Predicting Severity of Traffic Accidents in Seattle, WA

SO Cheuk Hei

31st August 30, 2020

Introduction

In the US, the main mode of transport is driving, and this can be seen by the advanced networks of highways and relatively simple public transit architecture in US cities. However, the reliance on highways meant that the traffic flow can be easily disrupted by accidents, and major efforts and resources would be needed to redistribute stuck traffic. This will extend to massive time and financial costs to both the Government and citizens.

To reduce future instances of such cost incurrence, a long-term demand has arisen for an Advisory that:

1. **advises drivers on the probability and severity of traffic accidents, proactive actions and alternative routes based on road condition and weather**, which helps drivers and commuters plan ahead and reduces inconveniences, and;
2. **advises medical workers to plan resources more efficiently to conduct more effective and timely rescues and treatment of casualties**

This report is targeted at the Government of Seattle and aims to illustrate the development of a simple machine learning algorithm which meets the aforementioned needs.

Data

The dataset used for this report ("Data") originated from and was recorded by Traffic Records Group of Seattle Police Department's Traffic Management Division. The Data includes all observations of traffic collisions from 2004 to the present and records all 37 attributes of each collision, including weather, collision address, road condition etc.

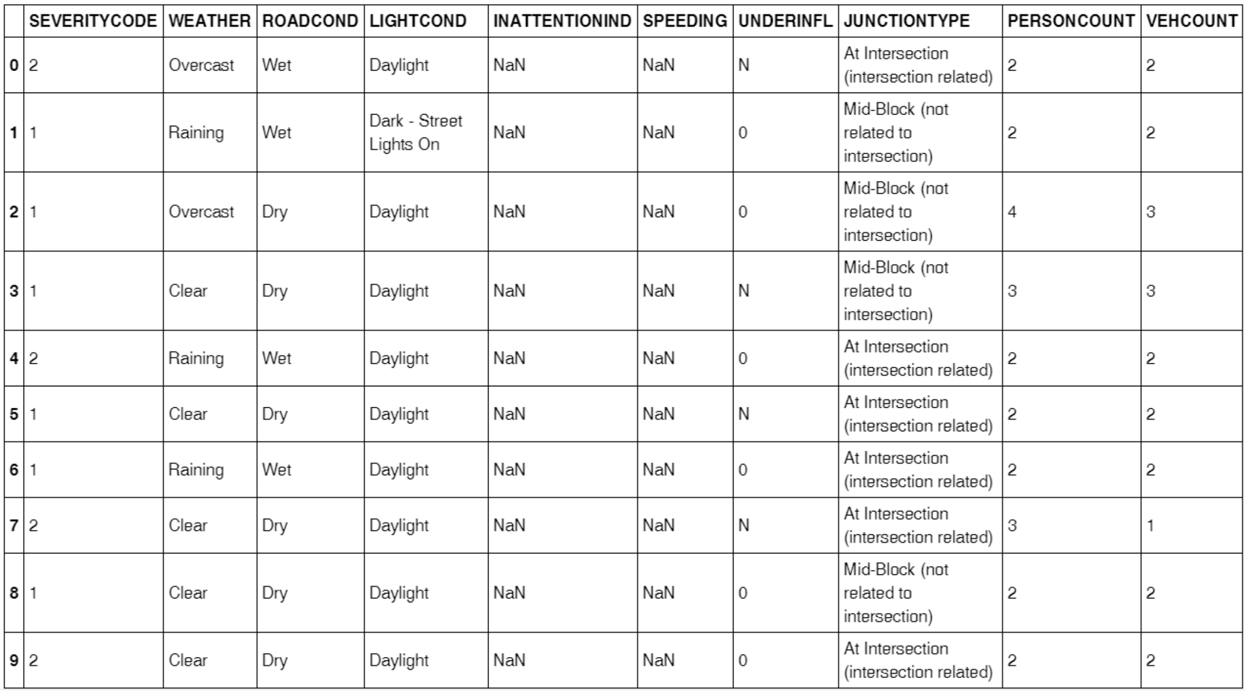
The Data includes 37 attributes for each observation and 194,674 observations from 2004 to the present. To enable an effective Traffic Advisory, several human and non-human attributes that are directly related to a traffic accident will be used for modelling. These 8 attributes include:

ROADCOND	The condition of the road during the collision.
WEATHER	A description of the weather conditions during the time of the collision.
LIGHTCOND	The light conditions during the collision.
INATTENTIONIND	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
SPEEDING	Whether or not speeding was a factor in the collision. (Y/N)
PERSONCOUNT	The total number of people involved in the collision
VEHCOUNT	The number of vehicles involved in the collision.

