Predicting the severity of motor vehicle collisions

1 INTRODUCTION

This project aims to develop a model that can be used to predict the expected severity of a motor vehicle accidents (should one occur) based on road conditions, weather and other environmental attributes derived from real world motor vehicle accident statistics. The model could be of interest to the following stakeholders

- Satellite Navigation providers who could combine this severity (consequence) model with model for computing accident likelihood based on locations, thus allowing the provider to deliver a navigation along the lowest risk route
- Emergency service dispatchers who could use the model for decision support or supplementary information when determining a course of action to take when calls are received with insufficient details about a motor vehicle collision

2 DATA

At a minimum, the following data is required to construct a model to estimate accident severity

- Collision statistics that include a severity measure
- Location information or road characteristics for each of the collisions to allow extrapolation to other similar sections of road
- Road surface condition and other environmental features that relate to each of the collisions

The viability of producing an accurate collision severity model will utilise the collision data from the Seattle Police Department accessible via the following link: Seattle Collision Data.

A description of the dataset can be found via the following link: Seattle Collision Metadata.

3 METHODOLOGY

3.1 EXPLORATORY DATA ANALYSIS

From the Metadata descriptions and inspecting the output of the head function, the following columns containing identifier and key values will not be investigated:

- OBJECTID
- INCKEY
- COLDETKEY
- REPORTNO

The remaining columns were examined through some high-level statistical analysis of the data to aid in narrowing down relevant features.

The following figures contain the result of applying the ‘describe’ method to the columns of the data.

	SEVERITYCODE	X	Y	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE
count	194673.000000	189339.000000	189339.000000	194673	192747	65070.000000	191996	84811
unique	NaN	NaN	NaN	2	3	NaN	24102	2
top	NaN	NaN	NaN	Matched	Block	NaN	BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB ...	
freq	NaN	NaN	NaN	189786	126926	NaN	276	79173
mean	1.298901	-122.330518	47.619543	NaN	NaN	37558.450576	NaN	NaN
std	0.457778	0.029976	0.056157	NaN	NaN	51745.990273	NaN	NaN
min	1.000000	-122.419091	47.495573	NaN	NaN	23807.000000	NaN	NaN
25%	1.000000	-122.348673	47.575956	NaN	NaN	28667.000000	NaN	NaN
50%	1.000000	-122.330224	47.615369	NaN	NaN	29973.000000	NaN	NaN
75%	2.000000	-122.311937	47.663664	NaN	NaN	33973.000000	NaN	NaN
max	2.000000	-122.238949	47.734142	NaN	NaN	757580.000000	NaN	NaN

