IBM Coursera Capstone Project On Applied Data Science

Nandan Choudhary

12th October 2020

1. Introduction:

1.1 Context

This data set is about accident's (car collisions) severity. This data includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. The data dates weekly from 2004 to present. The data has been collected from the Seattle Department of Transportation. All collisions provided by SPD and recorded by Traffic Records. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. For the data analysis I will look into road conditions, light conditions, address (where the accident took place), weather conditions. For deeper analysis I will use KNN, SVM, LR and Decision Tree algorithms.

1.2. Problem Statement/ Business Problem:

Intuitively, we might expect that some of the factors which influence the likelihood and severity of a road traffic accident include: the weather, road conditions(good or bad), time of day (and the presence or absence of street lights), and the number and type of vehicles in the area. Additional factors which may influence the severity of road traffic accidents include those which can be traced to individual negligence, such as driving under the influence of alcohol/narcotics, driving without due care and attention or driving at excessive speed. While it is intuitive that a combination of these factors may be important, intuition alone cannot determine the relative significance of these factors. Determining the relative significance of these competing factors is necessary if we are to fully understand the causes of road traffic accidents and devise new strategies to minimise their occurrence and severity.

1.3. Target Audience

The main target audiences for this work will be road/city planners and emergency service responders.
And also Police Departments, Common people, World Road Association and others.

2. Data

2.1 Data Acquisition

Data were obtained for all road traffic accidents recorded in the Seattle municipal area between Jan 2004–Aug 2020 by the Seattle Department of Transport (SDOT). The data were obtained from the Seattle Open Data Portal in Comma Separated Value (CSV) format and read into a Pandas Dataframe for analysis using the Pandas read_csv function. In total, the dataset comprises 221,006 rows (one for each road traffic accident in Seattle during this period) and 40 providing information about the accident, including the accident severitycode (i.e. the target variable for this analysis). Further information about the properties of the dataset (and the analysis thereof) is available in an online Jupyter Notebook, however a list of the columns present in the dataset is shown in Table 1 along with a brief description of each column's meaning. A full glossary of headings in the table 2 is available online at SDOT. Furthermore, to illustrate the format in which the raw data are recorded, the first row of the table is shown in Table 2. The target/dependent variable is severitycode which, in its original form, takes the values 0, 1, 2, 2b or 3. The definitions of these severity codes are provided in the "Attribute Information" metadata which accompany the data release and are as follows:

- 0: Unknown
- 1: Property/vehicular damage
- 2: Minor injury
- 2b: Serious injury
- 3: Fatality

2.2 Data Cleaning

In its original form, this dataset is not suitable for quantitative analysis. There are five principal reasons for this, which are explained in the following subsections.

Missing target data

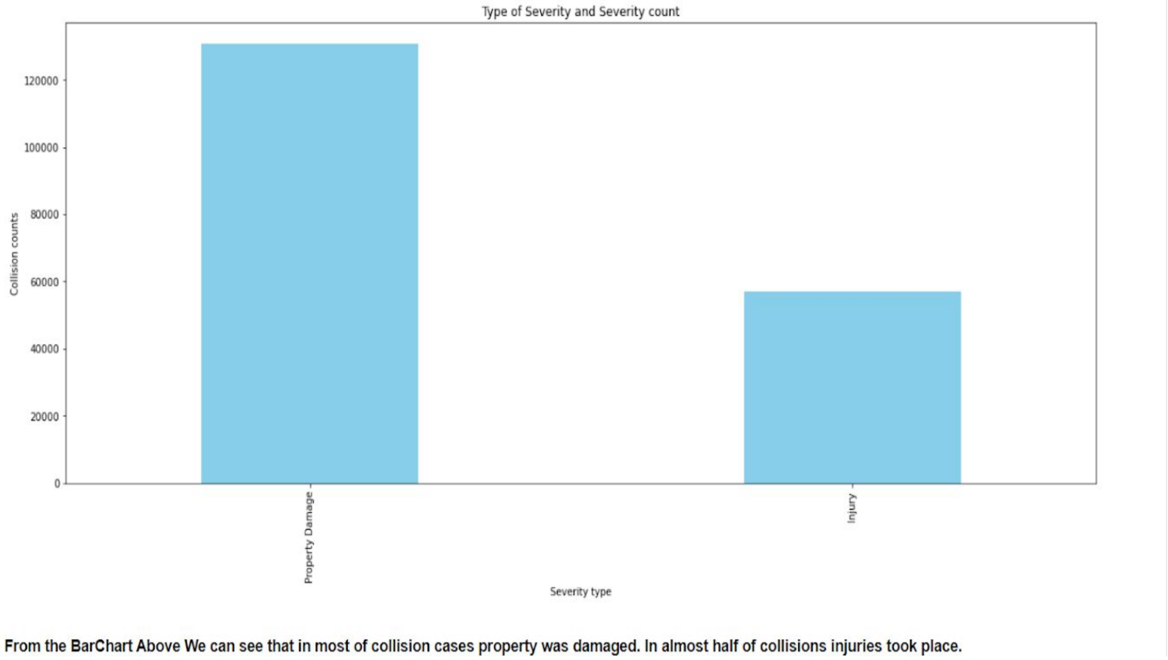
The target variable is severitycode, however nearly 20,000 accidents (9.8% of the dataset) are missing this vital information. As the purpose of the model is to predict the severity of an accident from the other features in the dataset, clearly accidents with no value for this vital variable are of no use to us, and need to be excluded.