Methodologies

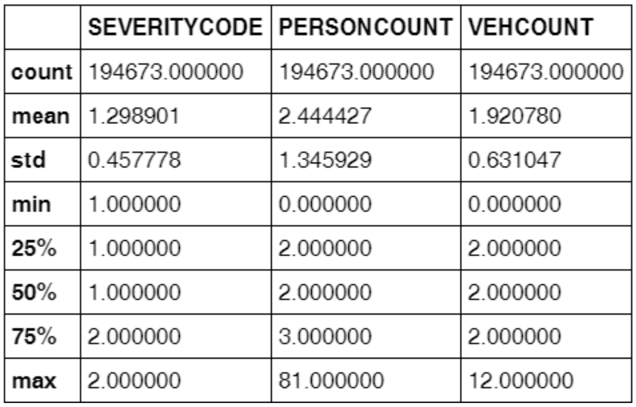
1. Data Understanding

Before making any changes to the Data, an exploratory data analysis on data labels and the key attributes was conducted to better familiarize with the Data and to predict what kinds of changes are needed to facilitate the subsequent modelling process.



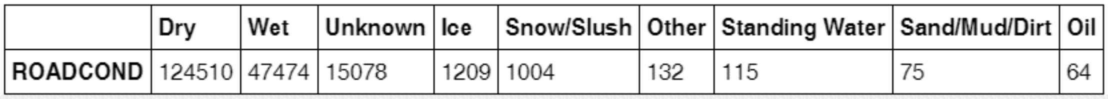
	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	INATTENTIONIND	SPEEDING	UNDERINFL	JUNCTIONTYPE	PERSONCOUNT	VEHCOUNT
0	2	Overcast	Wet	Daylight	NaN	NaN	N	At Intersection (intersection related)	2	2
1	1	Raining	Wet	Dark - Street Lights On	NaN	NaN	0	Mid-Block (not related to intersection)	2	2
2	1	Overcast	Dry	Daylight	NaN	NaN	0	Mid-Block (not related to intersection)	4	3
3	1	Clear	Dry	Daylight	NaN	NaN	N	Mid-Block (not related to intersection)	3	3
4	2	Raining	Wet	Daylight	NaN	NaN	0	At Intersection (intersection related)	2	2
5	1	Clear	Dry	Daylight	NaN	NaN	N	At Intersection (intersection related)	2	2
6	1	Raining	Wet	Daylight	NaN	NaN	0	At Intersection (intersection related)	2	2
7	2	Clear	Dry	Daylight	NaN	NaN	N	At Intersection (intersection related)	3	1
8	1	Clear	Dry	Daylight	NaN	NaN	0	Mid-Block (not related to intersection)	2	2
9	2	Clear	Dry	Daylight	NaN	NaN	0	At Intersection (intersection related)	2	2

The following summarizes the key statistics for quantitative attributes of the Data:

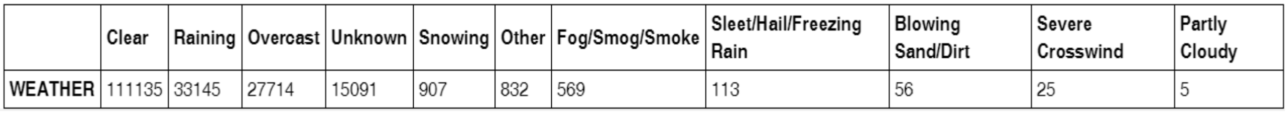


	SEVERITYCODE	PERSONCOUNT	VEHCOUNT
count	194673.000000	194673.000000	194673.000000
mean	1.298901	2.444427	1.920780
std	0.457778	1.345929	0.631047
min	1.000000	0.000000	0.000000
25%	1.000000	2.000000	2.000000
50%	1.000000	2.000000	2.000000
75%	2.000000	3.000000	2.000000
max	2.000000	81.000000	12.000000

