IBM Data Science Capstone Report

Car Accident Severity Analysis

Mayara Monteiro da Silva

September 7th, 2020

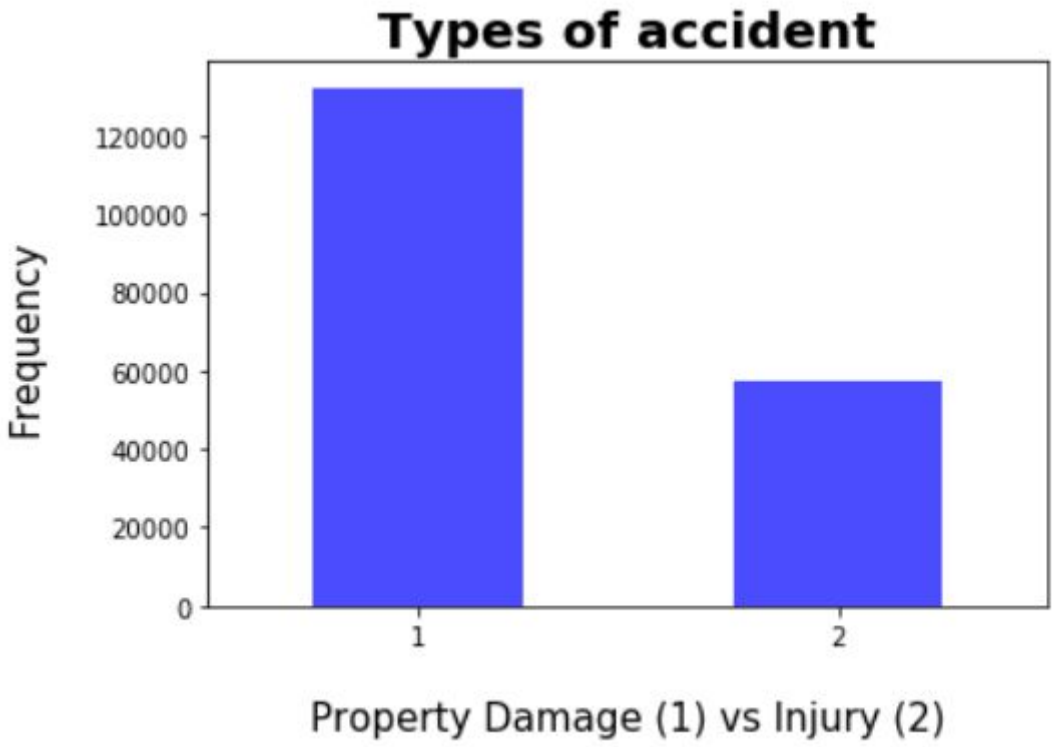
1. Introduction

According to the WHO, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. For children and young adults aged 5-29 years this is the leading cause of death.
Due to the severity of this situation, The2030 Agenda for Sustainable Development has set an ambitious target of halving the global number of deaths and injuries from road traffic crashes by 2020.
By developing an algorithm to predict the severity of an accident given the current weather, road and visibility condition, it will be possible to alert drivers about bad conditions, enabling them to be more careful. Therefore, the frequency of car accidents can be decreased.

