Ng Wen Jie

IBM Applied Data Science Capstone

27 Sep 2020

Predicting the Severity of a Car Accident

1. Introduction

1.1 Background

Seattle, one of the fastest-growing major city in United States (U.S.), is a seaport city situated on the West Coast of U.S. The metro area population of Seattle has been on the rise, increasing by 1.2% from about 3 339 000 in 2017 to 3 379 000 in 2018, and increasing by 0.8% to 3 406 000 in 2019.[1] As population increases, the amount of vehicles on the roads also rises. In 2017, the number of cars in Seattle reached 435 000, which is just over two thousand per square kilometres.[2] With more cars, the probability of car accidents increases as well. Car accidents generally happen about once a day in the transport corridors of Seattle.[3] Major car accidents almost always lead to huge traffic jams, resulting in commuters being late to their destinations. As such, it could be advantageous for drivers to be able to predict the severity of a car accident, so that they could avoid a potential massive traffic jam by changing their driving routes beforehand.

[1] (*Seattle Metro Area Population 1950-2020*)
[2] (Balk, *Booming Seattle is adding cars just as fast as people* 2017)
[3] (Baruchman, *Seattle traffic deaths and injuries down slightly last year; most of the fatalities were pedestrians* 2019)

1.2 Problem

Data that may contribute to the severity of a car accident may include the number of people involved, the conditions of the road such as light and wetness, the weather, and the type of road at which the accident occurred. This project aims to predict how severe a car accident is, depending on these data.

1.3 Interest

Drivers would be interested in accurate predictions of the severity of accidents, to minimise wastage of time and fuel if they were to be stuck in a traffic jam.

2. Data Acquisition and Cleaning

2.1 Data Sources

The dataset, consisting of records of past accidents from January 2004 to April 2020 and circumstances surrounding them, such as the collision type, locations and various conditions, can be found here.

2.2 Data Cleaning

Generally, the data set is rather complete with only a few features having missing values.

The oldest records compiled from 2004 had mostly intact information with only a few entries having missing values and hence there is no need to restrict the dataset by the year of the accident.

Based on the unique and secondary keys for the incidents, it seems that there were no duplicates of entry present in the dataset as well.

Some features had values such as ‘Unknown’ and ‘Others’ which do not hold much significance in modelling, due to them not describing much about the circumstances of the accidents. As such, entries holding such values are dropped.

2.3 Feature Selection

After going through some data cleaning, the dataset had 145 363 total entries and 38 features. Upon closer examination of the data, it was observed that some features were very similar and can be considered repetitions of each other. Some features also had too many missing values to be considered meaningful data.

Some of the features that are very closely linked would be those features which are pre-defined descriptions for a certain code. For instance, the feature for the code corresponding to the severity of the collision (SEVERITYCODE) and the feature for the detailed description of the collision are essentially the same (SEVERITYDESC). For easy data reading, the description counterpart of each pair will be dropped (see table 1).

Features with too many missing data can have reduced effectiveness in analysis and modelling. Features such as the key that corresponds to the intersection associated with a collision (INTKEY) and whether or not speeding was a factor in the collision (SPEEDING) had 129 603 and 185 340 missing values respectively. As a general guide, features with more than 10% of the total entries having missing values will not be selected (see table 1).

Table 1

Simple Feature Selection during Data Cleaning

Kept Features	Dropped Features	Reason for Dropping Features
EXCEPTRSNCODE, SEVERITYCODE, SDOT_COLCODE, ST_COLCODE	EXCEPTRSNDESC, SEVERITY.CODE1, SEVERITYDESC, SDOT_COLDESC, ST_COLDESC	Two similar features (one is description of the code in another feature)
All other features	INTKEY, EXCEPTRSNCODE, EXCEPTRSNDESC, INATTENTIONIND, SDOT_COLUMN, SPEEDING	Too many missing values (more than 10% missing)

3. Methodology

In this project, the target or dependent variable will be the feature named ‘SEVERITYCODE’, which is already present in the dataset. The data present in the dataset, known as the actual value, will be used to compare to the predicted values generated by the model.

