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

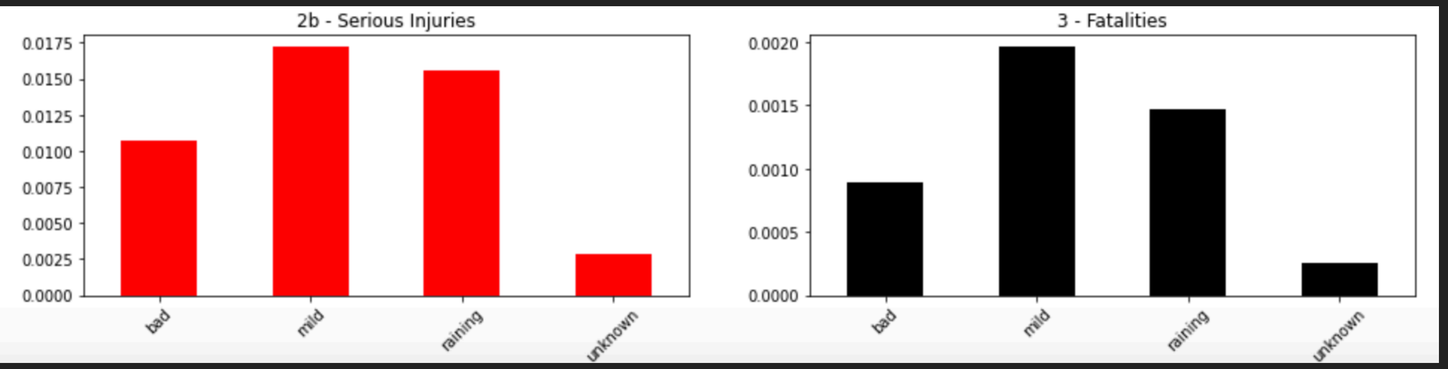
- Accident reports for accident severity levels
  - property damage only (code 1)
  - including injuries (code 2)
  - including severe injuries (2b)
  - including fatalities (3)
- Exploitable data contains information on
  - weather, road and lighting conditions
  - location data (junction type and GPS coordinates)
  - time/date (that is time of day, weekday and seasonality)
  - the fact that involved drivers were intoxicated and/or speeding

- 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, injuries taking most of the remaining parts and the other two categories sharing the last 2%.

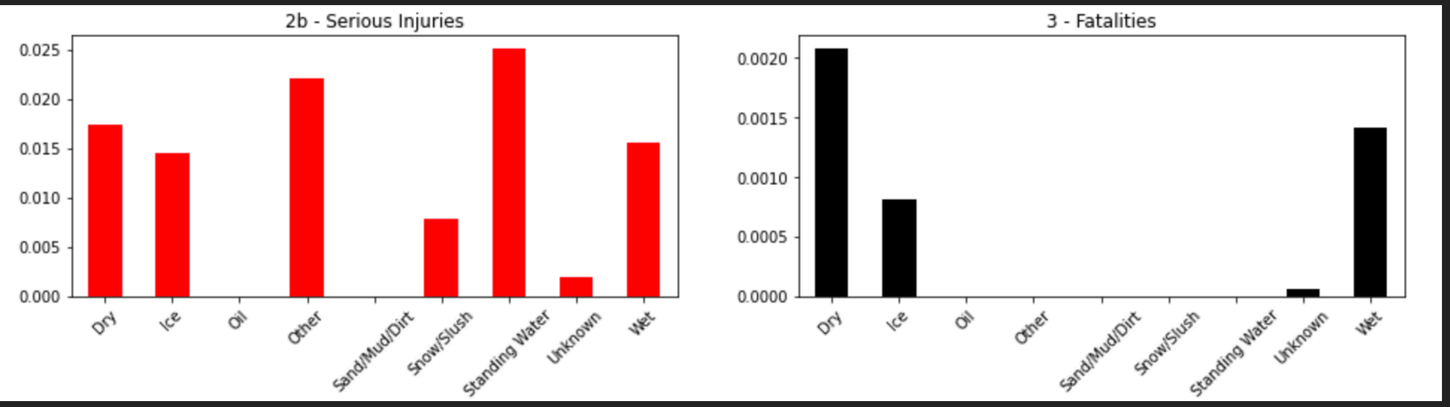
	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOL
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight	NaN	
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On	NaN	6354(
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight	NaN	4323
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight	NaN	
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight	NaN	4028(

