Capstone Project 1

In-depth Analysis (Machine Learning)

Greg Gibson 2020

As stated in prior reports, the dataset is located all together in one file, containing traffic accidents and weather metrics at the time of the event. The dataset has a four-level ranking of the severity of traffic caused by the accident. The goal is to condense the rankings into two levels, Low and High, or 0 and 1, and use a binary classifier, I hypothesize, to predict when the traffic could be severe based on weather conditions.

The Logistic Regression classifier was used first to produce the 0 and 1 outcomes. Categorical columns, Weather Group, Weekday, and Hour, were converted into binary dummy columns. Feedback also suggested turning low-variance columns Distance, Precipitation, and Visibility into binary to represent not-default / default values.

The final adjustment was to exclude Wind Chill as it correlated highly to Temperature and could interfere with the logistic regression. There are now 2.7M rows and 62 features.

The logistic regression was trained via a training subset of the data, with solver type set to 'saga' as a recommendation for larger datasets, and preprocessed with StandardScalar to balance the remaining float type columns. Testing accuracy was 69.4% on the first pass.

Feature Reduction

	coef
Traffic_Signal	-0.135802
Distance(Mi)	0.100135
Distancezero	-0.073732
Crossing	-0.061685
Daynight01	0.041559
Weekday_Sat	0.036909
Weekday_Sun	0.032850
Junction	0.032669
Precipitationzero	-0.026913
Hour_8	-0.021010
Pressure(In)	-0.020345
Weather_Group_Clear	-0.017879
Hour_7	-0.017450
Weekday_Tue	-0.017421
Weekday_Wed	-0.017111
Weekday_Mon	-0.015440
Wind_Speed(Mph)	0.014852
Station	-0.013350
Stop	-0.013228
Weekday_Thu	-0.012335

Of course, 62 features is unwieldy. The logistic regression is run again with an L1 regularization (Lasso) to reduce any unnecessary features' coefficients to zero. I print the features' coefficients via the .coef_ attribute.

Weather features did not make much of a showing! Only four features out of the first twenty are related to weather, starting at ranking nine.

Well, a third of the original features are +/- 0.01 away from zero and those will be kept for the next run, and the accuracy only declined 0.4%. Next, the model is evaluated.

Model Evaluation

A five-fold cross-validation yielded scores of 68.9, 69.3, 68.9, 68.5, and 68.5. The consistency shows the model is not over-fitting.

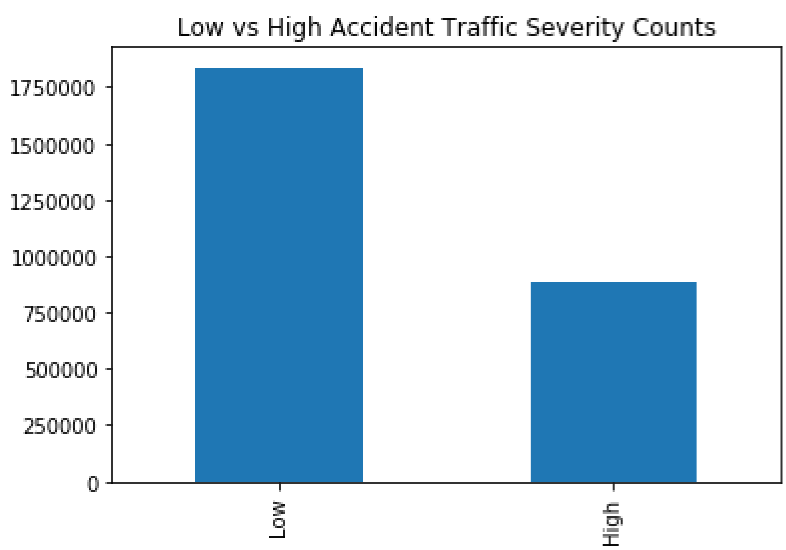
	precision	recall	f1-score	support
0	0.70	0.94	0.80	367866
1	0.57	0.16	0.25	175615
accuracy			0.69	543481
macro avg	0.64	0.55	0.53	543481
weighted avg	0.66	0.69	0.63	543481

The logistic regression has a great recall for Low (0) severity traffic, but only accurately identified High severity 16% of the time! Motorists cannot be sent into high severity traffic assuming it will be low. The logistic regression will be abandoned for a random forest classifier.

Random Forest's Turn

Training a random forest instead yielded better results as identifying High severity traffic improved to 44%.

Feedback noted the data is imbalanced with twice as many Low traffic severity accidents, and overtraining on identifying Low. Online research had suggested data balancing was for more extreme ratios, such as the medical field when very few patients have a rare condition.



Since there are plenty of records, a random selection of Low severity is chosen, in equal count to High severity: 880,347

Final Outcome

Data balancing made a significant improvement:

	precision	recall	f1-score	support
0	0.72	0.68	0.70	175525
1	0.70	0.74	0.72	176614
accuracy			0.71	352139
macro avg	0.71	0.71	0.71	352139
weighted avg	0.71	0.71	0.71	352139

The recall for High severity traffic reached 74%. It is feasible to use weather conditions to warn motorists when accident traffic may turn severe.