31 May 2021 Report

Brad Burkman

30 May 2021

Contents

1 Accomplishments This Week

- Almost finished a first pass through the *Accident Analysis & Prevention* articles.
- Remembered lessons from Dr. Aminul's class on accuracy, precision, and recall.
- Lots of ideas.
- One big idea.

2 Accuracy, Precision, Recall, Sensitivity, f1, False Alarm Rate

Consider the scenario where, in a large number of vehicles, if an airbag deploys, the vehicle's computer automatically contacts the police with vehicle description and location (and perhaps other data). The police want a system that will decide whether to automatically dispatch an ambulance. Ambulances are expensive and in finite supply, but we are willing to tolerate false positives (sending an ambulance when it's not needed) but not false negatives (not sending an ambulance when it is needed).

In a previous report I said I was getting 99% accuracy, and you agreed that it seemed suspicious.; I've now figured out why. I'm getting 99% accuracy because the dataset is so imbalanced. We have 681 fatal crashes out of 160,186, or 0.43%. If the ML model predicts that all of the crashes are non-fatal, it will get 99.57% accuracy. What we want to measure is *recall*, also called *sensitivity* and *detection rate*, which is the proportion of fatal crashes that we have correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

I'm amazed that most of the articles I'm reading talk primarily about accuracy. They may mention recall/sensitivity in Section 4 or 5, but the abstract, introduction, and conclusion only mention accuracy.

Confusion Matrix in SciKit-Learn.

		Prediction	
		N	P
Actual	N	TN	FP
	P	FN	TP

3 Big Idea: Limitations of Using Feature Selection Algorithms for Crash Data

A data set is *imbalanced* if the two classes of data (positive and negative) are not evenly distributed. In crash data, if we're looking for crashes among all car trips, crashes (positive class) are a tiny proportion (the minority class). If we're looking for fatal crashes among all crashes, the positive class, in our data set it's 681 out of 160,186, only 0.4%.

Do feature selection algorithms work with really imbalanced data sets? In the scikit-learn documentation for feature selection, in the first example, "suppose that we have a dataset with boolean features, and we want to remove all features that are either one or zero (on or off) in more than 80% of the samples." If we had a variable that perfectly predicted fatality, it would have zero in 99.6% of the samples.

In Roland [1], they use a classifier in scikit-learn to narrow the number of fields to fifteen. They eliminate Rain/cloudy/foggy/snow/clear but keep uvIndex and humidity.

It makes sense that accidents are more likely during rush hour, but less likely to be fatal during

rush hour, because most of the accidents are rear-end collisions in stop-and-go traffic. It makes sense that accident rates and accident fatality rates correlate to time of day.

The uvIndex and humidity also correlate strongly to time of day, and correlation is transitive (though it may be weaker or stronger), so uvIndex and humidity correlate to accident and fatality rates, but do those features give any new information?

The second example in the scikit-learn documentation for feature selection is univariate feature selection, which can select the fifteen variables that each correlate well to the target, but doesn't consider whether the chosen variables correlate well to each other.

What I think we want to find in feature selection are variables that correlate to the target, but not in the same way, so each feature contributes to our prediction.

Also with an imbalanced data set like crash data, we don't want feature selection that improves accuracy; we want to select a set of features that improves recall.

3.1 Test Data Set and Experiment

Create a balanced data set for classification with, say, 10,000 records, so 5,000 zeroes and 5,000 ones in the y-values. Create ten x-features of decreasing accuracy by starting with the y-values and randomly perturbing them to get 95%, 90%, ..., 50% accuracy. Then create some features with good recall but decreasing levels of accuracy. Take the y-values of $\{0, 1\}$, leave the 1's as 1 (or close), and perturb the 0's so that the accuracy falls to 95%, ..., 75%.

Run different feature selection algorithms and see which features get chosen.

Have two identical features and see if both get chosen.

Then recreate a similar data set, but imbalanced, starting with 9,900 zeroes and 100 ones in the target, and run the tests again.

4 Little Ideas

4.1 Train/Test Split for Imbalanced Data Set

For balanced data sets, the question in splitting into Train and Test data is whether you want a 50/50, 75/25, 80/20, or 90/10 split.

For this Louisiana crash data set, we have 160,168 records, only 681 of which are fatal crashes. We want to find needles in the haystack, but we don't have many needles. If we want an 80/20 split, should we make sure that the test set has 20% of the needles?

We could do that by first splitting the dataset into Positive and Negative sets, taking an 80/20 split of each set, then combining the splits into the Train and Test set. I've implemented that.

4.2 SMOTE: Synthetic Minority Oversampling TEchique

Used to balance an imbalanced dataset by creating new data points for the minority class.

"New synthetic data points are created by forming a convex combination of neighboring members." [2]

Might be something I should try.

4.3 Negative Samples

In Roland [1], in the lit review, they talk about previous authors creating *negative samples*. For each crash record, change one feature from {hour, day, unique road identification}, and see if there's an actual crash record with that set of features. If not, make it another record representing `not_a_crash`. The ideal proportions of positive (crash) and negative (generated non-crash) is a matter of debate.