5 rows × 38 columns

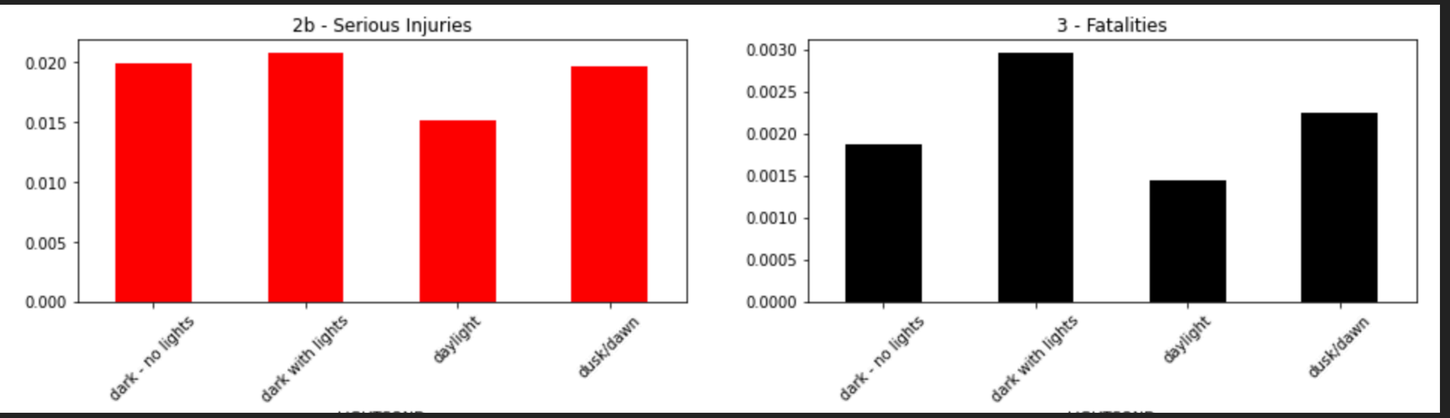
3. DATA ANALYSIS



- Mild and raining are highest risk weather situations for fatalities and serious injuries