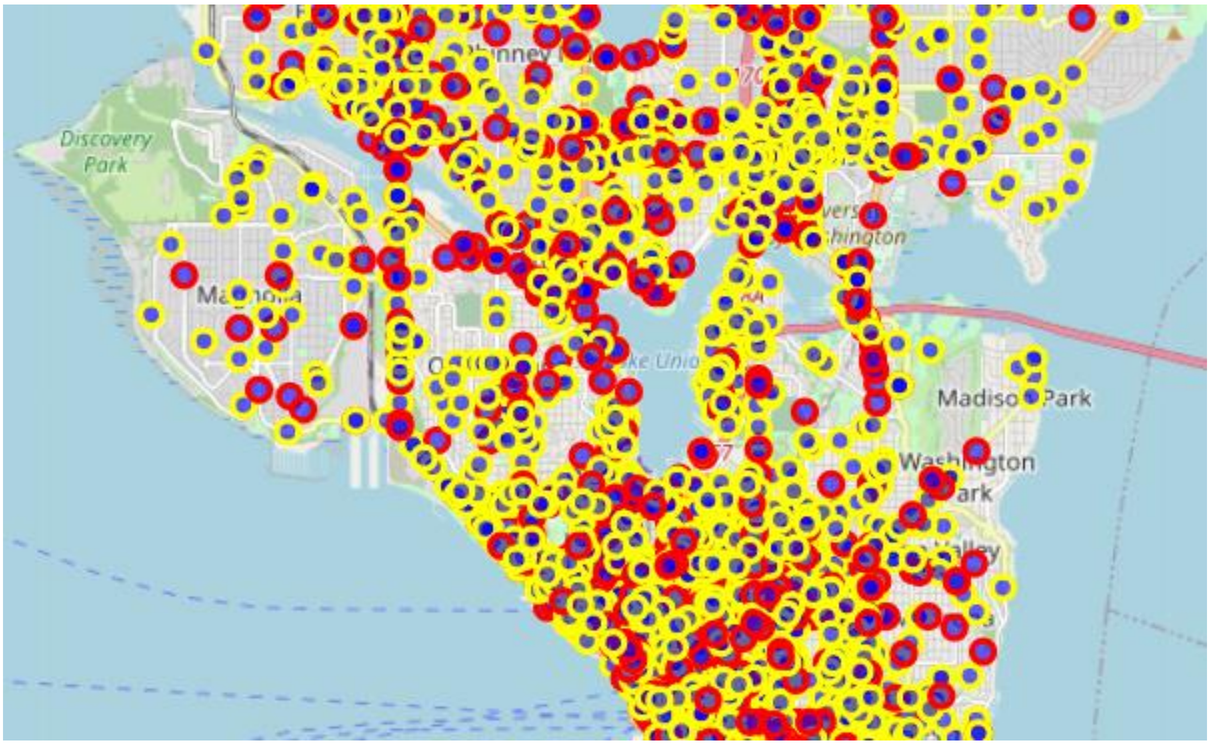
	EXCEPTRSNDESC	SEVERITYCODE.1	SEVERITYDESC	COLLISIONTYPE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	5638	194673.000000	194673	189769	194673.000000	194673.000000	194673.000000	194673.000000
unique	1	NaN	2	10	NaN	NaN	NaN	NaN
top	Not Enough Information, or Insufficient Locati...	NaN	Property Damage Only Collision	Parked Car	NaN	NaN	NaN	NaN
freq	5638	NaN	136485	47987	NaN	NaN	NaN	NaN
mean	NaN	1.298901	NaN	NaN	2.444427	0.037139	0.028391	1.920780
std	NaN	0.457778	NaN	NaN	1.345929	0.198150	0.167413	0.631047
min	NaN	1.000000	NaN	NaN	0.000000	0.000000	0.000000	0.000000
25%	NaN	1.000000	NaN	NaN	2.000000	0.000000	0.000000	2.000000
50%	NaN	1.000000	NaN	NaN	2.000000	0.000000	0.000000	2.000000
75%	NaN	2.000000	NaN	NaN	3.000000	0.000000	0.000000	2.000000
max	NaN	2.000000	NaN	NaN	81.000000	6.000000	2.000000	12.000000

	INCDATE	INCDTTM	JUNCTIONTYPE	SDOT_COLCODE	SDOT_COLDESC	INATTENTIONIND	UNDERINFL
count	194673	194673	188344	194673.000000	194673	29805	189789
unique	5985	162058	7	NaN	39	1	4
top	2006/11/02 00:00:00+00	11/2/2006	Mid-Block (not related to intersection)	NaN	MOTOR VEHICLE STRUCK MOTOR VEHICLE, FRONT END ...	Y	N
freq	96	96	89800	NaN	85209	29805	100274
mean	NaN	NaN	NaN	13.867768	NaN	NaN	NaN
std	NaN	NaN	NaN	6.868755	NaN	NaN	NaN
min	NaN	NaN	NaN	0.000000	NaN	NaN	NaN
25%	NaN	NaN	NaN	11.000000	NaN	NaN	NaN
50%	NaN	NaN	NaN	13.000000	NaN	NaN	NaN
75%	NaN	NaN	NaN	14.000000	NaN	NaN	NaN
max	NaN	NaN	NaN	69.000000	NaN	NaN	NaN

	WEATHER	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING	ST_COLCODE	ST_COLDESC	SEGLANEKEY	CROSSWALKKEY	HITPARKEDCAR
count	189592	189661	189503	4667	1.149360e+05	9333	194655	189769	194673.000000	1.946730e+05	194673
unique	11	9	9	1	NaN	1	63	62	NaN	NaN	2
top	Clear	Dry	Daylight	Y	NaN	Y	32	One parked--one moving	NaN	NaN	N
freq	111135	124510	116137	4667	NaN	9333	44421	44421	NaN	NaN	187457
mean	NaN	NaN	NaN	NaN	7.972521e+06	NaN	NaN	NaN	269.401114	9.782452e+03	NaN
std	NaN	NaN	NaN	NaN	2.553533e+06	NaN	NaN	NaN	3315.776055	7.226926e+04	NaN
min	NaN	NaN	NaN	NaN	1.007024e+06	NaN	NaN	NaN	0.000000	0.000000e+00	NaN
25%	NaN	NaN	NaN	NaN	6.040015e+06	NaN	NaN	NaN	0.000000	0.000000e+00	NaN
50%	NaN	NaN	NaN	NaN	8.023022e+06	NaN	NaN	NaN	0.000000	0.000000e+00	NaN
75%	NaN	NaN	NaN	NaN	1.015501e+07	NaN	NaN	NaN	0.000000	0.000000e+00	NaN
max	NaN	NaN	NaN	NaN	1.307202e+07	NaN	NaN	NaN	525241.000000	5.239700e+06	NaN

