Report: Predicting accident severity in Seattle

IBM Capstone Project – Applied Data Science

C. Cotton
October 4, 2020

Table of Contents

Introduction:

The number of accidents has been rising globally due to increases in population and transportation. Traffic accidents are a daily source of death, injury and property damage on roads and highways resulting in huge losses at economic and social levels.
According to the World Health Organisation (WHO), in 2018[1] around 1.35 million died as a result of a traffic accident.

If we look more precisely at the figures for Seattle, in 2017[2], there were 10,959 police reported collisions and a further 1516 self-reported ones.
Fatalities were 19, and serious injuries 168.

Business Understanding:

As the demand for vehicles rises, the number of vehicles on the road and traffic jams increase, especially during rush hours.
The local government of Seattle is trying to implement measures to alert vehicle users, police, traffic and health systems about critical situations – to reduce the number of accidents on the road.
I will attempt to build a model to predict the severity of an accident given the weather, road conditions and location.
By doing this, we should be able to predict the severity of an accident.

This will be useful to the traffic department also to implement measures to decrease the risk of an accident.

Sources:
1. WHO Global status report on road safety 2018: https://apps.who.int/iris/bitstream/handle/10665/277370/WHO-NMH-NVI-18.20-eng.pdf?ua=1

2. Seattle Department of Transport 2018 Traffic Report:
https://www.seattle.gov/Documents/Departments/SDOT/About/DocumentLibrary/Reports/2018_Traffic_Report.pdf

Data source & description:

The data used in this study was provided by Coursera, and had records on all types of collisions from 2004 through to present day.

The data is regarding the **severity of each accident** along with the time and conditions under which each accident occurred.

The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury) which were encoded to the form of *0* (Property Damage Only) and *1* (Physical Injury). Following that, *0* was assigned to the element of each variable which can be the least probable cause of severe accident whereas a high number represented adverse condition which can lead to a higher accident severity.
As there are null values in some records, the data needs to be pre-processed before any further processing.

The data will be analysed to predict the likelihood of a collision using 3 machine learning modules and it's severity given a number of variables such as weather conditions, road conditions, time of day, period of year, etc.
These predictions will help the Department of Transport of Seattle to implement preventive measures as required.

A quick look at the map of Seattle, shows that the majority of accidents occur downtown.

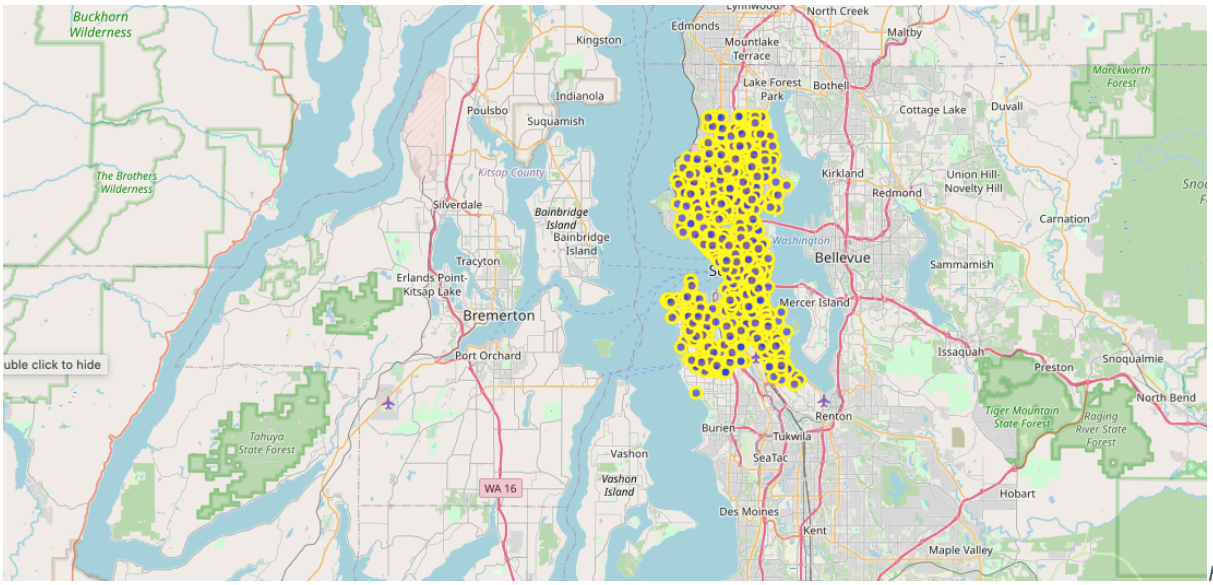

Figure 1: Map showing the 1st 1000 collisions

Methodology:

1. Data evaluation

The original data set had 37 columns and 194.673 rows of entries. The columns are (with the count of null values):