Is the number of crashes on a stretch of road at a particular time of day proportional to the traffic volume? If we know the traffic volume, should we put in negative samples in such a way that the dataset reflects the traffic volume at each time of day?

For most roads, we don't have much data on traffic volume per hour. What kinds of roads have really good records? Toll roads. Unfortunately, Louisiana doesn't have any, but Texas does.

4.4 More on Feature Selection

In Roland [1], they make a weird argument for including variables it acknowledges are redundant, like Lat/Lon and Grid_Num, because they use different scales. (?)

Could we normalize some of the weather data to be above/below average at that time of day and that time of year? Or use a daily value rather than an hourly value? I find it hard to believe that uvIndex is more predictive than rain. Roland's article suggests future work making a single variable to represent the weather for the day, as was done in Hébert [3].

The Dark Sky API that Roland used has been bought by Apple and is no longer accepting new signups, but there are others that do the same thing. The remaining documentation gives a forecast by year, month, day, and hour, returning the information in this format:

```
{
'ozone': 290.06,
'temperature': 58.93,
'pressure': 1017.8,
'windBearing': 274,
'dewPoint': 52.58,
'cloudCover': 0.29,
```

```
'apparentTemperature': 58.93,
'windSpeed': 7.96,
'summary': 'Partly Cloudy',
'icon': 'partly-cloudy-night',
'humidity': 0.79,
'precipProbability': 0,
'precipIntensity': 0,
'visibility': 8.67,
'time': 1476410400
}
```


The documentation says there are other attributes that may or may not exist in a particular record, including uvIndex, windGust, precipAccumulation, and precipType.

Kathleen has an ecologist friend at USGS in Lafayette. We saw him Saturday and asked whether he could advise me on how to get weather data. He offered to help.

4.5 Citing Sources for Common Knowledge

Many of the papers I've read cite sources for common knowledge in the field, like the definitions of recall, precision, and accuracy, and statements like "driving is dangerous." It seems silly to me.

5 More Data Sources

5.1 California Statewide Integrated Traffic Records System (SWITRS)

`https://iswitrs.chp.ca.gov/Reports/jsp/index.jsp`

Apparently one can create an account and get data for research purposes.

5.2 Highway Safety Information System

Data from seven states, run by US Department of Transportation. I don't see anything more recent than 2018.

`http://www.hsisinfo.org`

6 Ideas, but not Good Ideas

6.1 Dangerous Drivers

Using the SHRP-2 data, rather than identify whether a crash has occurred or how serious the injuries are, can we identify the drivers more likely to be involved in a crash? We would not use the data at the moment of the crash, but the drivers' driving history in the study.

Insurance companies also have data like this. They have programs where drivers can get a lower rate if they use an accelerometer to demonstrate that they have safe driving habits. I haven't seen that data in these papers.

Tselentis et al did a study like this [4] where they categorized drivers, but they did not correlate it to crashes.

6.2 Google Maps Data?

The dataset we have has lots of information that's only available after the crash, not available in real time. What if we combined the following data streams and focused on a particular stretch of road that have frequent traffic slowdowns, like I-10 around the Mississippi River Bridge in Baton Rouge or I-10 through Kenner (near the New Orleans airport)?

- Google Maps (or similar) crash reports, which would give time and location.
- Other Google Maps reports, such as speed traps, slowdowns, construction, lane closures, stalled vehicles, and objects on road.
- Google Maps real-time data on traffic speed.
- Louisiana DOTD traffic cameras, which would give traffic volume and speed. A well positioned camera could give information about a reckless driver, either because of
  - Higher speed than the traffic
  - Tailgating
  - Sudden or multiple lane changes
- Current weather (Python API *Dark Sky*)
- Time of day and day of week.

We expect slowdowns 7-9am and 4-6pm that correlate to high traffic volume, but the interesting parts are when traffic is not heavy then suddenly slows down.

7 References

[Rol+21] Jeremiah Roland, Peter D. Way, Connor Firat, et al. "Modeling and predicting vehicle accident occurrence in Chattanooga, Tennessee". In: *Accident Analysis & Prevention* 149 (2021), p. 105860. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2020.105860. URL: https://www.sciencedirect.com/science/article/pii/S0001457520316808.

[Par+19] Amir Bahador Parsa, Homa Taghipour, Sybil Derrible, et al. "Real-time accident detection: Coping with imbalanced data". In: *Accident Analysis & Prevention* 129 (2019), pp. 202–210. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2019.05.014. URL: https://www.sciencedirect.com/science/article/pii/S0001457519301642.

[Heb+19] Antoine Hebert, Timothee Guedon, Tristan Glatard, et al. "High-Resolution Road Vehicle Collision Prediction for the City of Montreal". In: *2019 IEEE International Conference on Big Data (Big Data)* (Dec. 2019). DOI: 10.1109/bigdata47090.2019.9006009. URL: http://dx.doi.org/10.1109/BigData47090.2019.9006009.

[TVY21] Dimitrios I. Tselentis, Eleni I. Vlahogianni, and George Yannis. "Temporal analysis of driving efficiency using smartphone data". In: *Accident Analysis & Prevention* 154 (2021), p. 106081. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2021.106081. URL: https://www.sciencedirect.com/science/article/pii/S0001457521001123.