Katharina Egert

k@egert.org

Oct 2020

IBM CAPSTONE FINAL PROJECT

PREDICTING CAR ACCIDENT SEVERITY

PREDICTING CAR ACCIDENT SEVERITY

- **Goal**: Warn drivers and traffic regulators about high risk driving conditions leading to severe accidents by identifying factors that correlate with high severity accidents.
  - Knowing these risk factors may help drivers to better schedule or reroute their trips.
  - Traffic planners may mitigate risks by taking appropriate measures via construction improvements or speed limits etc.
- **Method**: Use accident data history to understand common causes of severe accidents.

COLLISIONS DATA PROVIDED BY SEATTLE POLICE DEPARTMENT 2004–2020

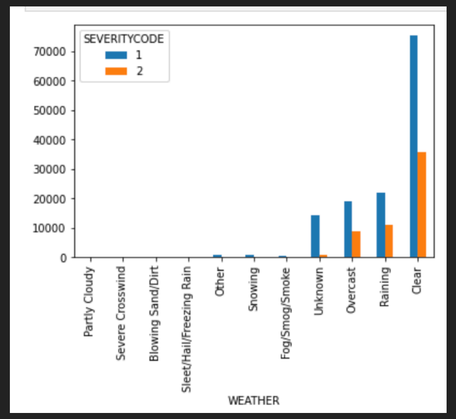
- Accident reports for accident severity levels
  - property damage only (code 1) and
  - including injuries (code 2)
- Exploitable data contains information on
  - weather, road and lighting conditions
  - location data (junction type and GPS coordinates)
  - time/date (that is time of day, weekday and seasonality)
- Data cleaning:
  - Removal of unnecessary technical columns
  - Removal of NaN (missing data) and unknown data rows
- Remark: Data is unbalanced: Property damage (code 1) occurring nearly twice as often

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	NaN	NaN
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354039.0	NaN
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323031.0	NaN
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	NaN	NaN
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028032.0	NaN

5 rows × 38 columns

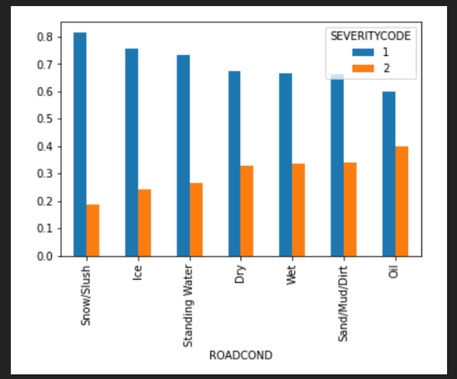
WEATHER

- Snowing, while intuitively making up for a big risk, has actually a lower risk than the globale average (19% instead of 23%) Perhaps drivers are already taking this risk into account and drive more carefully.
- Nice weather, e.g. 'Clear', 'Overcast' or 'Partly Cloudy' still accounts for a more than average amount of injuries.
- Raining is the most dangerous condition.



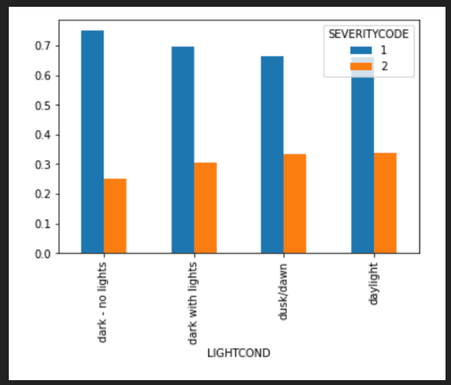
ROAD

- Snow/Slush and Ice are again surprisingly, low risk.
- Oil and Wet yield slippery conditions and are therefore high risk as one would guess.