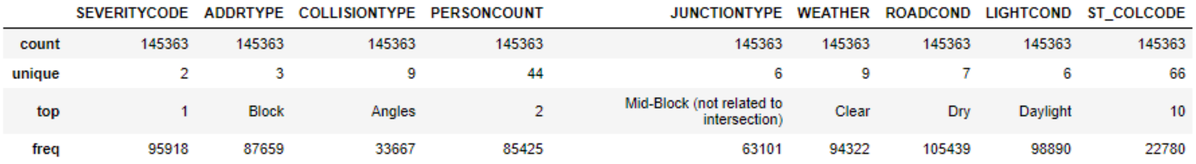
As for the independent variables to be input, I will first determine to correlation of selected features to the dependent variable. After which, strongly correlated features will be used for modelling.

Finally, classification models such as k-Nearest Neighbours, Decision Trees, Support Vector Machines and Logistic Regression will be built and their effectiveness on predicting the severity of an accident will be assessed. Metrics used to assess the models are Jaccard index, F1-score and Logarithmic Loss (if applicable).

4. Results

4.1 Exploratory Data Analysis

First, we look at the overview of the dataset (see figure 1) after data cleaning, displaying the count, number of unique values, and the highest frequency of values of each feature.



	SEVERITYCODE	ADDRTYPE	COLLISIONTYPE	PERSONCOUNT	JUNCTIONTYPE	WEATHER	ROADCOND	LIGHTCOND	ST_COLCODE
count	145363	145363	145363	145363	145363	145363	145363	145363	145363
unique	2	3	9	44	6	9	7	6	66
top	1	Block	Angles	2	Mid-Block (not related to intersection)	Clear	Dry	Daylight	10
freq	95918	87659	33667	85425	63101	94322	105439	98890	22780

Figure 1. Overview of Dataset after Data Cleaning

4.1.1 Relationship between Collision Address Type and Accident Counts

It can be said that the type of address (Intersection, Block, Alley) where the accident occurs is rather closely related, since the type of address corresponds to how important and frequently used the road is. This means that on critical roads, if accidents were to happen, more drivers are likely to be affected compared to roads that are less often used. Looking at the plot below (see figure 2), it can be inferred that most accidents occur at Blocks, followed by Intersections then Alleys, but the amount of Injury Collisions (Severity Code 2) occurred at Intersections are slightly more than that at Blocks. It can also be inferred that a higher proportion of Injury Collisions (Severity Code 2) occur at Intersections compared to Property Damage Only Collisions (Severity Code 1).

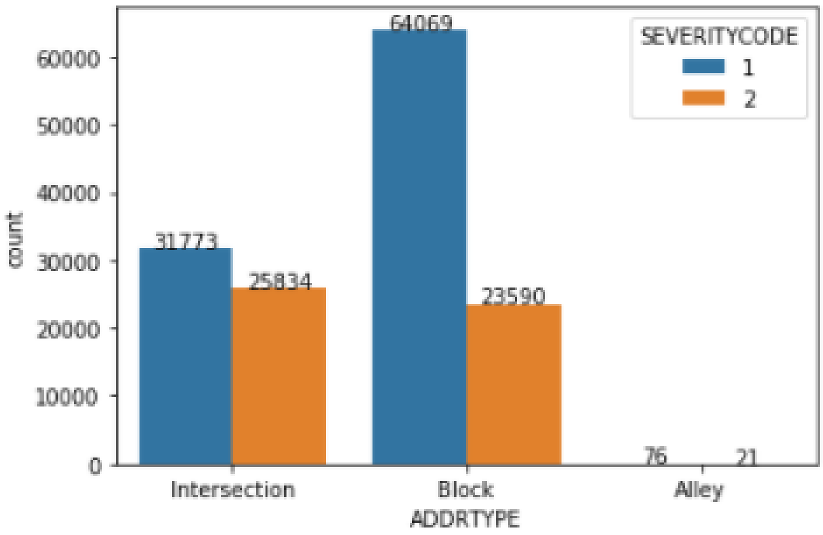
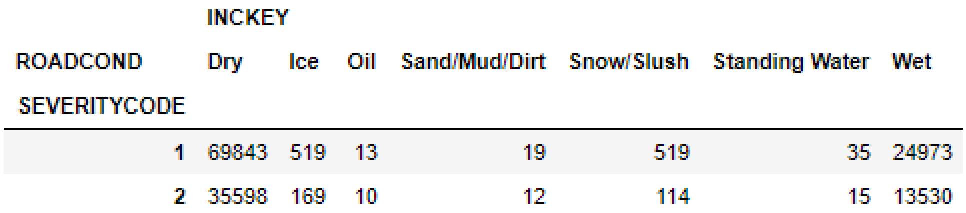


