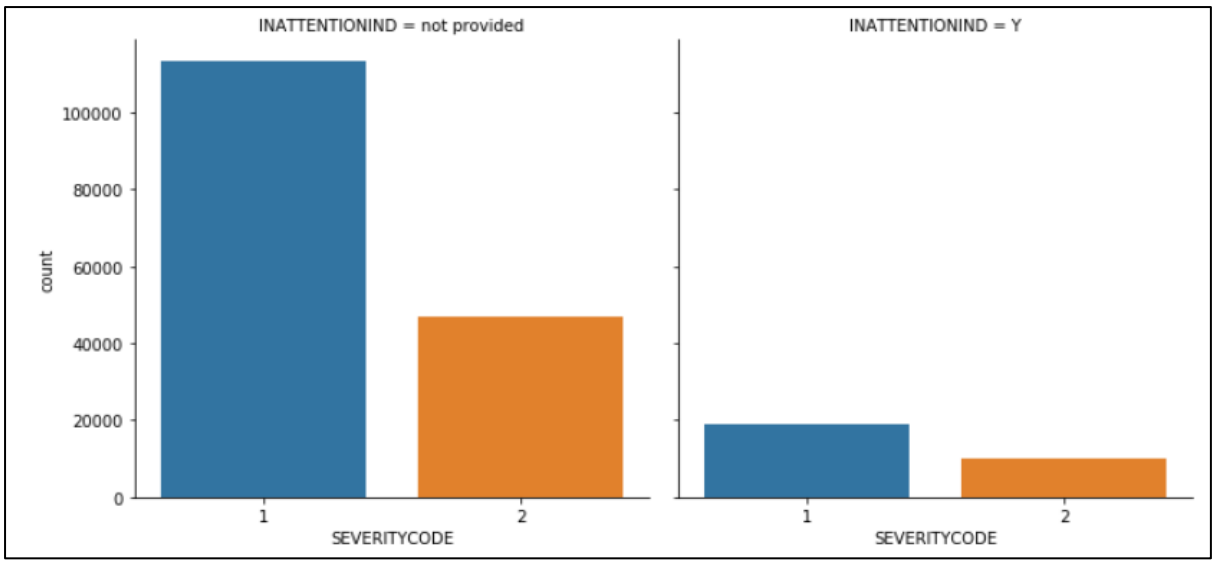
Person Count versus Inattention; under influence and Speeding Analysis: It was observed that Inattention, driving under the influence of alcohol or drugs and speeding results in a number of people involved in an accident. However, these are not directly linked to the number of accidents that happen.







Inattention ID versus Severity Code: It was observed that most injury collisions happen when motorists are well attentive as compared to when they are not paying attention. On the other hand, during times of inattention injury collisions do happen more as well.