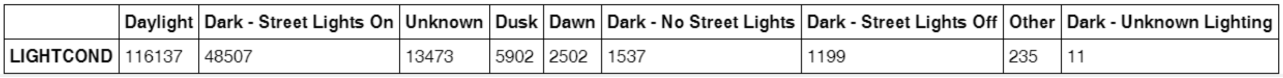
A count of values was also conducted for each of the 8 attributes:



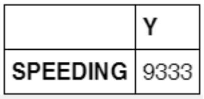
	Dry	Wet	Unknown	Ice	Snow/Slush	Other	Standing Water	Sand/Mud/Dirt	Oil
ROADCOND	124510	47474	15078	1209	1004	132	115	75	64



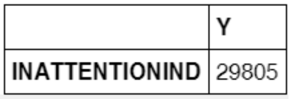
	Clear	Raining	Overcast	Unknown	Snowing	Other	Fog/Smog/Smoke	Sleet/Hail/Freezing Rain	Blowing Sand/Dirt	Severe Crosswind	Partly Cloudy
WEATHER	111135	33145	27714	15091	907	832	569	113	56	25	5



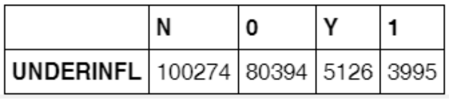
	Daylight	Dark - Street Lights On	Unknown	Dusk	Dawn	Dark - No Street Lights	Dark - Street Lights Off	Other	Dark - Unknown Lighting
LIGHTCOND	116137	48507	13473	5902	2502	1537	1199	235	11



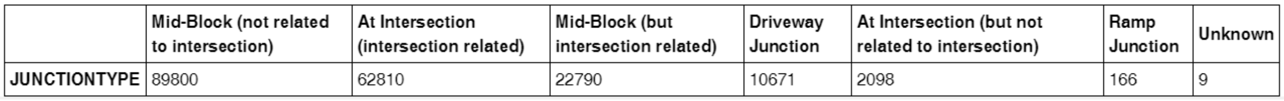
	Y
SPEEDING	9333



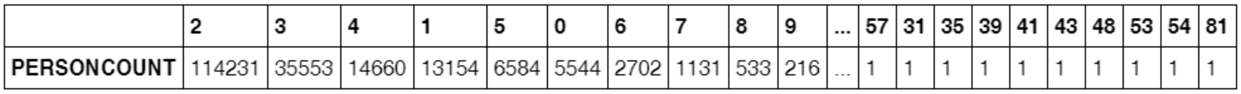
	Y
INATTENTIONIND	29805



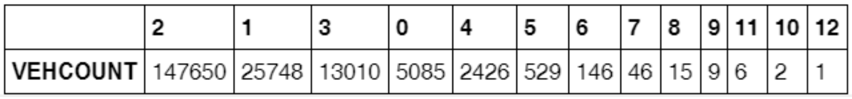
	N	0	Y	1
UNDERINFL	100274	80394	5126	3995



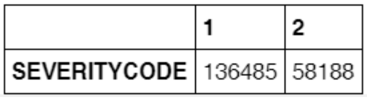
	Mid-Block (not related to intersection)	At Intersection (intersection related)	Mid-Block (but intersection related)	Driveway Junction	At Intersection (but not related to intersection)	Ramp Junction	Unknown
JUNCTIONTYPE	89800	62810	22790	10671	2098	166	9



	2	3	4	1	5	0	6	7	8	9	...	57	31	35	39	41	43	48	53	54	81
PERSONCOUNT	114231	35553	14660	13154	6584	5544	2702	1131	533	216	...	1	1	1	1	1	1	1	1	1	1