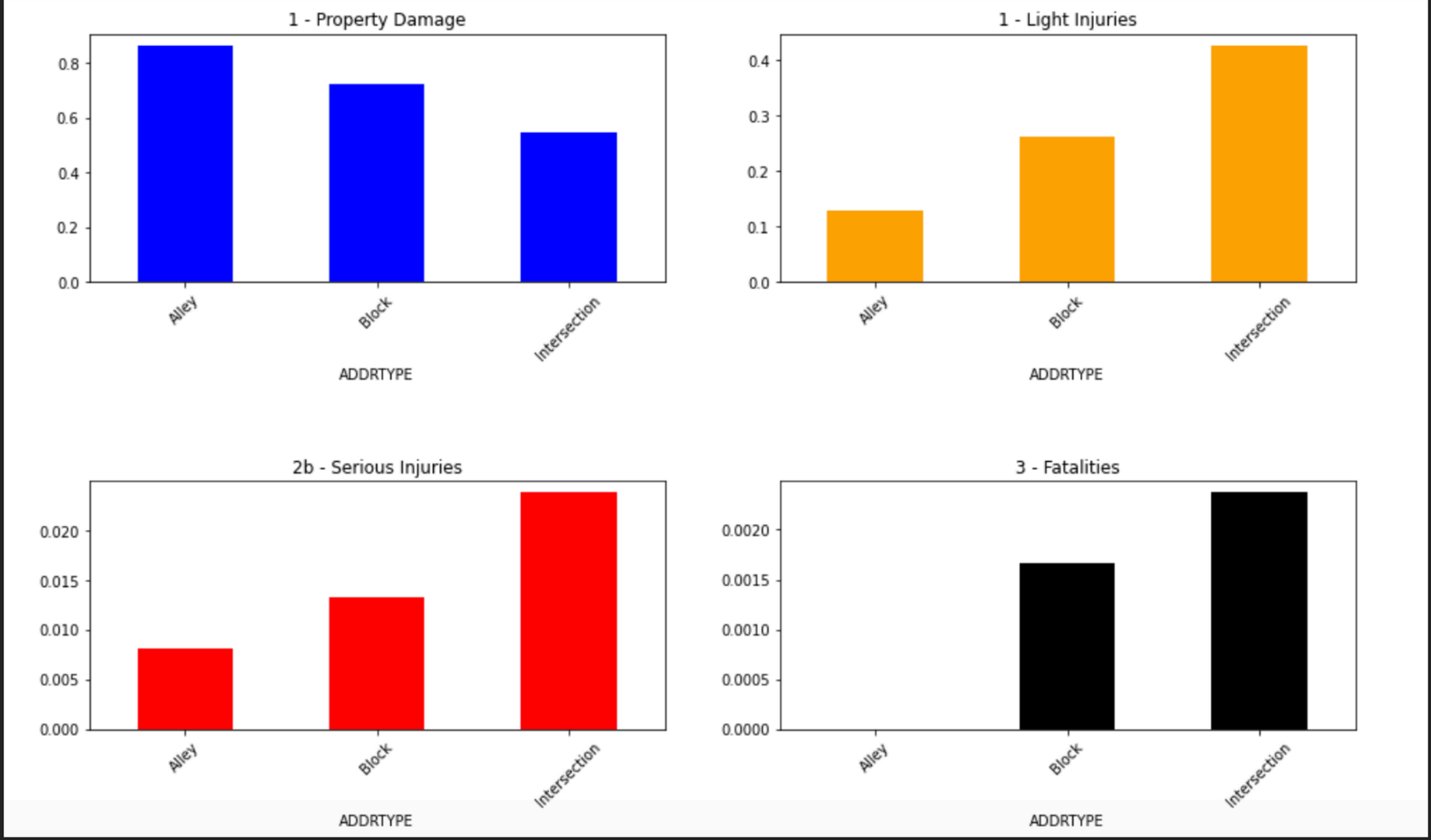


- Dry, wet and ice are high risk for very severe accidents
- ‚medium bad' weather is indicative for serious injuries without fatalities



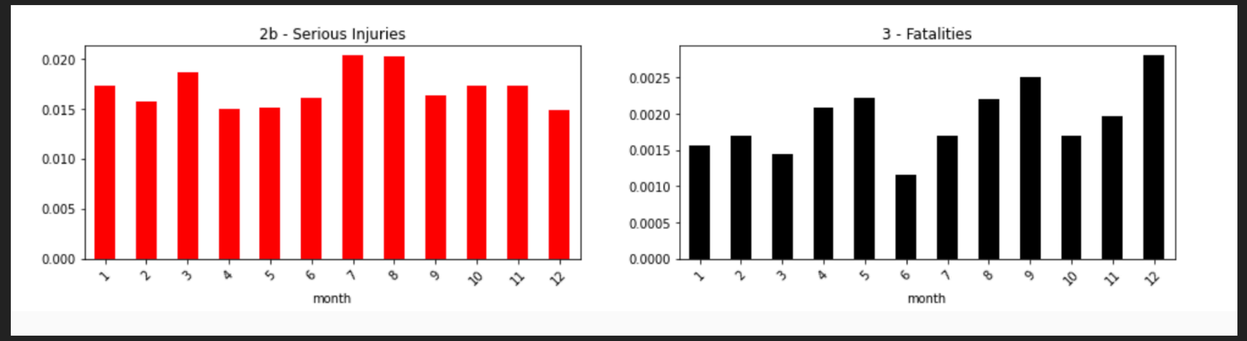
- Fatalities happen most often when it's dark and there are lights and during dusk and dawn