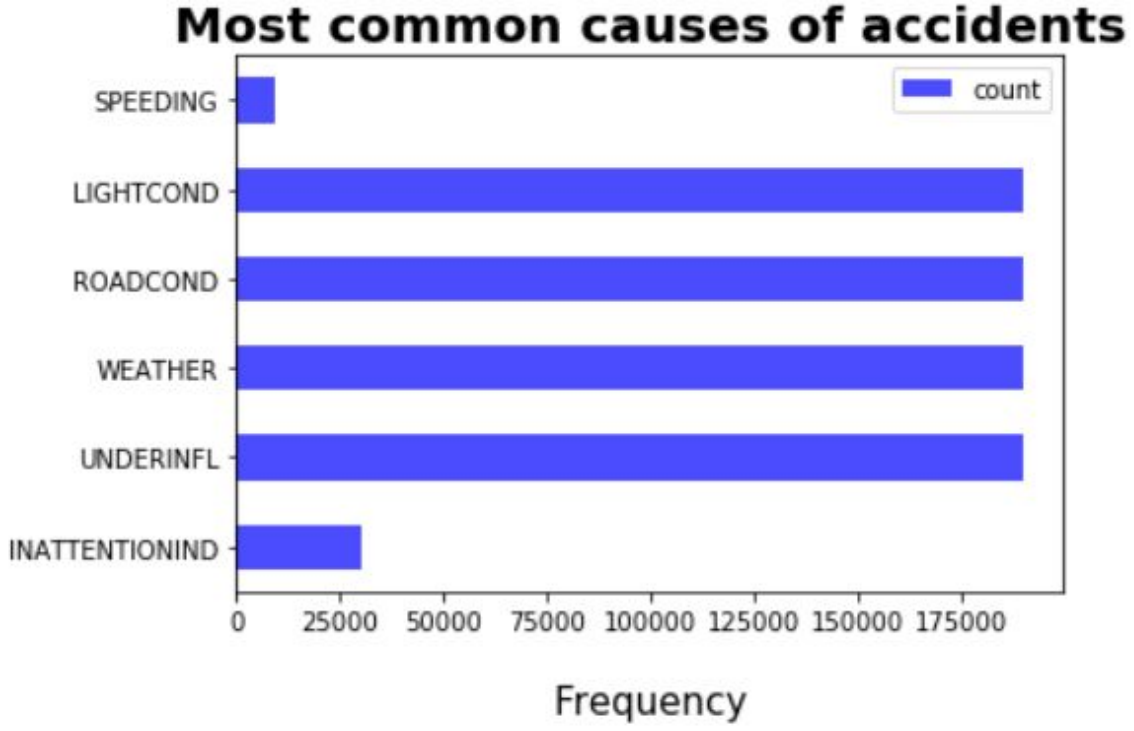
2. Data

The data collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to present consists of 37 independent variables and 194,673 rows.
The variable, “SEVERITYCODE”, classifies the level of severity caused by an accident as:

1: Property Damage
2: Chance of Injury



The dataset provides a lot of different information about the circumstances in which the accidents took place. The ones related to possible causes of accidents are speeding, light conditions, road conditions, weather, the driver being under influence and lack of attention.



After analysing the data, it's noticeable that Light Condition, Road Condition, Weather and Under Influence are the main ones. Alongside with that, it was also noticed the need to prepare the data, since there were many unneeded columns and information missing. To solve this and other problems, the following steps were taken:

- Remove unnecessary columns.
- Create new columns where the variables that are strings will be replaced by numbers
- Drop Nan values
- Balance the dataset (since the class 1 of variable "SEVERITYCODE" is almost three times the size of the class 2).

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHERRCODE	ROADCONDCODE	LIGHTCONDCODE
0	2	Overcast	Wet	Daylight	2	1	0
1	1	Raining	Wet	Dark - Street Lights On	1	1	1
2	1	Overcast	Dry	Daylight	2	0	0
3	1	Clear	Dry	Daylight	0	0	0
4	2	Raining	Wet	Daylight	1	1	0

The prepared dataset

3. Methodology

The purpose of this project was to build Machine Learning models in order to try to predict car accidents. To do that, Github was used as a repository and Jupyter Notebook was used to preprocess data and build the models using Python and some of its packages, such as Pandas, NumPy and Sklearn.
As shown on the Data Section, after loading data into Pandas Dataframe and checking the feature names and their data types, the most important features to predict the severity of accidents were selected and the target variable was balanced. They are:

- WEATHER
- ROADCOND
- LIGHTCOND
- SEVERITYCODE (the target variable).

Once the dataset was ready, it was split into one Train Set and one Test set. The test size was 0.30 and the random state was 42.

