Crash Test Data Exploration

Brad Burkman

26 February 2021

1 Introduction

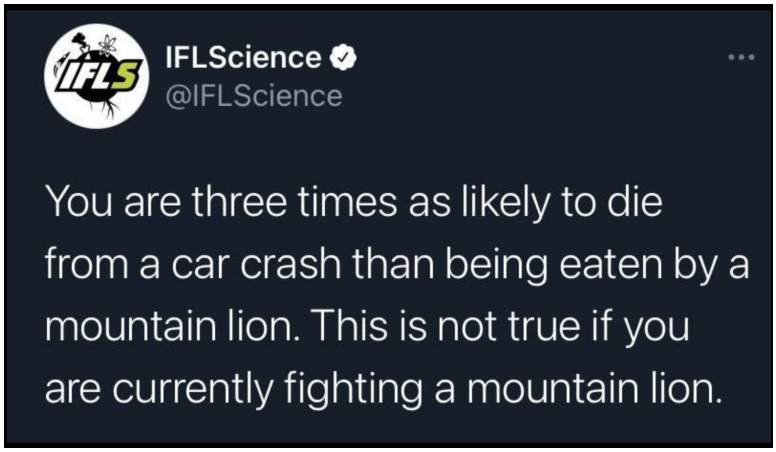
1.1 Goal

Given the `2019 Crash 1 Database.csv` from Dr. XiaoDuan Sun and her student Malek Abuhijeh, I'm trying different machine learning algorithms to make sense of the data.

1.2 Initial Question

Of the many factors in each crash, and the range of values for those factors, which ones most heavily correlate to the crash being fatal?

1.3 Requisite Nerdy Image of Negligible Relevance



Contents

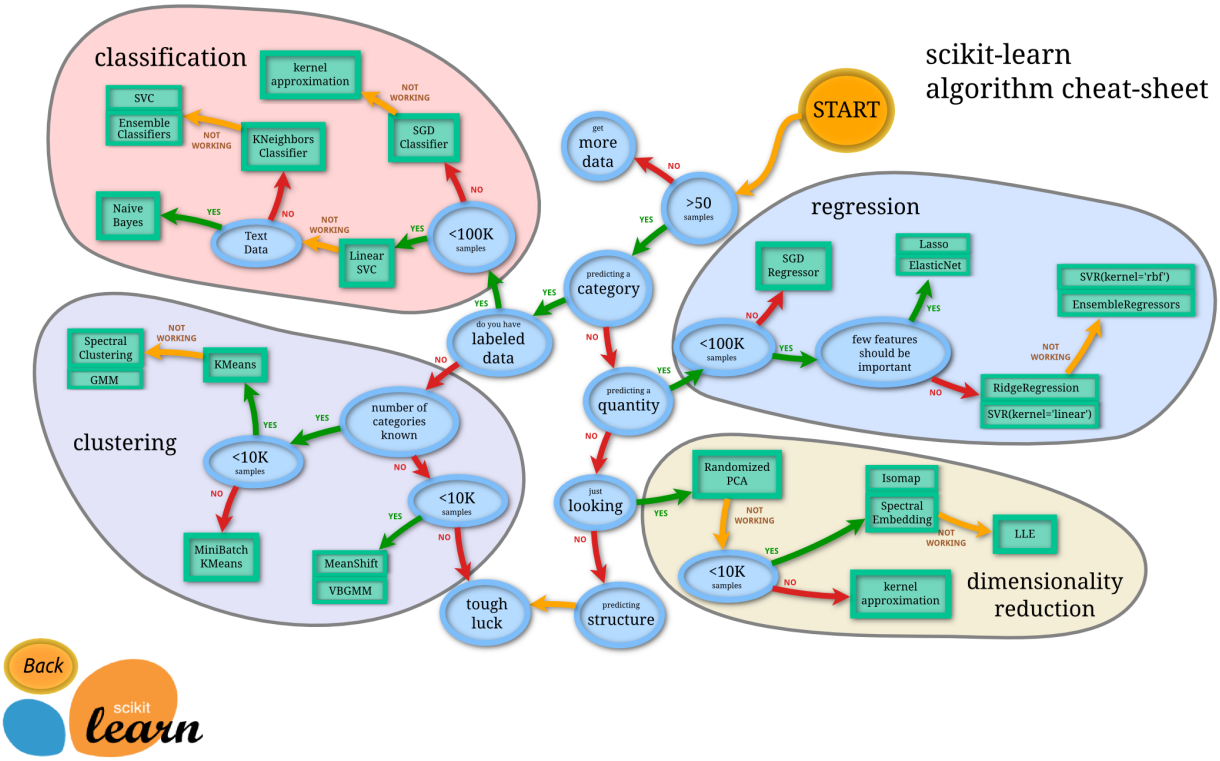
2 Data

The data was 160,186 records of crashes in Louisiana in 2019, with 79 fields per record. Some of the fields had a reasonable number of unique elements, like , like "Principal Contributing Factor" with thirteen unique values (and blank and "1," which didn't seem to mean anything), but others had many different elements, like "Principal Road Name." Presuming that we were more likely to find strong correlations in fields with a small number of unique values, I arbitrarily chose twenty unique values as the cutoff, tossing out the fields with more variety.

I used one-hot encoding to make each of the 415 remaining unique values into its own $\{0, 1\}$ vector.

3 Tools

I will start with scikit-learn, and work through its machine learning algorithm selection diagram.