INTERSECTIONS ARE MOST DANGEROUS

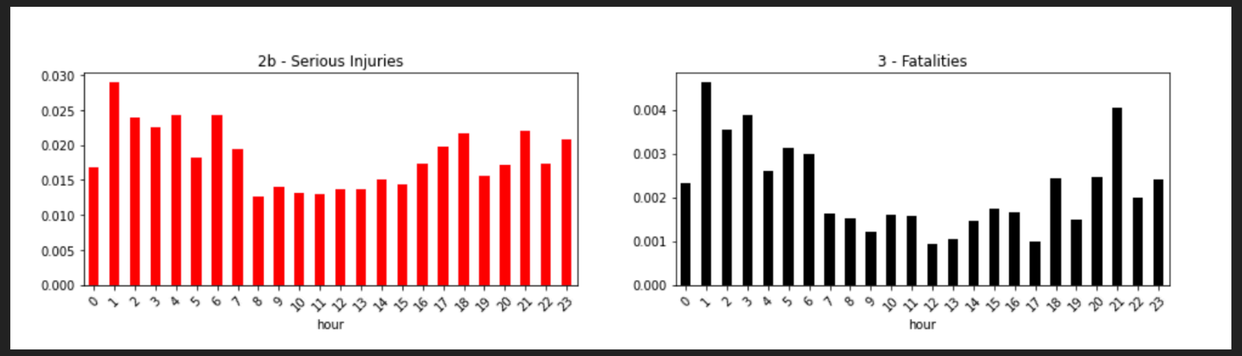


Likelihood of each severity category at a given address type.

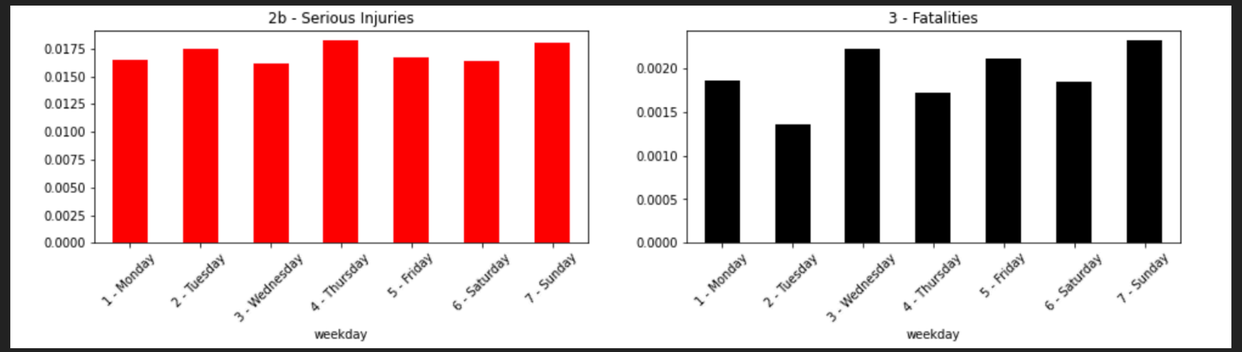
TIME DIMENSION



- Serious injuries much more likely in July, August and March
- Highest fatalities in December



- Less severe accidents during the day from 7am to 5pm
- Peaks in the evening 9pm and high fatality risk at night peaking at 1am