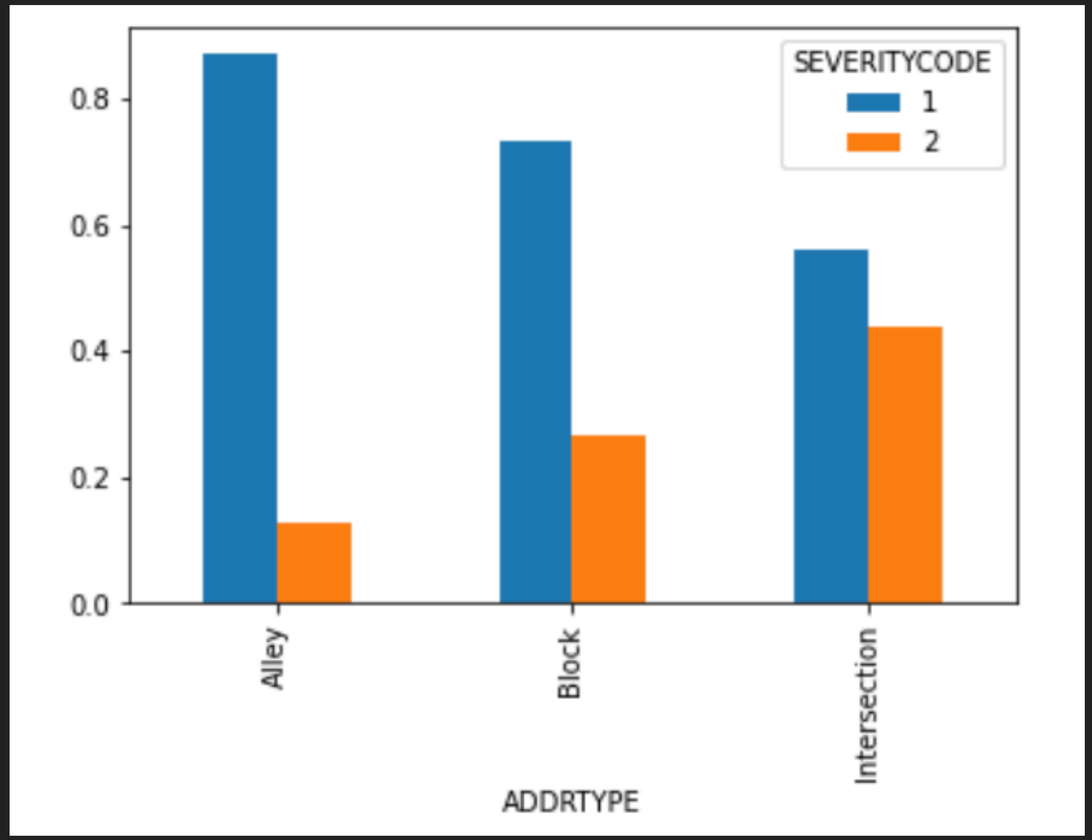


LIGHTING

- As one would expectThe darker the riskier.
- Most accidents happen at daylight, so lighting does not fully explain higher severity accidents.
- Dusk/dawn are similar, also no streetlights when dark, so we can group these together.

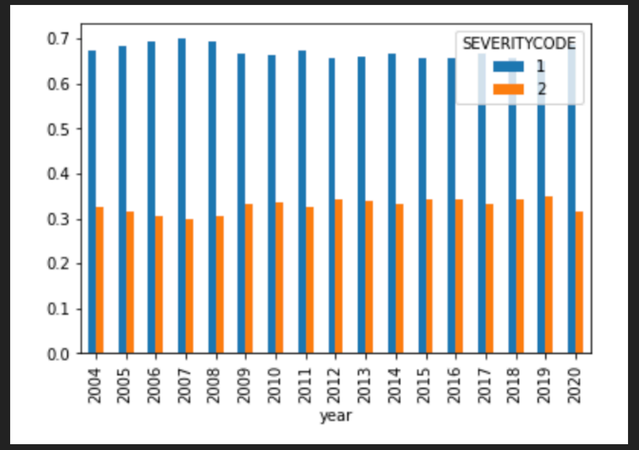


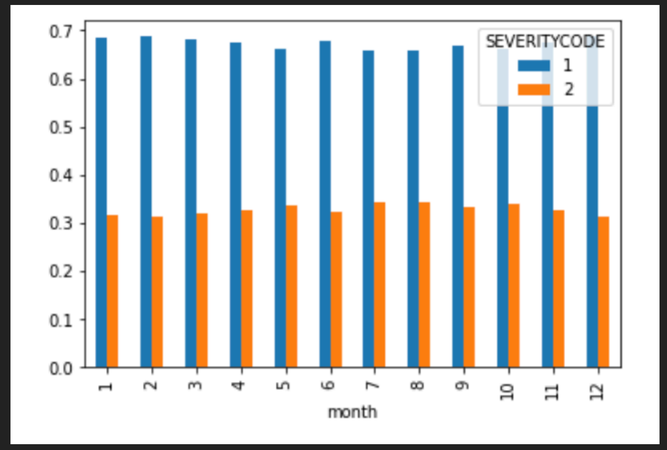
INTERSECTIONS ARE MOST DANGEROUS

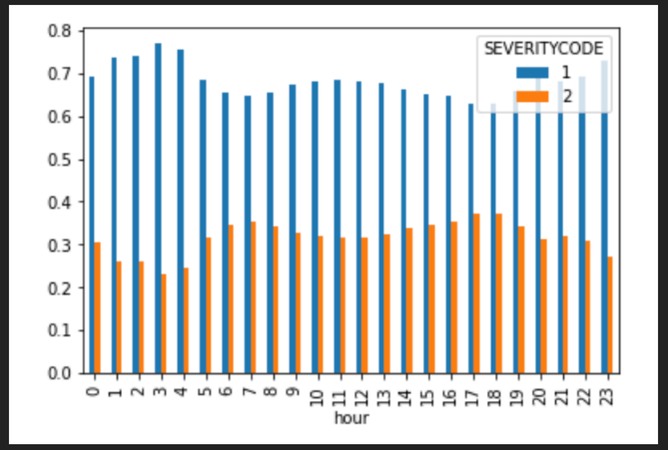


TIME DIMENSION

- Some years are more dangerous than others, however no clear trend.
- Accidents are more severe in summer than in winter.
- Rush hour (between 5AM to 9AM and 3PM to 5PM) is most dangerous.
- Sundays are safest.