Figure 2. Bar graph of Address Type against Accident Counts

4.1.2 Relationship between Road Conditions and Accident Counts

It is widely accepted that the conditions of the road is closely related with the risk of an accident. For instance, when roads are slippery from rain, drivers tend to drive slower and more carefully as the braking distance of vehicles increases in such road conditions. However, the data seems to disagree that slippery road conditions suggest higher accident counts. From the figure below (see figure 3), it can be observed that most accidents occur when the road is dry as compared to other road conditions.

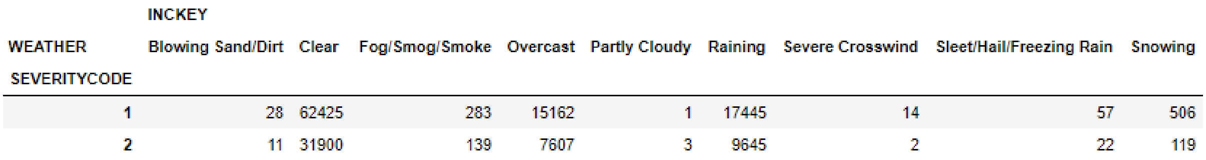


	INCKEY						
ROADCOND	Dry	Ice	Oil	Sand/Mud/Dirt	Snow/Slush	Standing Water	Wet
SEVERITYCODE							
1	69843	519	13	19	519	35	24973
2	35598	169	10	12	114	15	13530

Figure 3: Pivot Table of Road Conditions against Accident Counts

4.1.3 Relationship between Weather Conditions and Accident Count

It is generally accepted that during inclement weather, the risk of accidents while driving increases. This could be due to reduced visibility of other vehicles near the driver, or even slippery roads depending on the weather conditions. However, the data (see figure 4) seems to disagree as it is observed that majority of accidents occur during clear weather. There also seems to be no clear indication of which weather condition will result in more severe accidents.

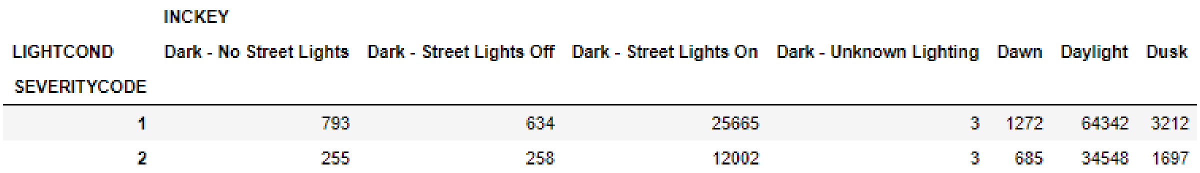


	INCKEY								
WEATHER	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing
SEVERITYCODE									
1	28	62425	283	15162	1	17445	14	57	506
2	11	31900	139	7607	3	9645	2	22	119

Figure 4: Pivot Table of Weather Conditions against Accident Counts

4.1.4 Relationship between Light Conditions and Accident Count

It is also widely agreed that poor light conditions can easily result in higher risks of accidents as the drivers may not be able to see surrounding vehicles due to poor visibility. Similar to Weather Conditions, the data (see figure 5) also disagrees with the above statement. It is observed that a lot more accidents occur in environments of good lighting – such as Daylight and Dark with Street Lights On. It is also observed to be difficult to determine which light condition results in more severe accidents.



