I. Introduction

Road accidents have become very common nowadays. As more people are buying automobiles, the road accidents are increasing day by day. Furthermore, people have also become more careless now. Not many people follow the traffic rules. Especially in big cities, there are various modes of transports. Moreover, the roads are becoming narrower and the cities have become more populated. Due to the increasing number of accidents everyday, there should be more warning signs which could tell the drivers that this road is dangerous under certain conditions as rain or dim light or any other condition which could increase the possibility of an accident and how severe it could be so the driver could be alerted and alarmed that they should drive more carefully when passing through this road.

Business problem:

The objective of this capstone project is to analyze and predict the possibility of an accident and it's severity relative to a place. Using data science methodology and machine learning techniques like classification, this project aims to predict the severity of an accident which could occur given some details such as light condition, the type of junction where the accident occurred. This prediction could be used to reduce the number of accidents as signs could be put on each road to warn drivers that this road is dangerous and they should take care when they are passing through. Roads could be closed at certain occasions such as rain, storms … etc if these conditions proved to be dangerous. If the reason behind the accidents is the lights or road condition, it could be fixed to minimize the number of accidents.

II. Data

To solve this problem, we will need the following data:

- A list of accidents occurred in Seattle, Canada.
- Conditions of the road, weather and light.
- Information about each accident.

These data is offered by Ibm cloud (https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). This data contains the accidents occurred in Seattle, Canada from 2004 to present. It contains date of the accident, people and vechiles involved in the accident, the exact location with coordinates and description of the place and also the condition of the road, weather and light when the accident occurred. There is more than 190k of accidents in this data-set which is more than enough to train the model and predict the severity of an accident which could occur on a road.

Data Cleaning:

The column EXCEPTRSNCODE values are NEI or NaN values. The NaN values imply that this accident has enough information. The NEI values imply that the accident report doesn't have enough information. The accidents with NEI value are removed from the dataset.

The column UNDERINFL values are “0”, “1”, Y, N. The values N and “0” are replaced with 0. The values Y and “1” are replaced with 1.

The columns which have many NaN values are dropped from the dataset

The rows which have NaN values are also dropped from the dataset.

III. Feature Selection:

After cleaning the data, there were 178252 sample and 38 feature in the data. There were some redundancy in the features. Some features like OBJECTID, INCKEY, COLDETKEY...etc served the same purpose which is giving each accident unique value to define so, I only kept OBJECTID as an index to the dataframe.

Some other features had similar meaning where one feature like SEVERITYCODE had another feature like SEVERITYDESC to describe it. For example, SEVERITYCODE values were numbers to describe the severitylevel of the accident and SEVERITYDESC values were a text description for the severity level. I only kept one feature from similar features and dropped other features.

Other type of features contained more NaN values than it should be. For example, the SPEEDING feature had only 9333 Y value while 185340 NaN value.

Other features like PERSONCOUNT, PEDCOUNT, INCDATE...etc were not useful to predict the severity of the accidents so, they had to be removed from the data frame.

After dropping all irrelative features or useless data, 10 features remained which are all relative to the target and could be useful to build and train the model.

IV. Predictive Modeling

The best model to predict the severity of an accident is the classification as we are trying to put the severity of an accident in a predefined values which goes from 1 to 5 according to the accident. I used two classification model to predict the target, Logistic Regression and Decision Tree.

Classification Models:

1.Decision Tree:

I used Jaccard score and f1-score to check the model and the results came as in the following figures.

DT Jaccard index: 0.74
DT F1-score: 0.68

	precision	recall	f1-score	support
1	0.73	0.99	0.84	24610
2	0.88	0.19	0.31	11041
micro avg	0.74	0.74	0.74	35651
macro avg	0.81	0.59	0.57	35651
weighted avg	0.78	0.74	0.68	35651

2. Logistic Regression:

I used the same tests as the decision tree model for the logistic regression model and the results came as in the following figures:

DT Jaccard index: 0.74
DT F1-score: 0.68

	precision	recall	f1-score	support
1	0.73	0.98	0.84	24610
2	0.81	0.21	0.33	11041
micro avg	0.74	0.74	0.74	35651
macro avg	0.77	0.59	0.59	35651
weighted avg	0.76	0.74	0.68	35651

V.Conclusion

I analyzed the data from the seattle accidents from 2004 till the present and built two classification models two predict the severity of an accident. These models are useful to warn people about the severity of an accident which could happen at each road under some conditions in order to be careful while driving. These models will help in reducing the number of accidents.