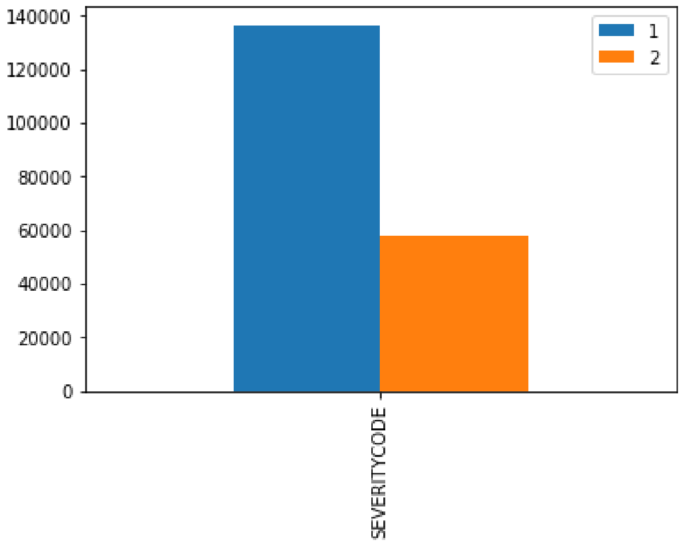


	2	1	3	0	4	5	6	7	8	9	11	10	12
VEHCOUNT	147650	25748	13010	5085	2426	529	146	46	15	9	6	2	1



	1	2
SEVERITYCODE	136485	58188

Upon closer examination, many deficiencies have been found that would potentially affect the precision of the machine learning model. Not only were parts of the observations labelled as “Unknown” or “Other”, but the Data was also severely imbalanced, with the number of observations marked with Severity Code 1 exceeding that marked with Code 2 by more than 1.1 times, as seen below:



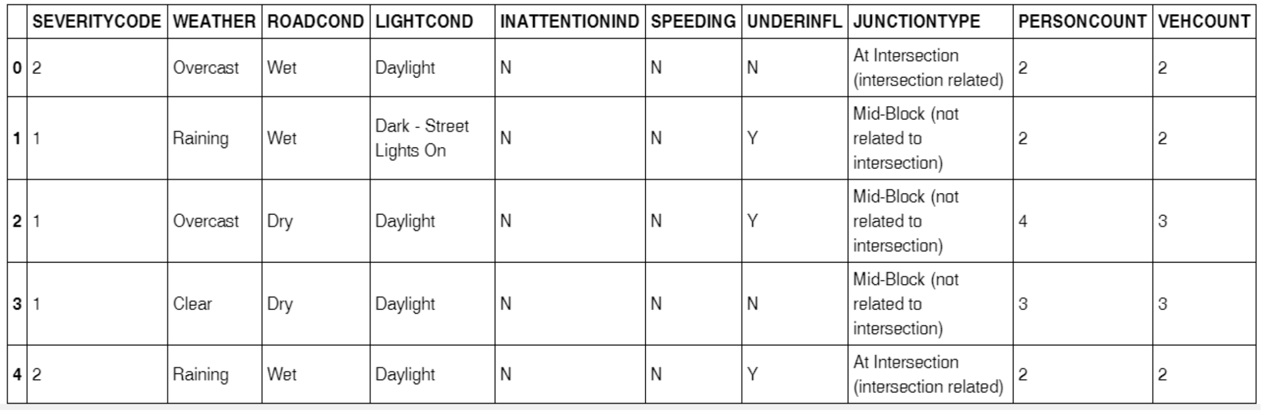
Hence, to avoid developing a biased machine learning model, the Data would be processed further by removing all observations labelled with “Unknown” and “Others”, and balancing the dataset.

2. Data Preparation

2.1. Feature Engineering

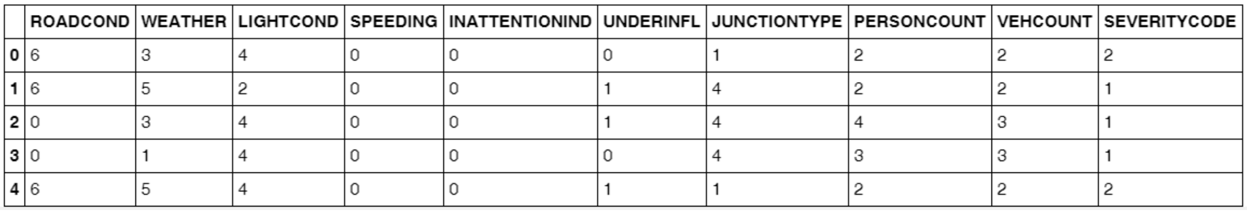
The first objective was to remove all observations labelled with “Unknown” and “Others”, as these observations did not provide sufficient information to help an unbiased machine learning model. For binary attributes such as “**SPEEDING**” and “**INATTENTIONIND**”, however, it was important that the negative, unmarked values were not overlooked and wrongly deleted as missing values.