The Machine Learning models chosen to solve the problem were:

- K-Nearest Neighbour (KNN)

The value of K used was 18.

- Decision Tree

The criteria was Entropy and the Max Depth was 7.

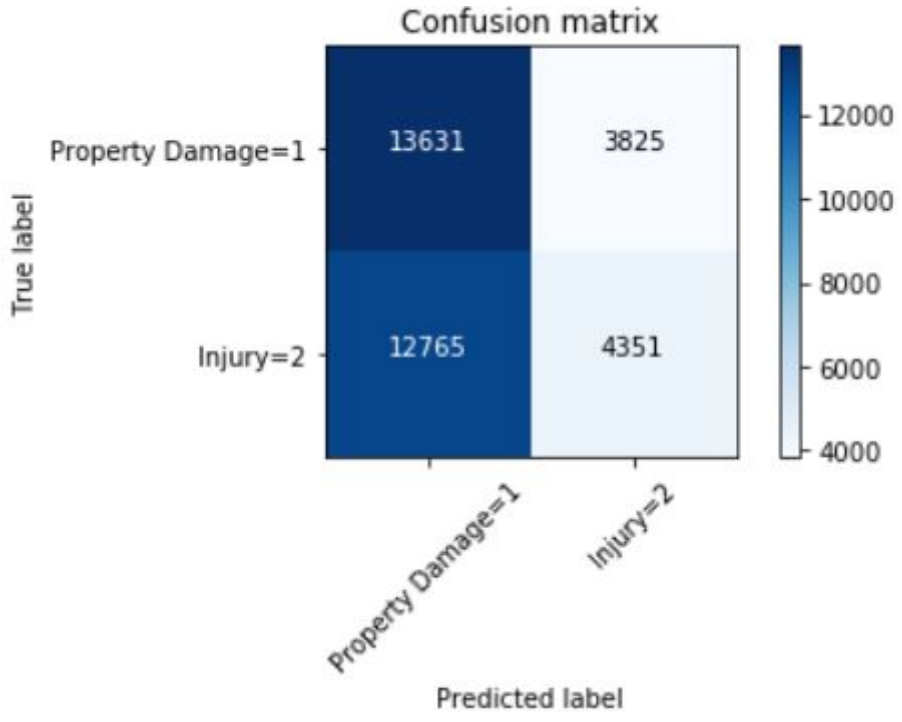
- Logistic Regression

The value of C used was 6 and the solver was liblinear.

To evaluate the results of the Machine Learning models, the Jaccard Score, T1 Score and Accuracy Score were used.