Most important part of data analysis is data cleaning and understanding. Severity of the accidents was segmented mainly into 2 groups:
1) collisions which only involved property damage;
2) collisions which involved injury.
All rows that contained not accurate data, or data that would not help us, was discarded. For example, values that were named „Unknown“, missing values.

3. Exploratory Data Analysis

3.1. Relationship between severity types

From Figure 1 we can clearly see that in most collision cases property was damaged. Property damage was in 136485 collisions, injuries were in 58188 collisions. That clearly illustrates Figure 1.



From the BarChart Above We can see that in most of collision cases property was damaged. In almost half of collisions injuries took place.

3.2. Relationship between address and collision count

From Figure 2 we can see that most of the accidents occured in the blocks, less accidents occurred at the intersections. Least accidents occurred at the alley.

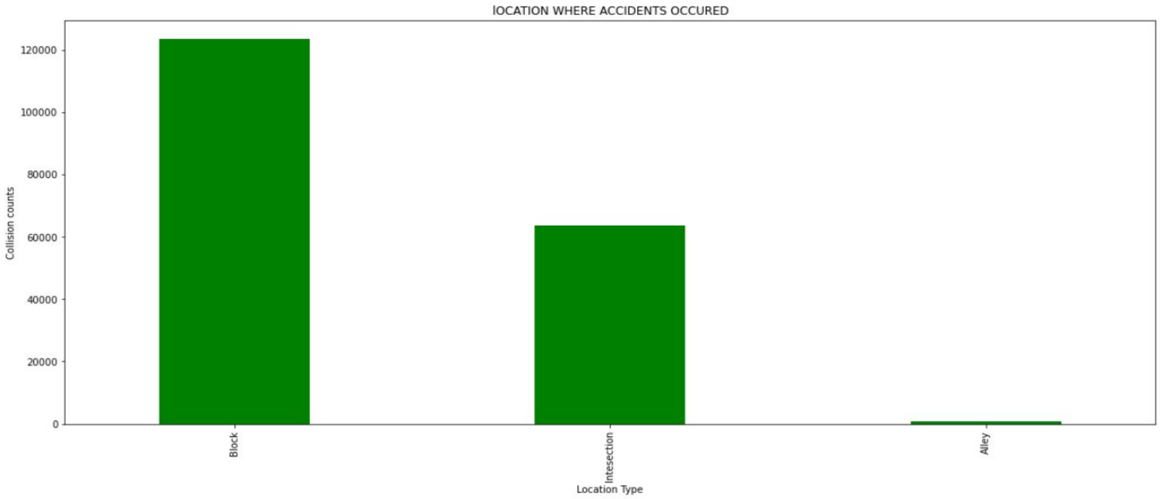


Table 1: Values of Collision on type of Location

Block	107780
Intersection	61406
Alley	595

3.3. Relationship between weather conditions and collision count

From the Fig.3 we can see that most of the accidents occured in the good weather conditions. (Most accidents happened in the 'clear' weather condition.)