To achieve this, the unmarked values were replaced by an “N” to indicate negative values, leaving behind the “pure” missing values that were dropped with **df.dropna()** to produce the following:



	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	INATTENTIONIND	SPEEDING	UNDERINFL	JUNCTIONTYPE	PERSONCOUNT	VEHCOUNT
0	2	Overcast	Wet	Daylight	N	N	N	At Intersection (intersection related)	2	2
1	1	Raining	Wet	Dark - Street Lights On	N	N	Y	Mid-Block (not related to intersection)	2	2
2	1	Overcast	Dry	Daylight	N	N	Y	Mid-Block (not related to intersection)	4	3
3	1	Clear	Dry	Daylight	N	N	N	Mid-Block (not related to intersection)	3	3
4	2	Raining	Wet	Daylight	N	N	Y	At Intersection (intersection related)	2	2

Next, the categorical attributes were encoded to become quantitative labels with **LabelEncoder** from Sci-Kit Learn. The dataframe with the encoded attributes is shown here:

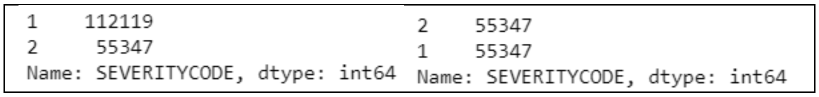


	ROADCOND	WEATHER	LIGHTCOND	SPEEDING	INATTENTIONIND	UNDERINFL	JUNCTIONTYPE	PERSONCOUNT	VEHCOUNT	SEVERITYCODE
0	6	3	4	0	0	0	1	2	2	2
1	6	5	2	0	0	1	4	2	2	1
2	0	3	4	0	0	1	4	4	3	1
3	0	1	4	0	0	0	4	3	3	1
4	6	5	4	0	0	1	1	2	2	2

2.2. Balancing Dataset

As shown previously, the Data is an imbalanced dataset, with Code 1 accidents severely outnumbering Code 2 accidents. Such an imbalanced dataset would lead to a machine learning model that would be biased towards producing Code 1 predictions. As such, the dataset was balanced by down-sampled the majority class, i.e. Code 1 class.

This was done by the **resample** function from Sci-Kit Learn and the number of observations from Code 1 class was reduced by randomly removing observations from the class. After balancing, the number of Code 1 observations became the same as Code 2 observations.



```
1    112119                              2    55347
2     55347                              1    55347
Name: SEVERITYCODE, dtype: int64         Name: SEVERITYCODE, dtype: int64
```

3. Modelling

3.1. Pre-processing

In the modelling process, the predictor variables (X) included the 8 attributes chosen from the Data, i.e. **"ROADCOND", "LIGHTCOND", "WEATHER", "SPEEDING", "INATTENTIONIND", "JUNCTIONTYPE", "PERSONCOUNT", "VEHCOUNT"**, and they are normalized. The target variable (Y) was set to **"SEVERITYCODE".**

Using the **Train/Test Split** function, the Data was divided into 2 mutually exclusive datasets: the testing set and training set, the former of which represented 30% of all data.

3.2. Fitting the Model

3 machine learning algorithms, all of which are classification algorithms, were used. These 3 algorithms are Logical Regression (LR), K-Nearest Neighbors (KNN) and Decision Tree. These algorithms are used for their popularity of use.

After training and fitting the model with the aforementioned training and testing sets, the models were evaluated using a new testing set that was randomly selected from the Data and comprised of 30% of the total Data. Each model was tested and their

Jaccard Index, F1 Scores and Log Loss (Logical Regression only) were calculated to evaluate the accuracy of the models.