- Target: Binary classification
- Independent variables: All categorical
- Identified issue: Low recall for severity code = 2

LOGISTIC REGRESSION

Classification Report: Logistic Regression

	precision	recall	f1-score	support
1	0.73	0.70	0.71	22890
2	0.43	0.48	0.45	11100
micro avg	0.62	0.62	0.62	33990
macro avg	0.58	0.59	0.58	33990
weighted avg	0.63	0.62	0.63	33990

Accuracy: 62.47% (correctly classified test data)
F1: 0.714462 (weighted average of recall and precision)

RANDOM FOREST

Classification Report: Random Forest

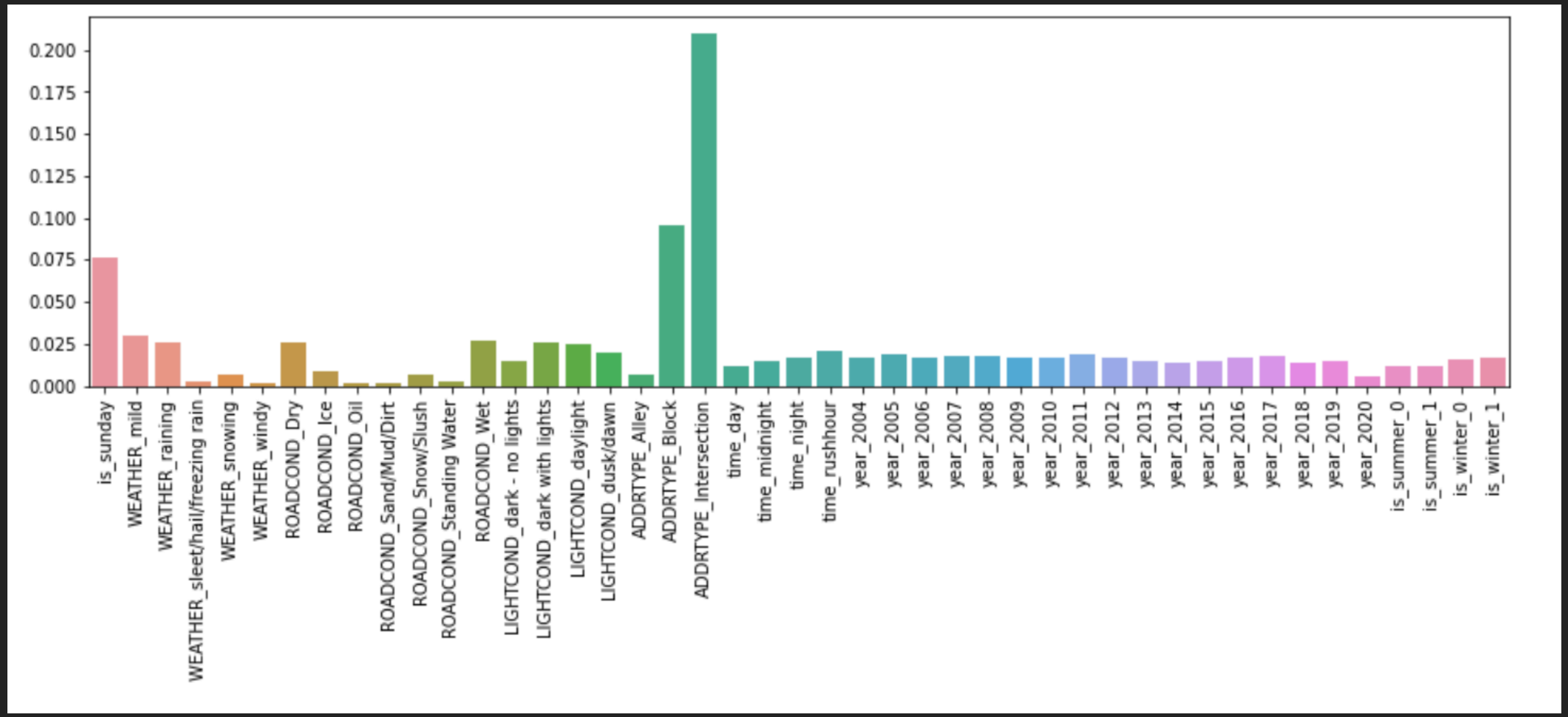
	precision	recall	f1-score	support
1	0.73	0.70	0.71	22890
2	0.43	0.48	0.45	11100
micro avg	0.62	0.62	0.62	33990
macro avg	0.58	0.59	0.58	33990
weighted avg	0.63	0.62	0.63	33990

Accuracy: 62.47% (correctly classified test data)
F1: 0.714462 (weighted average of recall and precision)

Pretty equal,

slight advantage random forests

FEATURE IMPORTANCES



KEY INSIGHTS

- Relative importance: Location matters the most (intersections!)
- Drivers already avoid obvious risks (snowing, darkness etc.) but tend to misjudge slightly less adverse seeming conditions such as raining
- Traffic regulators should try and make intersections safer.

POSSIBLE MODELING IMPROVEMENTS

- Recall (code 2) still low: adjust weights to account for higher cost of misclassifying injury-accidents.
- Introduce numerical scales for weather, light or road conditions from light to severe.