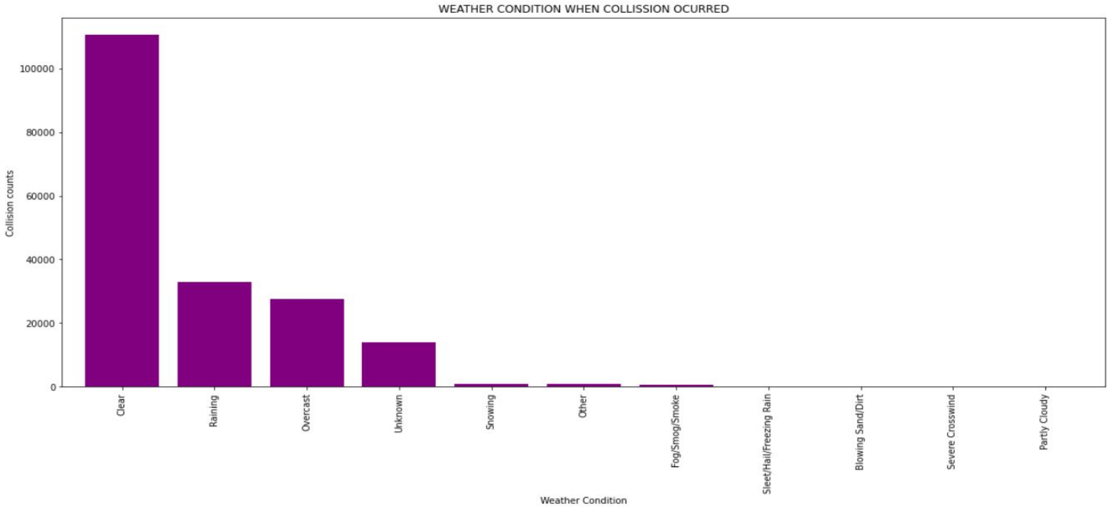


Table 2: Values of accidents in some weather condition

Clear	108507
Raining	32599
Overcast	26863
Snowing	827
Fog/Smog/Smoke	549
Other	253
Sleet/Hail/Freezing Rain	110
Blowing Sand/Dirt	43
Severe Crosswind	25
Partly Cloudy	5

3.4. Road condition influence to accidents

In Figure 4 we can see that most collisions happen in dry road conditions.

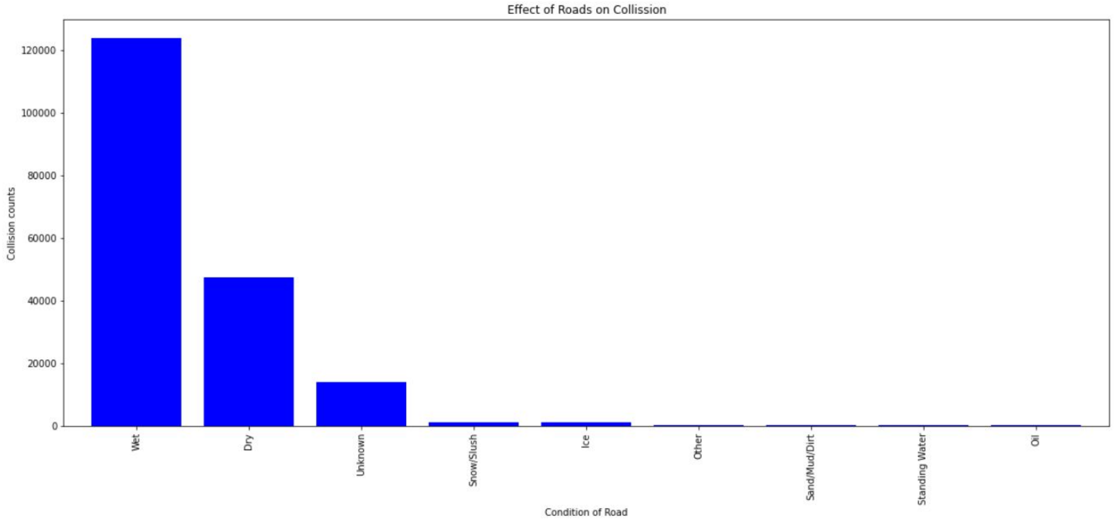


Table 3: Counts of Collision Based on Road Condition

Dry	123736
Wet	47223
Unknown	14009
Ice	1193
Snow/Slush	992
Other	124
Standing Water	111
Sand/Mud/Dirt	73
Oil	64

3.5. Light Conditions influence to accidents

From Figure 5. we can see that most accidents happened in the daylight. That can happen because most of the traffic happens also in the daylight. People go to jobs, schools and so on.

