Predicting the severity of motor vehicle collisions

Problem

Predict the expected severity of a motor vehicle accidents (should one occur) using

- Road conditions
- Weather
- Other environmental attributes

Why useful?

The model could be of interest to

- • Satellite Navigation providers
  - Combine with accident likelihood model to provide navigation along the lowest risk route
- • Emergency service dispatchers
  - Decision support when calls are received with insufficient details about a motor vehicle collision

Dataset used

- Utilise collision data from the Seattle Police Department:
  - Seattle Collision Data
- Description of dataset:
  - Seattle Collision Metadata

Data Analysis (High Level Overview)

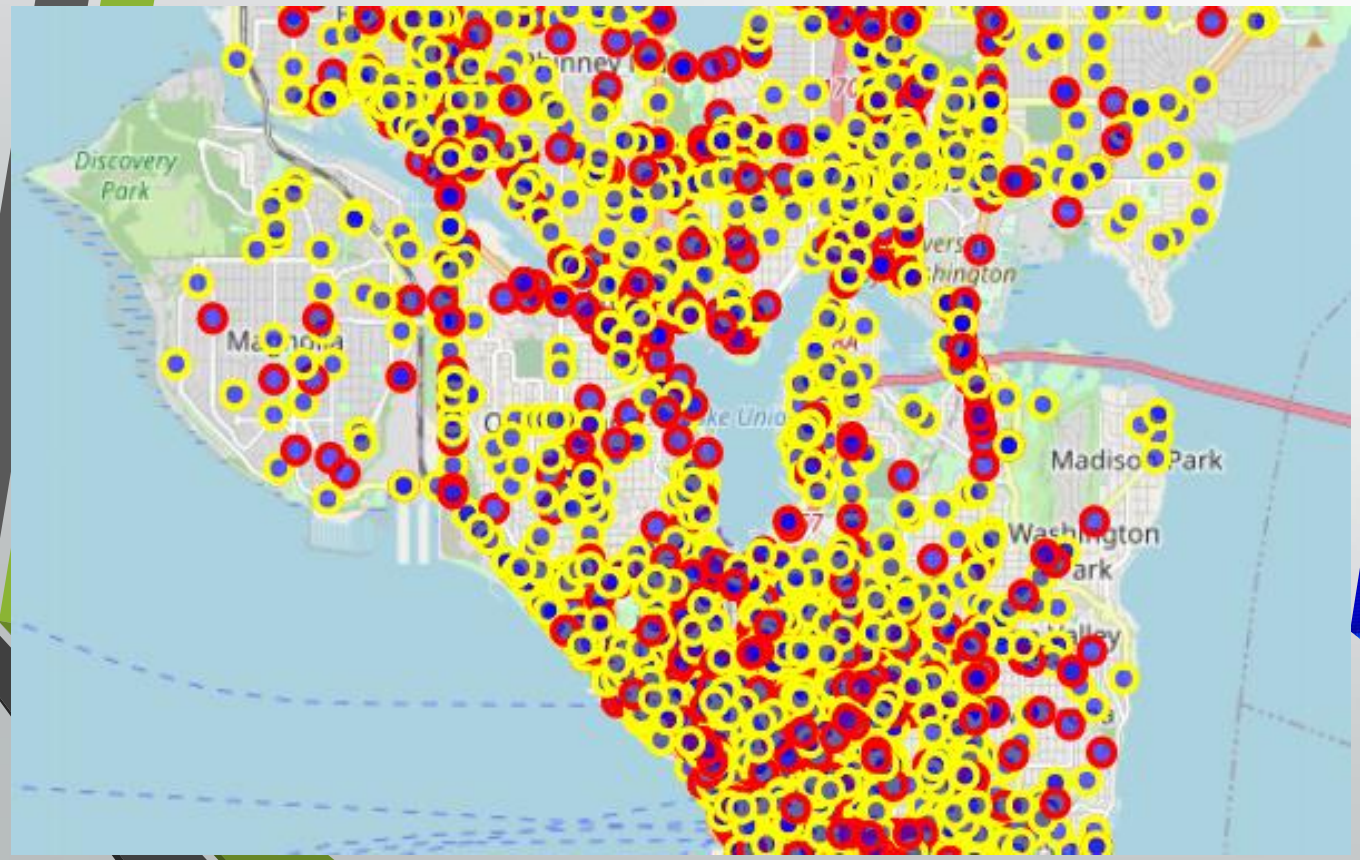
	SEVERITYCODE	X	Y	STATUS	ADDRTYPE	INTKEY	LOCATION	EXCEPTRSNCODE
count	194673.000000	189339.000000	189339.000000	194673	192747	65070.000000	191996	84811
unique	NaN	NaN	NaN	2	3	NaN	24102	2
top	NaN	NaN	NaN	Matched	Block	NaN	BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB ...	
freq	NaN	NaN	NaN	189786	126926	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

Data Analysis (Geospatial)



Data Analysis (Discrete features)

LIGHTCOND	SEVERITYCODE count	SEVERITYCODE mean	SEVERITYCODE std
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

ROADCOND	SEVERITYCODE count	SEVERITYCODE mean	SEVERITYCODE std
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

WEATHER	SEVERITYCODE count	SEVERITYCODE mean	SEVERITYCODE std
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

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

JUNCTIONTYPE	SEVERITYCODE count	SEVERITYCODE mean	SEVERITYCODE std
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

Data Preparation

- Create a data-frame with the SEVERITYCODE, ADDRTYPE, WEATHER, ROADCOND and LIGHTCOND columns
- Drop null
- Encode discrete values as dummies

Result: 33 columns and 187525 rows

Model Development

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%

Conclusion

- High accuracy model unlikely