3.1.1 Geospatial View

A plot of Seattle with an overview of property damage (yellow) and injury (red) for the first five thousand data points was produced to see if location was significant in the outcome of an incident.



The overview of the first five thousand collisions did not appear to show an obvious bias based on location so will not be used in modelling.

3.1.2 Analysing Discrete Features

Several the columns contain discrete values which merit further investigation. The value_counts method is used to provide a quick overview of the data. An example of the output examined is:

```
In [16]: df['ADDRTYPE'].value_counts()

Out[16]: Block           126926
         Intersection     65070
         Alley              751
         Name: ADDRTYPE, dtype: int64
```

After assessing all remaining columns, the following are candidates for further analysis:

- ADDRTYPE appears useful for generic prediction along routes as Block, Intersection and Alley are relatively easy to determine for other road networks
- JUNCTIONTYPE may be used if there is correlation with severity
- WEATHER, ROADCOND and LIGHTCOND are likely to be useful and will require further analysis

The following columns may be useful during data cleansing and subsequently improving the model:

- EXCEPTRSNCODE and EXCEPTRSNDECS may be a useful detail to identify and drop incomplete information
- INCDATE may be evaluated further to determine whether season or month can improve the accuracy of the model beyond just weather, road condition or light
- INCDTTM may be used in place of INCDATE if date-based improvements are required

The remaining columns will not be investigated further:

- LOCATION appears too specific for a general-purpose prediction.
- SEVERITYCODE.1 and SEVERITYDESC appear to be duplicates of the SEVERITY column and will not be evaluated further
- COLLISIONTYPE is unlikely to be useful as a prediction of the collision type may be difficult to predict but it may be analysed further during modelling as it may be correlated with other features useful for determining routes (e.g. Left Turn at an intersection may be more likely to result in an injury which may require an alternate route)
- The counts will not be further evaluated as they are a consequence of a collision and are unlikely to predict severity
- INATTENTIONIND and UNDERINFL will not be used for predicting severity as they will not be an input into route planning
- All other columns will be excluded

3.1.3 Further Analysis

The candidate data columns require further analysis before inclusion in the model.

Firstly, the relationship between SEVERITYCODE and each of the candidates as well as some basic statistics (count, average and standard deviation). The mean and standard deviation are relevant as severity code is either 1 or 2, so a mean closer to 2 indicates more likelihood of an injury.

The figures below capture the results from each assessed column:

	SEVERITYCODE		
	count	mean	std
ADDRTYPE			
Alley	751	1.109188	0.312082
Block	126926	1.237115	0.425315
Intersection	65070	1.427524	0.494723

NOTE: ADDRTYPE will be used in the model as there is a significant severity ratio difference between Alley, Block and Intersection.

	SEVERITYCODE		
	count	mean	std
JUNCTIONTYPE			
At Intersection (but not related to intersection)	2098	1.296949	0.457023
At Intersection (intersection related)	62810	1.432638	0.495446
Driveway Junction	10671	1.303064	0.459604
Mid-Block (but intersection related)	22790	1.320184	0.466557
Mid-Block (not related to intersection)	89800	1.216080	0.411572
Ramp Junction	166	1.325301	0.469905
Unknown	9	1.222222	0.440959

NOTE: JUNCTIONTYPE may be added to the model after the first iteration if accuracy needs to be improved because it looks like it overlaps with ADDRTYPE

	SEVERITYCODE		
	count	mean	std
WEATHER			
Blowing Sand/Dirt	56	1.267857	0.446850
Clear	111135	1.322491	0.467432
Fog/Smog/Smoke	569	1.328647	0.470135
Other	832	1.139423	0.346596
Overcast	27714	1.315544	0.464741
Partly Cloudy	5	1.600000	0.547723
Raining	33145	1.337185	0.472756
Severe Crosswind	25	1.280000	0.458258
Sleet/Hail/Freezing Rain	113	1.247788	0.433651
Snowing	907	1.188534	0.391353
Unknown	15091	1.054072	0.226167

NOTE: WEATHER will be used in the model as there appears to be enough variation across the different weather conditions that it may be useful.