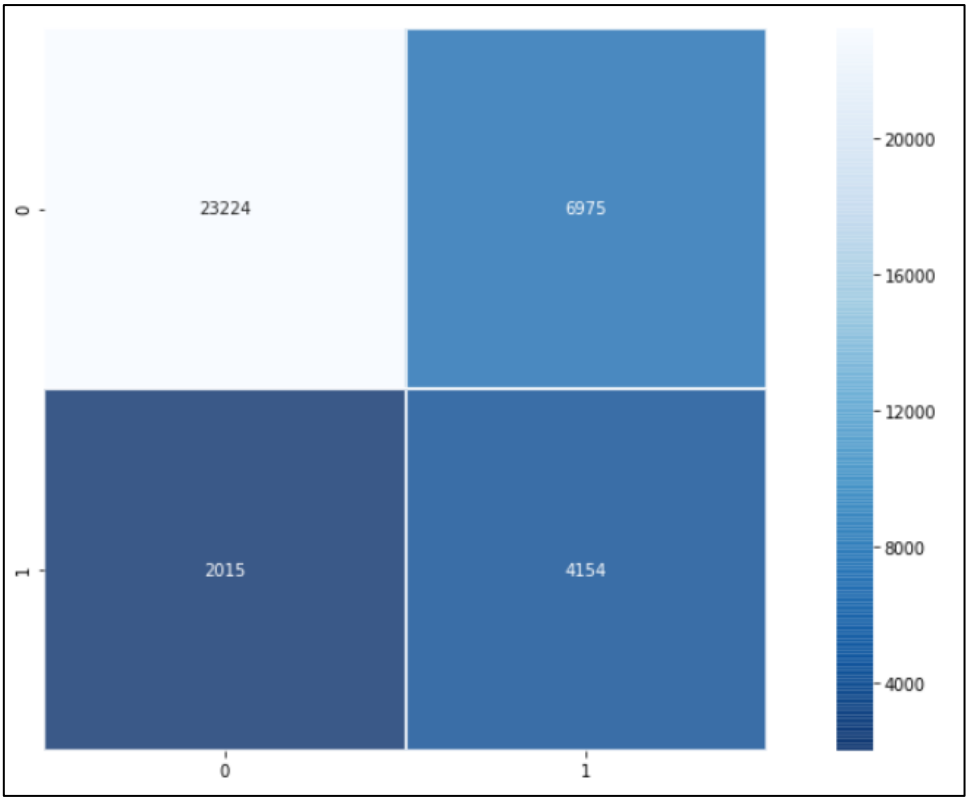


7. Results

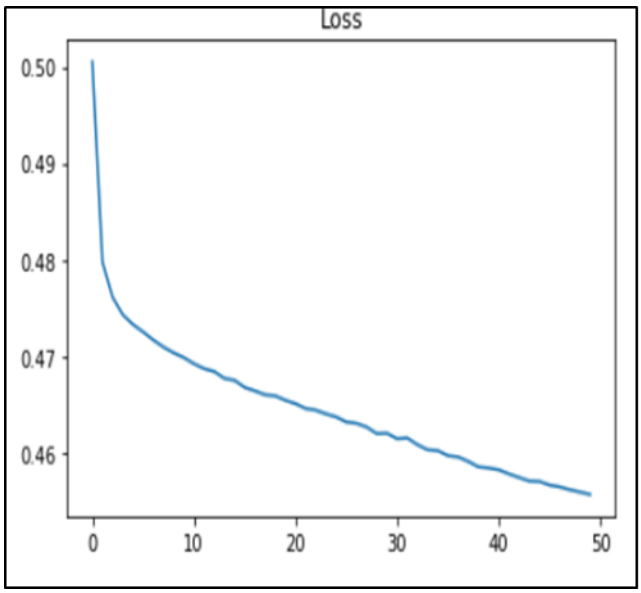
From the Exploratory Data Analysis that we performed, it was observed that the general view that wet roads, overcast weathers and poor lighting have larger contribution to road accidents is inaccurate. From the data analysis performed it was observed that in Seattle accidents do happen more in dry roads, clear weathers and during the day without and artificial form of lighting.

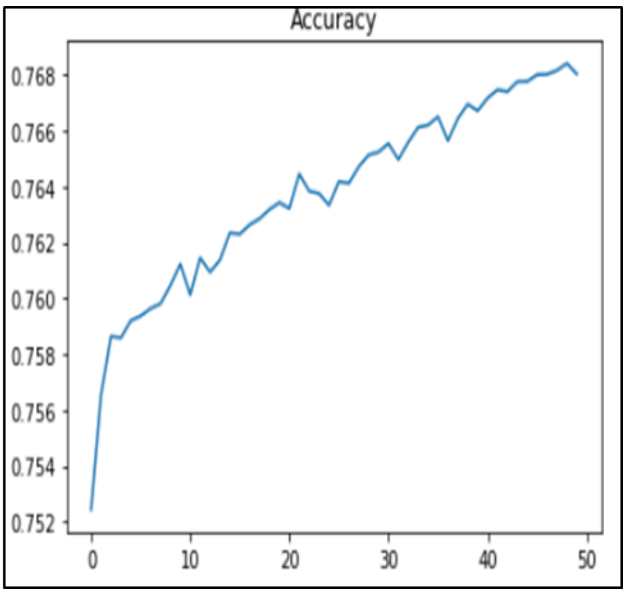
A clear interpretation to this implies that the authorities should do a deep dive into what actually takes place during the day. Could be an issue of day traffic in terms of vehicles and pedestrians that might be contributing to these accidents. There is real need for the Department of transport and other relevant stakeholders to take action and strategize on how accidents can be minimised during these periods and conditions.

The classification algorithm retained an accuracy of 75% having a total of 27,378 true predictions out of a total number of 36,368.



Accuracy and Loss results can be summarised by the graphs below;





Below is a summary of the results from the confusion matrix;

	precision	recall	f1-score	support
0	0.77	0.92	0.84	25239
1	0.67	0.37	0.48	11129
micro avg	0.75	0.75	0.75	36368
macro avg	0.72	0.65	0.66	36368
weighted avg	0.74	0.75	0.73	36368

8. Conclusion

The purpose of the project was to predict the severity of an accident based on the accident data for Seattle City. By applying supervised learning in the form of a multi-classification algorithm, we were able to able an accuracy of 75% which might indicate that the algorithm to a larger extent will be able to predict and help various stakeholders in planning against these accidents. The end result we all wish for will be a decline in the rate of these accidents. Exploratory Data Analysis apart from the model itself was done to give a clear picture to the authorities and various stakeholders on the certain conditions that are mostly resulting in accidents. Exploratory data analysis shows the correlations and behaviour of the variables in relation to accident statistics and severity of the accidents.

9. Future direction

- Data sharing with motorists and pedestrians proactively and in advance before they approach what can be termed a “black spot” to ensure that they proactively take precaution.
- The use of artificial intelligence in self-driving cars to calculate the probability of an accident occurring and assess the risk associated with a particular location and advise accordingly on the best route to take.