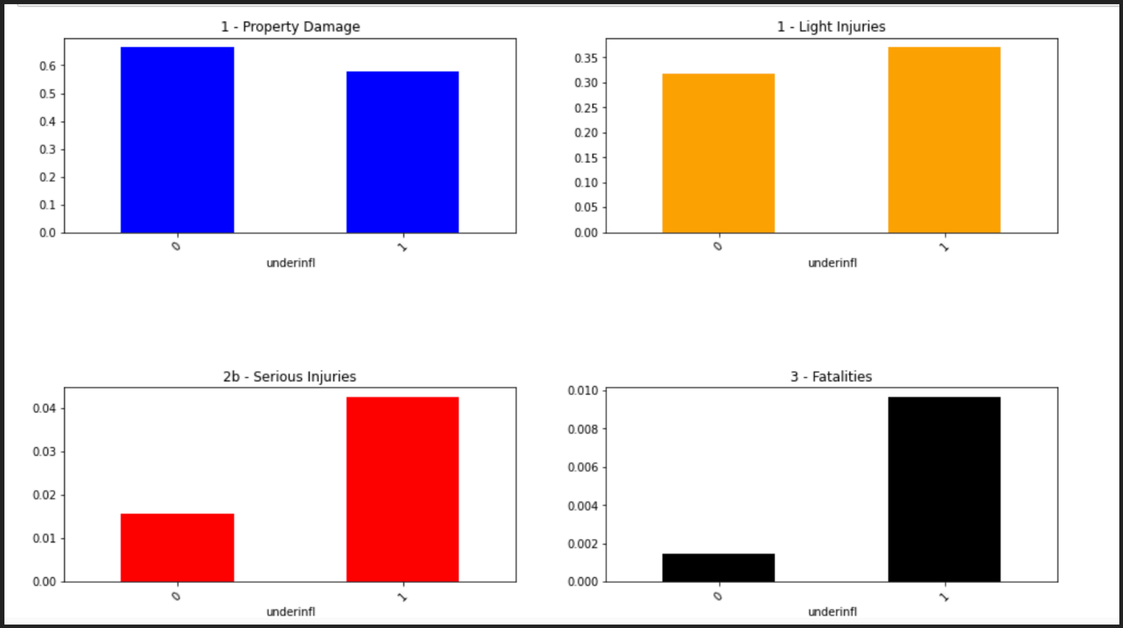


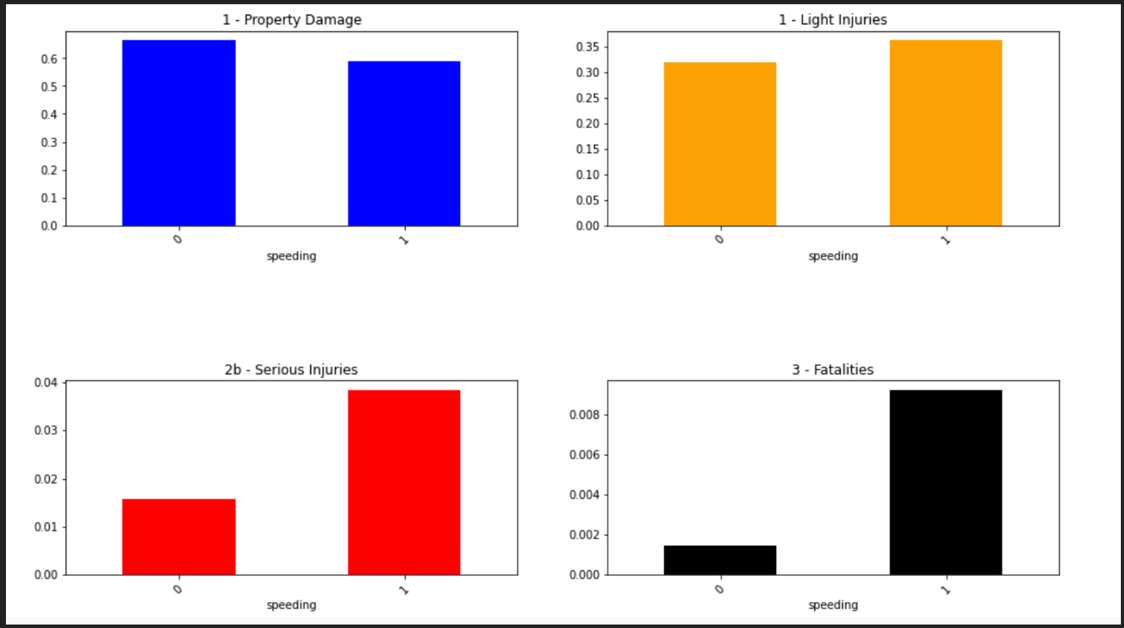
- Least risk on Mondays, Tuesdays and Thursdays for fatalities
- Highest risk on Sundays

UNDER INFLUENCE

SPEEDING

Likelihood of each severity category if drivers were speeding or under influence reps.