	SEVERITYCODE		
	count	mean	std
ROADCOND			
Dry	124510	1.321773	0.467158
Ice	1209	1.225806	0.418285
Oil	64	1.375000	0.487950
Other	132	1.325758	0.470443
Sand/Mud/Dirt	75	1.306667	0.464215
Snow/Slush	1004	1.166335	0.372566
Standing Water	115	1.260870	0.441031
Unknown	15078	1.049675	0.217280
Wet	47474	1.331866	0.470888

Note: ROADCOND will be used in the model as there appears to be enough variation across the different road conditions that it may be useful.

	SEVERITYCODE		
	count	mean	std
LIGHTCOND			
Dark - No Street Lights	1537	1.217306	0.412547
Dark - Street Lights Off	1199	1.263553	0.440743
Dark - Street Lights On	48507	1.298411	0.457565
Dark - Unknown Lighting	11	1.363636	0.504525
Dawn	2502	1.329337	0.470066
Daylight	116137	1.331884	0.470892
Dusk	5902	1.329380	0.470028
Other	235	1.221277	0.415992
Unknown	13473	1.044905	0.207102

NOTE: LIGHTCOND will be used in the model as there appears to be enough variation across the different light conditions that it may be useful.

3.2 Data Preparation

Data preparation consisted of the following steps:

- Create a data-frame with the SEVERITYCODE, ADDRTYPE, WEATHER, ROADCOND and LIGHTCOND columns
- Drop all rows that have null entries in any of the columns – it is assumed that there are sufficient data points that dropping data will have minimal impact on the model
- Replace all columns of discrete values with ‘dummies’ to represent each of the values.

The resultant data-frame has 33 columns and 187525 rows.

3.3 Model Development

The intent of the model is to classify collisions as either ‘damage’ or ‘injury’, hence classification machine learning models will be employed to predict injury. The selected models are Decision Trees, K-Nearest Neighbours (KNN), Support Vector Machines (SVM) and Logistic Regression. Note that Logistic Regression is applicable as there are only two labels in the dataset.

The prepared data-frame was split into Training and Test datasets consisting of 168772 and 18753 samples, respectively.

The test dataset contained 12920 damage and 5833 injury samples.

As the samples are biased, the baseline for the model performance should be an improvement on always selecting ‘damage’ which would yield accuracy of 68.9%.

As this is a classification problem, the following models were produced, analysed and summarised in the following table:

Model	Parameters	Test set accuracy
KNN	K = 4	66.3%
Decision Tree	criterion="entropy", max_depth = 6	68.9%
SVC	kernel='rbf'	68.9%
Logistic Regression	C=0.01, solver='liblinear'	68.9%

It should be noted that none of these models was an improvement on always selecting property damage.

After the initial modelling, attempts were made to use different encoding schemes for the discrete values as well as using the StandardScaler and neither improved the predictions.

The correlations were produced for the training dataset and inspecting these showed that the correlations between the SEVERITYCODE and each to the 32 columns confirmed that there is low correlation between any individual column and the occurrence of injury.

4 Results

The result of the investigation is that although nearly 30% of collisions result in an injury, it is not feasible to predict whether or not a collision will result in an injury based on the type of road, weather, condition of road and light conditions.

5 Discussion

Given the poor correlation between any individual feature and injury, it is highly unlikely that an accurate model can be produced from the analysed data.

Although using the actual injury likelihood at a given location may result in a better prediction, it is not useful for a general route planning application.

Further investigation could be conducted into whether temporal information could play a significant role in predicting injury, such as time of day, day of week or season of year, but this is unlikely to be useful for route planning or service allocation as it is not location based.

6 Conclusion

Based on the data analysed and the classification models investigated, it was not possible to build a classification model that could outperform always predicting that an injury did not occur. Instead, other data or modelling techniques are required to build a reliable predictive model.