	INCKEY						
LIGHTCOND	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk
SEVERITYCODE							
1	793	634	25665	3	1272	64342	3212
2	255	258	12002	3	685	34548	1697

Figure 5: Pivot Table of Light Conditions against Accident Counts

4.1.5 Relationship between Junction Types and Accident Count

It can be hypothesised that junction types that are more frequently used will have higher accident counts. The data (see figure 6) seems to suggest that accidents occur at Intersections and Mid-Blocks more frequently compared to the other junction types. It can also be seen that accidents occurred at Intersections and Intersection-related Mid-Blocks are usually more severe than accidents at other junction types.

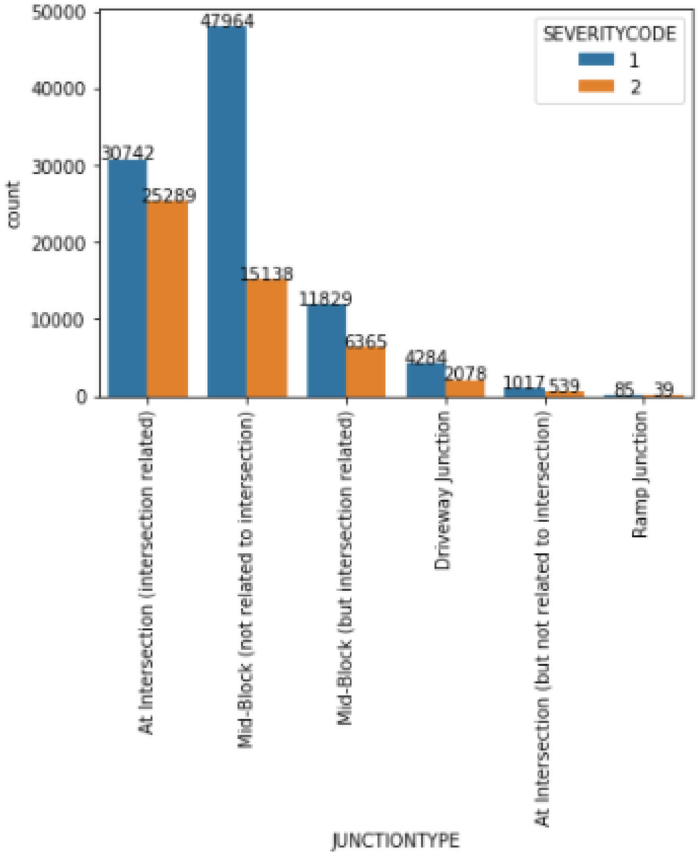
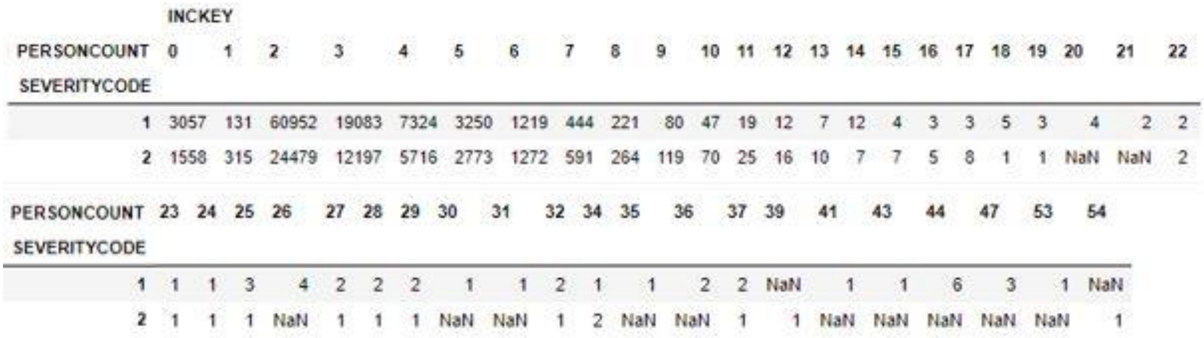


Figure 6: Count Plot of Junction Types against Accident Counts

4.1.6 Relationship between Number of People Involved and Accident Count

The data (see figure 7) suggests that the accidents usually involve only a few people and accidents involving 1 person are likely to be more severe. However, there seems to be no concrete correlation between the number of people involved and severity of accidents, possibly due to accidents involving many people being rare and uncommon.