- Fatalities more than 10x as likely if drivers were speeding or under influence!
- Severe injuries up to 5x as likely if drivers were speeding or under influence!

4. MODELING

- Target: Multi-category classification
- Independent variables: All categorical
- Identified issue: Low recall for higher severity

LOGISTIC REGRESSION

Classification Report: Logistic Regression

	precision	recall	f1-score	support
1	0.73	0.62	0.67	23348
2	0.42	0.28	0.33	11307
2b	0.03	0.17	0.04	586
3	0.01	0.41	0.02	83
micro avg	0.50	0.50	0.50	35324
macro avg	0.30	0.37	0.27	35324
weighted avg	0.62	0.50	0.55	35324

Accuracy: 49.91% (correctly classified test data)
F1: 0.549666 (weighted average of recall and precision)

RANDOM FOREST

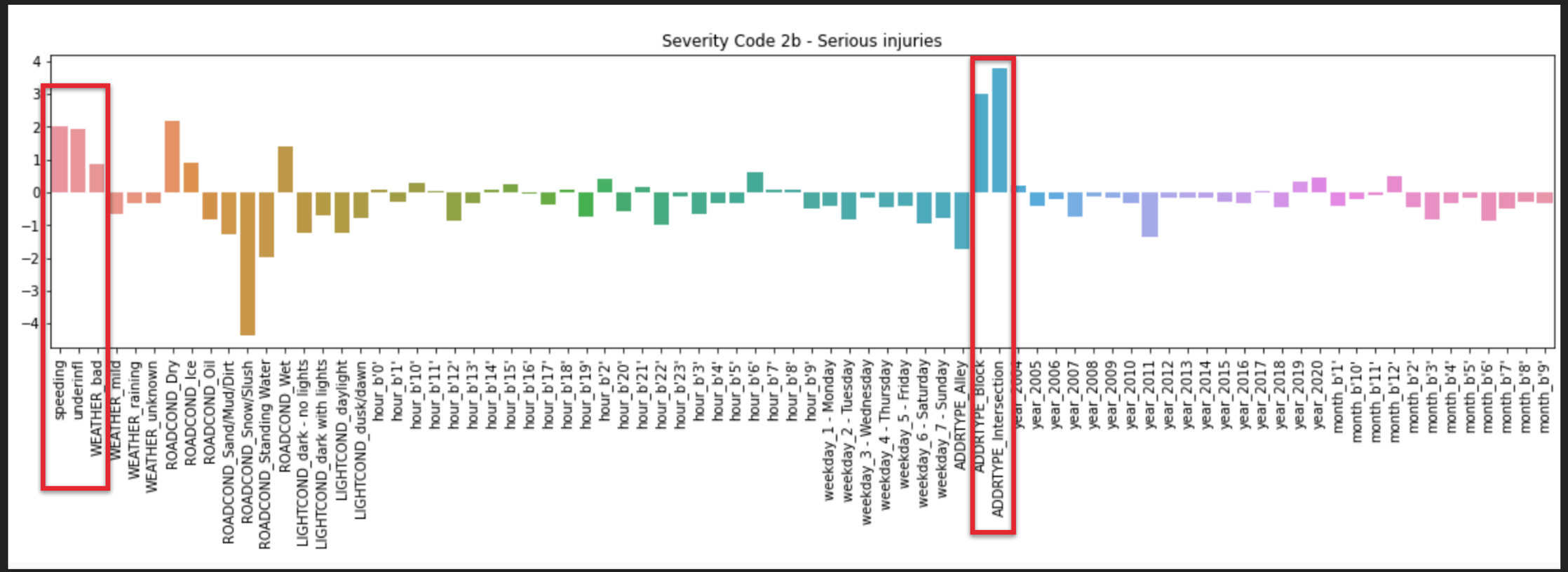
Classification Report: Random Forest

	precision	recall	f1-score	support
1	0.73	0.68	0.70	23348
2	0.42	0.46	0.44	11307
2b	0.04	0.07	0.05	586
3	0.02	0.05	0.03	83
micro avg	0.60	0.60	0.60	35324
macro avg	0.30	0.31	0.30	35324
weighted avg	0.62	0.60	0.60	35324