	NaN Count
SEVERITYCODE	0
X	5334
Y	5334
OBJECTID	0
INCKEY	0
COLDETKEY	0
REPORTNO	0
STATUS	0
ADDRTYPE	1926
INTKEY	129603
LOCATION	2677
EXCEPTRSNCODE	109862
EXCEPTRSNDESC	189035
SEVERITYCODE.1	0
SEVERITYDESC	0
COLLISIONTYPE	4904
PERSONCOUNT	0
PEDCOUNT	0
PEDCYLCOUNT	0
VEHCOUNT	0
INCDATE	0
INCDTTM	0
JUNCTIONTYPE	6329
SDOT_COLCODE	0
SDOT_COLDESC	0
INATTENTIONIND	164868
UNDERINFL	4884
WEATHER	5081
ROADCOND	5012
LIGHTCOND	5170
PEDROWNOTGRNT	190006
SDOTCOLNUM	79737
SPEEDING	185340
ST_COLCODE	18
ST_COLDESC	4904
SEGLANEKEY	0
CROSSWALKKEY	0
HITPARKEDCAR	0

We can drop those columns that have more than 3% of null values as well as those that are not relevant to our study:
- INTKEY
- EXCEPTRSNCODE
- EXCEPTRSNDESC
- INATTENTIONIND
- PEDROWNOTGRNT
- SDOTCOLNUM
- SPEEDING
- INCDATE

2. Data Exploration

As we want to clean the data, we can start by looking at the “critical” variables:
- weather
- light conditions
- road conditions
- location of accident

Number of collisions

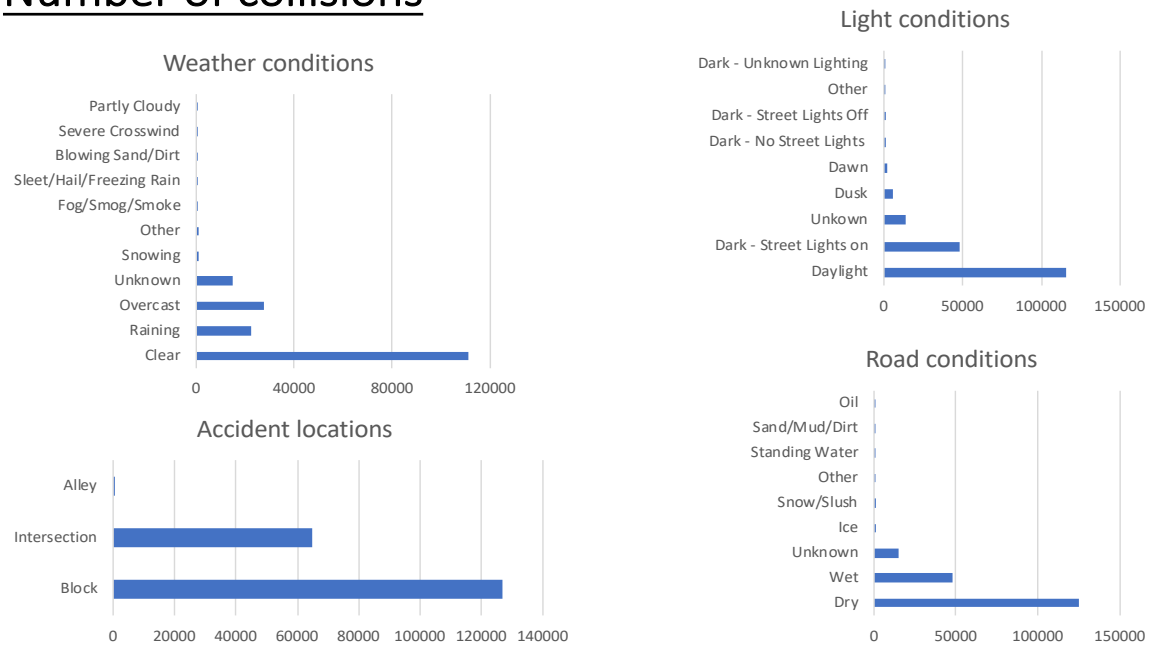

Figure 2: Accidents analysed with different variables

The majority of accidents did not involve injuries, but there were some that did.