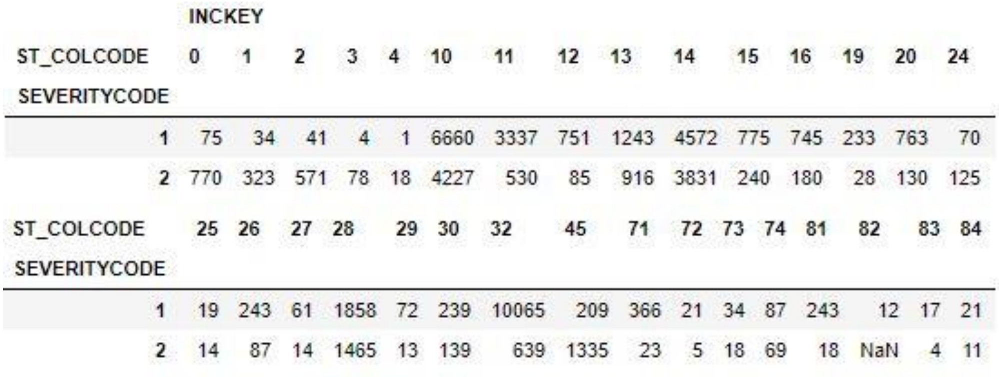
	INCKEY																						
PERSONCOUNT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
SEVERITYCODE																							
1	3057	131	60952	19083	7324	3250	1219	444	221	80	47	19	12	7	12	4	3	3	5	3	4	2	2
2	1558	315	24479	12197	5716	2773	1272	591	264	119	70	25	16	10	7	7	5	8	1	1	NaN	NaN	2

PERSONCOUNT	23	24	25	26	27	28	29	30	31	32	34	35	36	37	39	41	43	44	47	53	54
SEVERITYCODE																					
1	1	1	3	4	2	2	2	1	1	2	1	1	2	2	NaN	1	1	6	3	1	NaN
2	1	1	1	NaN	1	1	1	NaN	NaN	1	2	NaN	NaN	1	1	NaN	NaN	NaN	NaN	NaN	1

Figure 7: Pivot Table of Number of People Involved against Accident Count

4.1.7 Relationship between State Collision Code Dictionary and Accident Count

The State Collision Code Dictionary (ST_COLCODE) can be found here. From the data, it seems to suggest that accidents involving pedestrians (Code 0 to 4) are often more severe than not. The opposite is true to accidents involving vehicles facing the same direction (Code 10 to 23, 71 to 74 and 81 to 84). For accidents involving cars facing each other, severity of accidents is highly dependent on the specific circumstances of the accident. For instance, accidents where both cars moving toward each other colliding head on tend to be more severe as compared to accidents where one car was turning left and the other was going straight from opposite directions. Some of the other Codes not shown describes rare accidents.



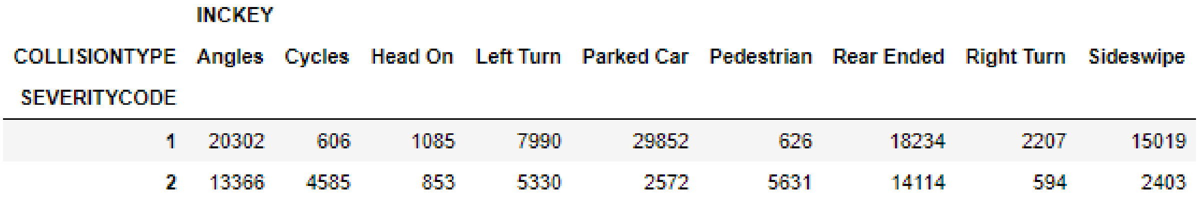
	INCKEY														
ST_COLCODE	0	1	2	3	4	10	11	12	13	14	15	16	19	20	24
SEVERITYCODE															
1	75	34	41	4	1	6660	3337	751	1243	4572	775	745	233	763	70
2	770	323	571	78	18	4227	530	85	916	3831	240	180	28	130	125