Accuracy: 59.52% (correctly classified test data)
F1: 0.604981 (weighted average of recall and precision)

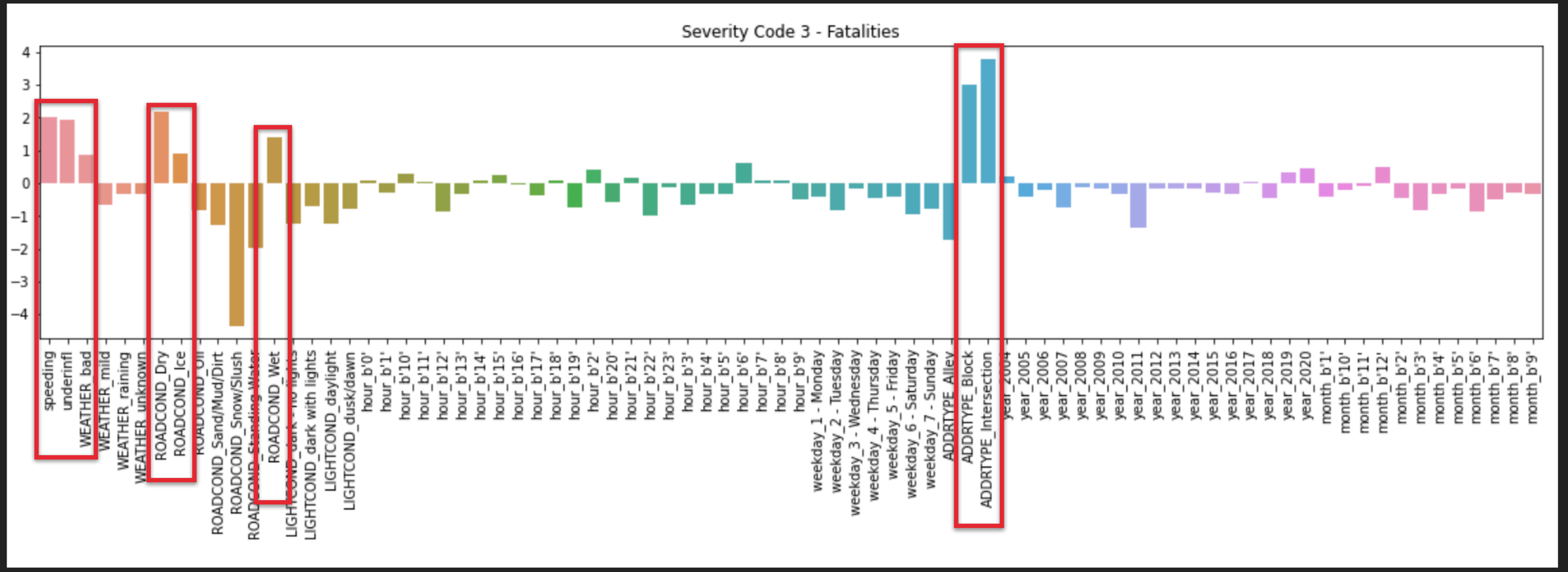
Random forest has better f1, but performs badly at high severity categories, hence we choose logistic regression

FEATURE IMPORTANCES: SEVERE INJURIES



Confounding factors: speeding, intoxication and blocks and intersections!

FEATURE IMPORTANCES: FATALITIES



Confounding factors: speeding, intoxication and blocks and intersections, as well as dry, icy and wet conditions.

KEY INSIGHTS

- Traffic regulators can try and introduce speed measurements and do drug texting in already at risk situations (e.g. intersections when it's raining)
- 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 still low: adjust weights to account for higher cost of misclassifying severer accidents.
- Introduce numerical scales for weather, light or road conditions from light to severe.