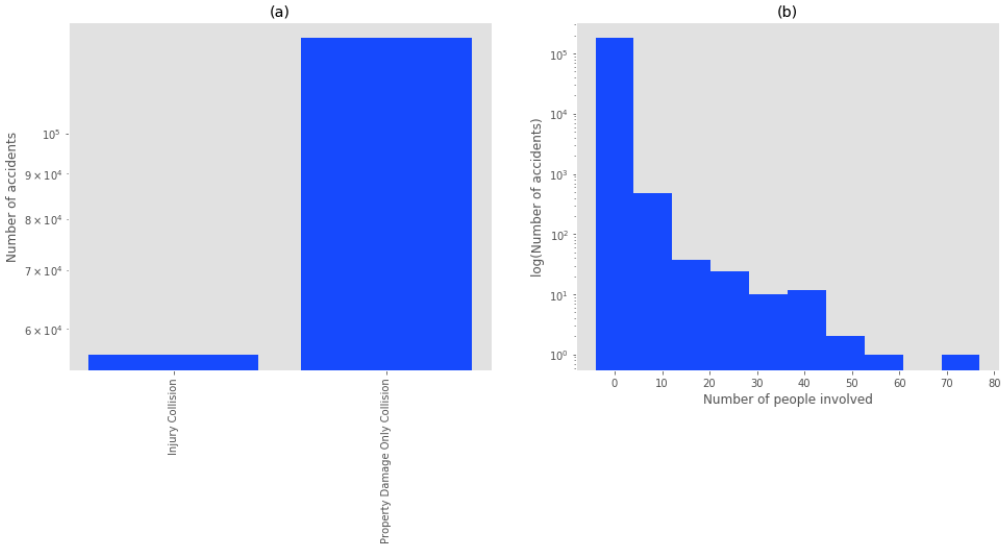

Figure 3: Type of collision & number of people involved

We can see that Fridays were the day with the most accidents, and October was the month.

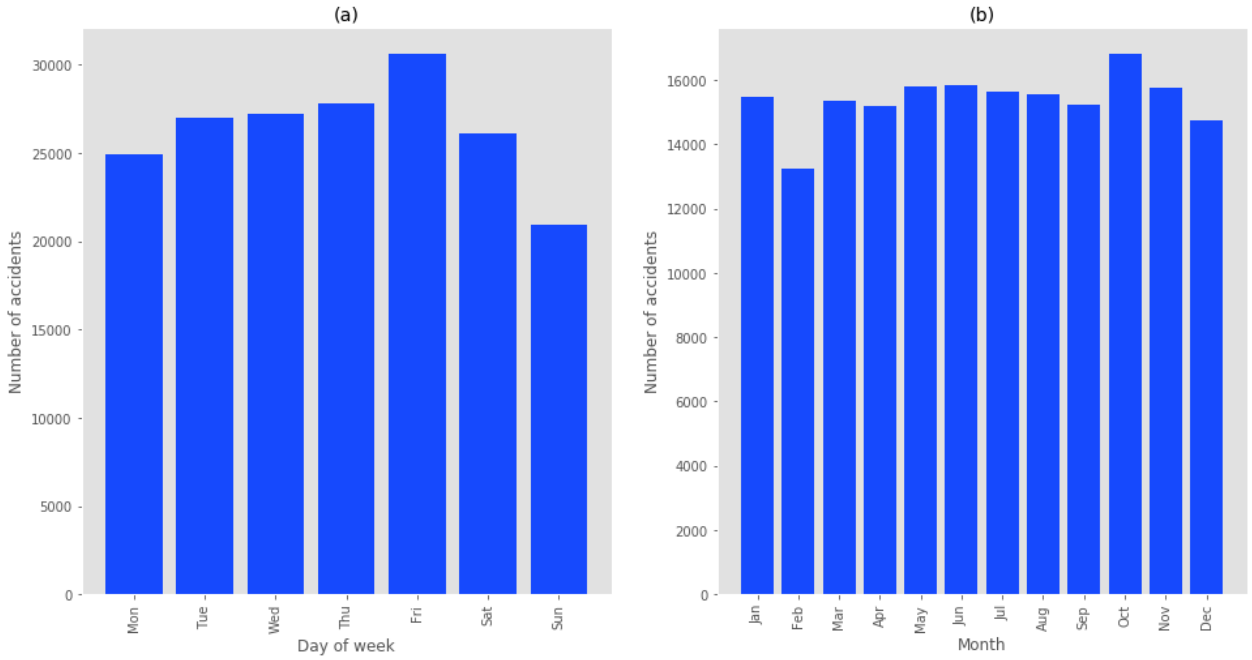

Figure 4: When the collisions took place

3. Data Pre-processing

The dataset providing the Seattle collisions information from 2004 – present day are not suitable for quantative data analysis in its original form, as:

- Data is incomplete. Around 15% of recorded accidents are missing one or more key features, including in some cases the target variable (accident severity code). These needed to be removed from the dataset.

- Dataset includes redundant columns. These also were removed.

- Data are imbalanced and non-standardised.

4. Model Development and Evaluation

To develop a model for predicting accident severity, the re-sampled and cleaned dataset was split into testing and training sets (containing 30% and 70% of the samples). By using the scikit learn "train_test_split" method, 3 models were trained and evaluated.

- Decision Tree model
- K-Nearest Neighbors (kNN) model
- Logistic Regression model

It was found that the Decision Tree produced the highest accuracy.

	Algorithm	Jaccard	F1-score	Precision
0	KNN	0.73	0.68	0.71
1	Logistic Regression	0.7	0.58	0.67
2	Decision Tree	0.75	0.69	0.77

Conclusions:

This work highlights that machine learning techniques can be applied to historical data to be able to make reliable predictions about the outcome of road traffic accidents.

The model can be extended to include new features so that city planners can gain insight into the conditions associated with high accident severity and use this to improve road or traffic flow design.