ST_COLCODE	25	26	27	28	29	30	32	45	71	72	73	74	81	82	83	84
SEVERITYCODE																
1	19	243	61	1858	72	239	10065	209	366	21	34	87	243	12	17	21
2	14	87	14	1465	13	139	639	1335	23	5	18	69	18	NaN	4	11

Figure 8: Pivot Table of State Collision Code Dictionary and Accident Count

4.1.8 Relationship between Collision Types and Accident Count

It can be observed that some of the accidents that occurred much more frequently such as those involving Parked Cars were generally not as severe as accidents that occurred less, such as those involving Pedestrians and Cycles. It can also be seem that Sideswipes were also usually not very severe. Other Collision Types had a significant amount of accidents resulting in both Severity Codes 1 and 2, such as Left Turns, Rear Ended and Head On.



	INCKEY								
COLLISIONTYPE	Angles	Cycles	Head On	Left Turn	Parked Car	Pedestrian	Rear Ended	Right Turn	Sideswipe
SEVERITYCODE									
1	20302	606	1085	7990	29852	626	18234	2207	15019
2	13366	4585	853	5330	2572	5631	14114	594	2403

Figure 9: Pivot Table of Collision Types against Accident Counts

4.2 Classification Models

Since the purpose of the model is to categorise accidents into the likely severity codes depending on the characteristics of the accidents, I chose to use classification models over regression models.

The applications of classification models were quite straightforward as I simply needed to split the original data set into training and test sets and run them through the models. The models applied here are k-Nearest Neighbours, Decision Trees, Support Vector Machines (SVM) and Logistic Regressions. Generally, the metrics used to evaluate the models are Jaccard Index, F1 Score and Logarithmic Loss (Log Loss).

4.2.1 Performance of Classification Models

As seen from the results (see table 2), Classification model k-Nearest Neighbours is the best model as it had the highest Jaccard Index of 0.707 and highest F1 score of 0.670. (Highlighted in yellow in table 2)

Table 2

Evaluation Results of Classification Models

Evaluation Metrics	k-Nearest Neighbours	Decision Tree	SVM	Logistic Regression
Jaccard Index	0.707	0.698	0.684	0.657
F1 Score	0.670	0.624	0.640	0.524
Log Loss	N.A.	N.A.	N.A.	0.623

5. Conclusion

In this study, I analysed the relationship between the severities of vehicle accidents and the circumstances surrounding the accidents. I identified collision, address, junction types, and road, light and weather conditions to be some of the most important features that affect the severity of an accident. Classification models were built to predict the severity of accidents, so that drivers can use these predictions to gauge the impact of the accident on traffic. This could be useful to drivers as they can change their route of advancements early so avoid being stuck in traffic jams should an accident occur.

6. Future Directions

Despite the classification models achieving about 70.7% accuracy, the circumstances around accidents are very dependent on the drivers themselves and how careful and alert they are while driving. Although accidents are not always preventable, we should actively try to reduce the chances of an accident by driving safely, especially when there are conditions that make driving riskier, such as rain and poor visibility. Ultimately, even though this model can help to reduce inconvenience to some, one should not take lives for granted and instead consider the consequences of reckless driving.

Works Cited

Balk, Gene. “Booming Seattle Is Adding Cars Just as Fast as People.” The Seattle Times, The Seattle Times Company, 9 Aug. 2017, www.seattletimes.com/seattle-news/data/booming-seattle-is-adding-cars-just-as-fast-as-people/.

Baruchman, Michelle. Seattle Traffic Deaths and Injuries down Slightly Last Year; Most of the Fatalities Were Pedestrians. 12 Mar. 2019, www.seattletimes.com/seattle-news/transportation/seattle-traffic-deaths-and-injuries-down-slightly-last-year-most-of-the-fatalities-were-pedestrians/.

“Seattle Metro Area Population 1950-2020.” MacroTrends, www.macrotrends.net/cities/23140/seattle/population.