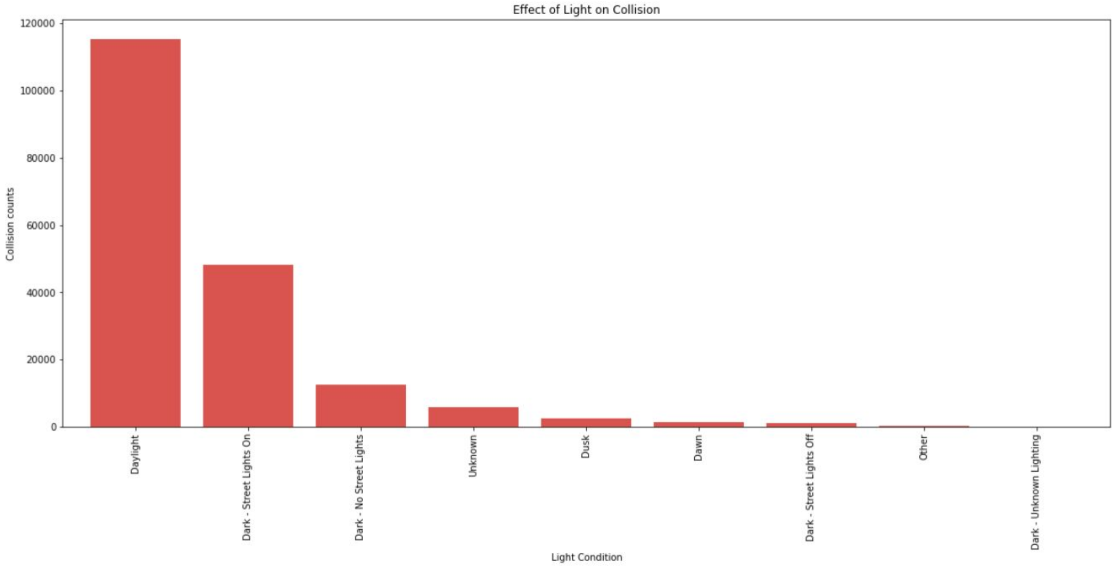


Table 4: Counts of Collision based on Light Condition

Daylight	115408
Dark - Street Lights On	48236
Unknown	12599
Dusk	5843
Dawn	2491
Dark - No Street Lights	1526
Dark - Street Lights Off	1184
Other	227
Dark - Unknown Lighting	11

3.6. Light Conditions influence to accidents

Majority of the accidents took place in daylight (property damage and injuries) and in dark (with street lights on). This may conclude that most accidents happen not because of daytime, but because of that people are unaware of the situation, not paying full attention.

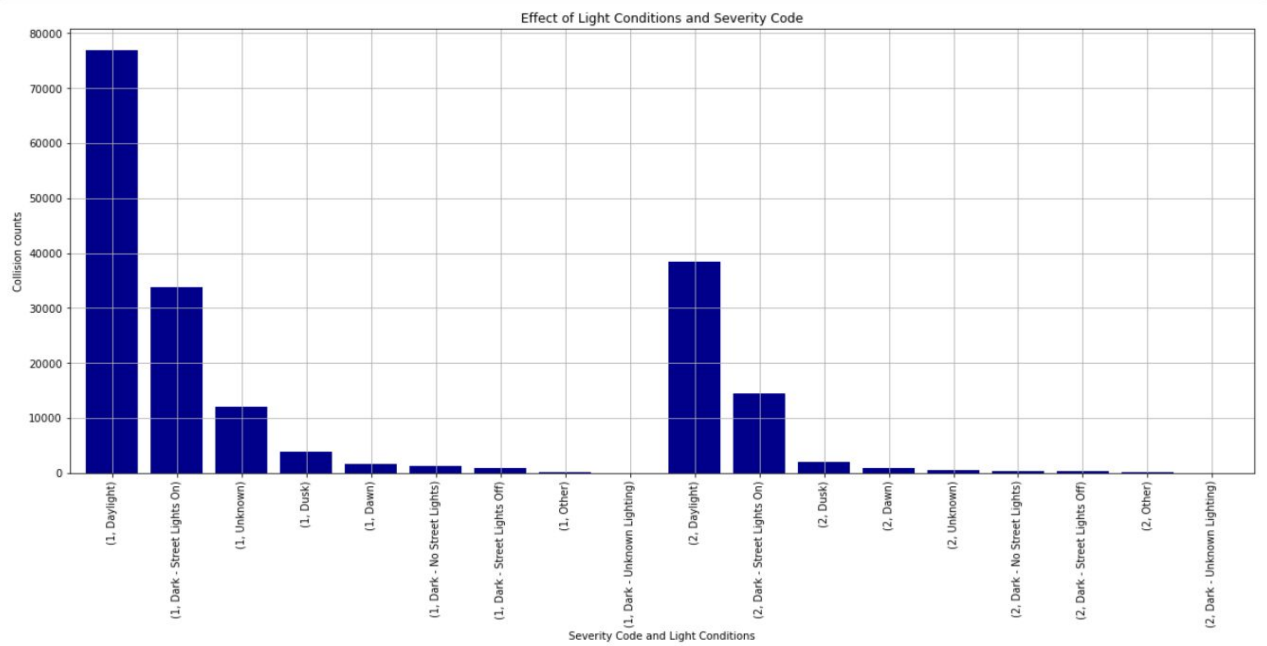


Table 5 Effect of Severity code and Light Condtion On Collision

SEVERITYCODE	LIGHTCOND	Counts
1	Daylight	76998
	Dark - Street Lights On	33816
	Unknown	12010
	Dusk	3907
	Dawn	1668
	Dark - No Street Lights	1192
	Dark - Street Lights Off	869
	Other	175
	Dark - Unknown Lighting	7
2	Daylight	38410
	Dark - Street Lights On	14420
	Dusk	1936
	Dawn	823
	Unknown	589
	Dark - No Street Lights	334
	Dark - Street Lights Off	315
	Other	52
	Dark - Unknown Lighting	4