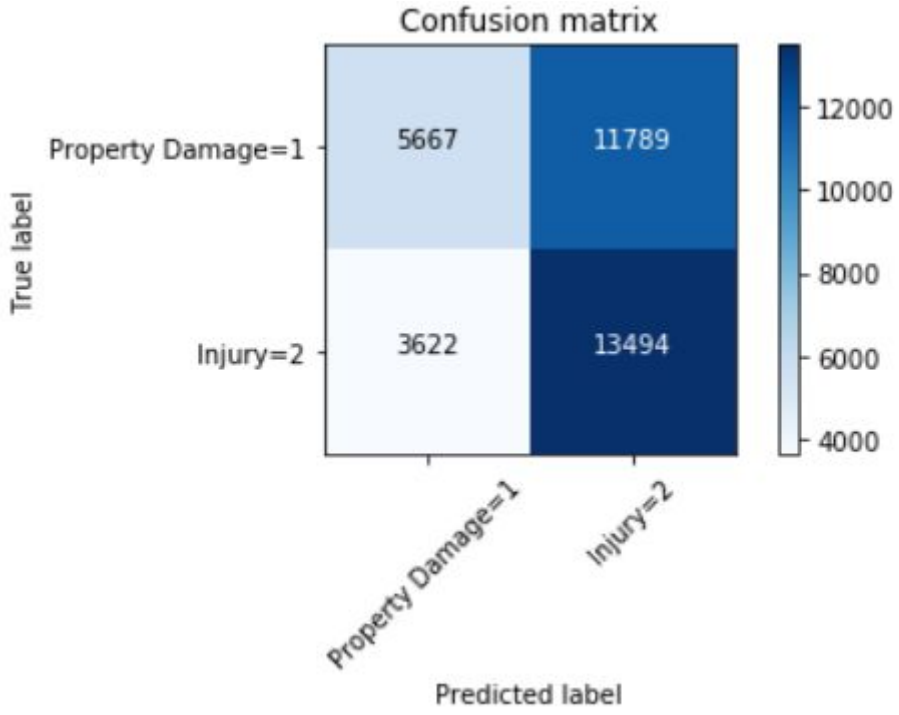
4. Results and Discussion

- K-Nearest Neighbour (KNN)



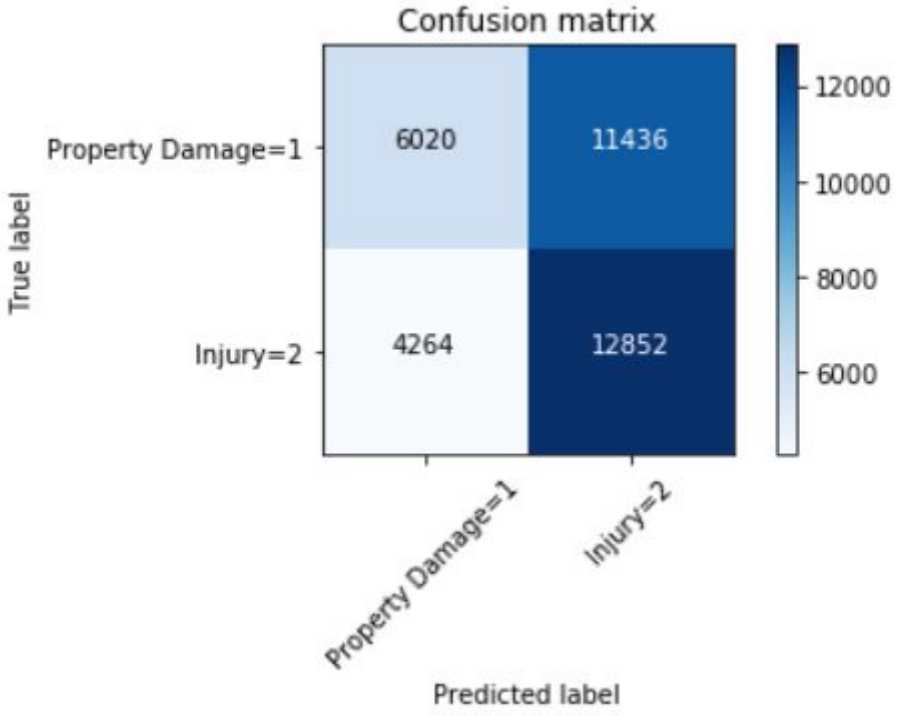
Jaccard Score	T1 Score	Accuracy Score
0.52	0.48	0.52

- Decision Tree



Jaccard Score	T1 Score	Accuracy Score
0.55	0.53	0.55

- Logistic Regression



Jaccard Score	T1 Score	Accuracy Score
0.54	0.52	0.54

Comparing the results:

The final results of the model evaluations can be summarized in the following table:

Jaccard Score	T1 Score	Accuracy Score
0.52	0.48	0.52
0.55	0.53	0.55
0.54	0.52	0.54

5. Conclusion

The careful study of the dataset and the accuracy results, leads to the conclusion that there are multiple factors that can contribute to the severity of a Car accident. The ones chosen for this project can help predict the level severity, but it’s important to keep in mind that no matter what the conditions are, being careful while driving is imprescindible.
After analysing the Jaccard Score, F1 Score and Accuracy score for the three Machine Learning models developed, it's noticeable that the Decision Tree is the best option.