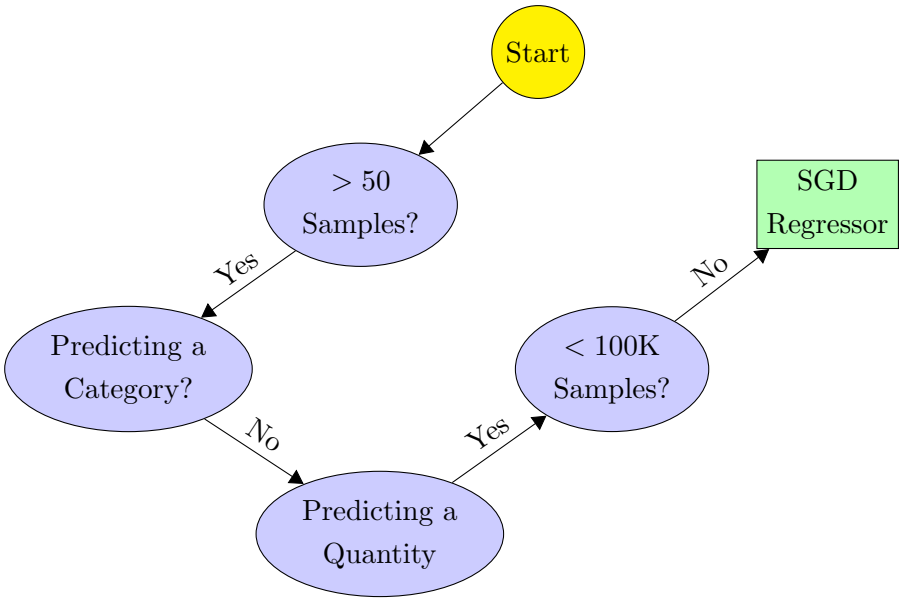


4 First Try: SGD Regressor

We're going to treat the severity of the accident as a real number in $[0, 1]$. The raw data set gives the severity in five categories, which we will convert to a real number.

Description	Raw Data	Feature-Engineered Data
FATAL	A	0.9
SEVERE	B	0.7
MODERATE	C	0.5
COMPLAINT	D	0.3
NO INJURY	E	0.1

The scikit-learn machine learning algorithm cheat sheet says the Stochastic Gradient Descent (SGD) Regressor algorithm is appropriate. We can also use SGDClassifier if we treat "FATAL" as 1 and the others as -1.



5 Output, Trying to Understand Linear Coefficients

It makes sense that the number of people killed in the accident being zero would negatively correlate with the accident being fatal, and there being more fatal crashes with one death than four, so the coefficients seem to be ordered according to our expectations.

coef_	Category	Value	Key
0.1585	num_tot_kil	1	
0.0163	num_tot_kil	2	
0.0039	num_tot_kil	3	
0.0005	num_tot_kil	4	
-0.1299	num_tot_kil	0	
0.0493	num_tot_kil	Sum	

It makes sense that head-on collisions would be most fatal, and left turns more deadly than right.

coef_	Category	Value	Key
0.0049	man_coll_cd	C	HEAD-ON
0.0042	man_coll_cd	A	NON-COLLISION WITH MOTOR VEHICLE
0.0042	man_coll_cd	J	SIDESWIPE - SAME DIRECTION
0.0042	man_coll_cd	B	REAR END
0.0042	man_coll_cd	Z	OTHER
0.0041	man_coll_cd	K	SIDESWIPE - OPPOSITE DIRECTION
0.0041	man_coll_cd	G	LEFT TURN - SAME DIRECTION
0.0041	man_coll_cd	E	LEFT TURN - ANGLE
0.0041	man_coll_cd	D	RIGHT ANGLE
0.0041	man_coll_cd	F	LEFT TURN - OPPOSITE DIRECTION
0.0040	man_coll_cd	H	RIGHT TURN - SAME DIRECTION
0.0031	man_coll_cd	I	RIGHT TURN - OPPOSITE DIRECTION
0.0493	man_coll_cd	Sum	

This one also satisfies expectations, because while snowy and icy conditions are dangerous, they are rare in Louisiana.

coef_	Category	Value	Key
0.0115	surf_cond_cd	B	WET
0.0114	surf_cond_cd	A	DRY
0.0100	surf_cond_cd	Y	UNKNOWN
0.0052	surf_cond_cd	2	
0.0031	surf_cond_cd	D	ICE
0.0029	surf_cond_cd		
0.0026	surf_cond_cd	E	CONTAMINANT (SAND, MUD, DIRT, OIL, ETC.)
0.0012	surf_cond_cd	Z	OTHER
0.0008	surf_cond_cd	C	SNOW/SLUSH
0.0007	surf_cond_cd	1	
0.0493	surf_cond_cd	Sum	

The driver's age doesn't seem to be correlated with the severity of the accident. Note that I changed the ages from years to decades to get fewer distinct values. Many records list the age as 200 years. (?)

coef_	Category	Value	Key
0.0051	dr_age_1	80	
0.0051	dr_age_1	70	
0.0051	dr_age_1	10	
0.0050	dr_age_1	30	
0.0050	dr_age_1	40	
0.0050	dr_age_1	60	
0.0050	dr_age_1	50	
0.0050	dr_age_1	20	
0.0042	dr_age_1	90	
0.0038	dr_age_1	200	
0.0008	dr_age_1	0	
0.0002	dr_age_1		
0.0000	dr_age_1	110	
0.0000	dr_age_1	100	
0.0493	dr_age_1	Sum	

The driver's gender also doesn't seem to play a role.

coef_	Category	Value	Key
0.0165	dr_sex_1		
0.0165	dr_sex_1	M	
0.0164	dr_sex_1	F	
0.0493	dr_sex_1	Sum	

6 Deflated Hopes

I had hoped that the sum of the coefficients for each category would be different, indicating which categories were had the strongest correlation to the severity of the accident, but they're all the same.

7 Request

Can you recommend a source I can read that will explain how to interpret the results? I emailed Henry, who taught me to use scikit-learn, but I have not heard back.

Thank you!