3.7. Road Conditions influence to accidents

Most car accidents where property was damaged or people got injured, took place in dry road conditions.

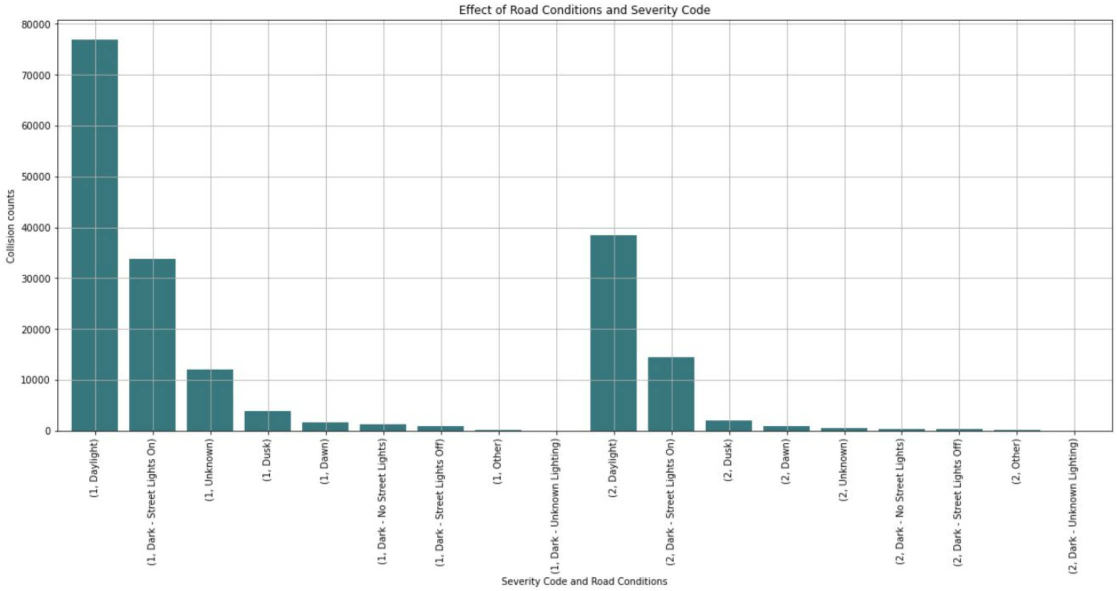


Table 6:Effect of Road condition and Severity code on Collision

SEVERITYCODE	ROADCOND	Counts
1	Dry	83835
	Wet	31523
	Unknown	13279
	Ice	923
	Snow/Slush	827
	Other	82
	Standing Water	82
	Sand/Mud/Dirt	51
	Oil	40
2	Dry	39901
	Wet	15700
	Unknown	730
	Ice	270
	Snow/Slush	165
	Other	42
	Standing Water	29
	Oil	24
	Sand/Mud/Dirt	22

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audiences might prefer one over the other.

I applied Linear Regression, Support Vector Machine, K –Nearest Number, Decision Tree models.

Table 7: Accuracy of different Classifiers.

	KNN Model	Decision Tree	SVM Model	Logistic Regression
Training Accuracy	0.6695	0.6970	0.6969	0.6957
Testing Accuracy	0.6700	0.6959	0.6958	0.6949

5. Conclusion

Most of accidents happened in the daylight, dry road conditions. That concludes that most accidents happen because of human influence (not paying attention, lack of sleep and so on).
Data was cleaned and prepared for data analysis and model building.

The four models we built are all very similar in terms of prediction and accuracy. The highest prediction accuracy is about 69.69%.

Most accurate models were "Support Vector Machine", "Logistic Regression" and "Decision Tree".

In this project, we have found the major environmental factors and road conditions that affect car accidents. Also we found a building a model that can help predict the severity of car accidents based on these conditions. Based on the data analysis and results, we can make some recommendations to improve the safety of drivers, pedestrians and others. Most helpful advice would be to pay attention and to watch the signs.