Results

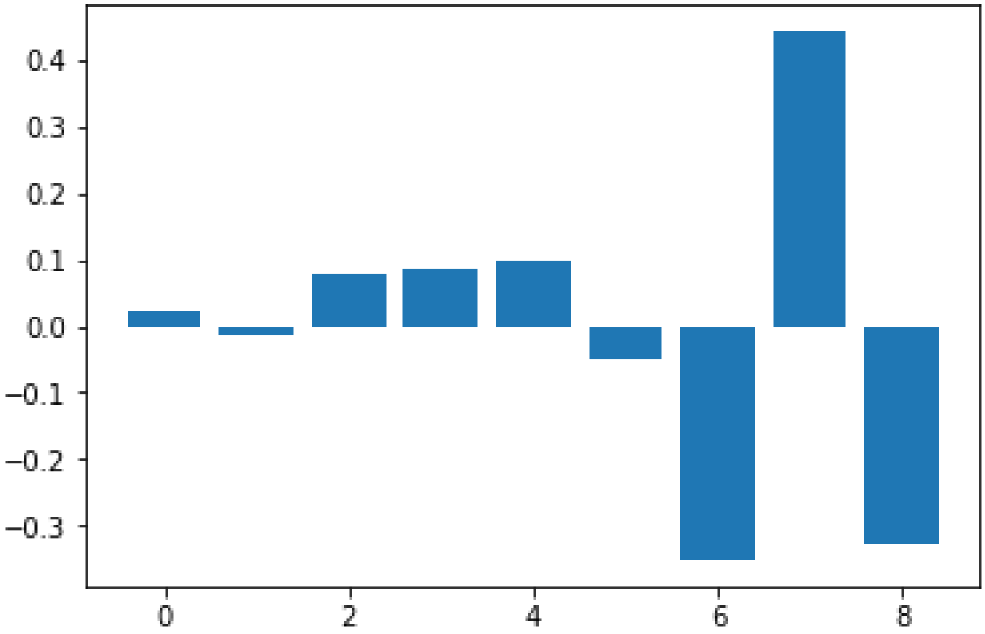
The Jaccard Indices, F1 scores and Log Loss of the models are shown below:

	Jaccard	**F1**	**Log Loss**
KNN	0.64	0.64	N/A
LR	0.62	0.62	0.66
Decision Tree	0.67	0.67	N/A

As seen above, the Decision Tree emerged as the most accurate model with 67% accuracy.

The most important attributes that would affect the severity of a Traffic Accident are:

- Number of people involved. **("PERSONCOUNT", F7)**
- Whether speeding was involved. **("SPEEDING", F4)**
- Whether or not collision was due to inattention. **("INATTENTIONIND", F3)**
- The light conditions during the collision. **("LIGHTCOND", F2)**
- The condition of the road during the collision. **("ROADCOND", F0)**



Discussions

As shown before in the Data Understanding section, many observations were unusable due to them being labelled with "unknown" and "other" in some of the attributes. The ambiguity in observations might contribute to a lowered accuracy in the final machine learning model. Besides, down-sampling was used to balance the Data due to limited cloud computing capabilities. It is noteworthy that up-sampling the minority class might increase the accuracy of the final model.

The results provide actionable insights that would reduce the probability of traffic accidents. It is recommended that, in the short term, the Government increase promotion on safe driving to reduce instances of speeding and inattention (e.g. texting while driving). In addition, the Government should be aware of the importance of maintaining road conditions and sufficient illumination on all types of roads. In the long-term, the Government should set up a public advisory which will advise drivers and commuters on potential accidents based on observable attributes such as road condition; the hospitals should also be granted access to the advisory to plan resources accordingly to minimize casualties.

Conclusion

This report summarized the need for a system that could provide advice on the probability and severity of traffic accidents for drivers and medical workers. Through the analysis of past traffic data and machine learning, a decision tree model has emerged as the optimal model to predict traffic accident severity. The data science process also provided actionable insights that could improve road safety in the short and long-term.

Appendix

Github Code:

https://github.com/31245678/Coursera_Capstone/blob/master/CasptoneFinal.ipynb