5 July 2021 Report

Brad Burkman

5 July 2021

This week I'll double down on my idea about balanced precision. It might not only give us better prediction power for injury/fatality, but it might be an original theoretical contribution to ML in general.

Contents

1 Activities this Week

- Started to read *Multivariate Methods*
- Thoughts about Feature Engineering
- Thoughts about our data set
- More thinking about Balanced Precision

2 Feature Engineering

2.1 Time of Day

Time of day is a continuous variable, but the correlation between time of day and [anything] is nonlinear. We could do some kind of data transformation, perhaps taking the ratio of the number of accidents to the typical traffic density at that time of day, but the typical car trip at 3 am on a Wednesday may be different in character than a car trip at 7 am on a Saturday, even if the traffic volumes are similar. Perhaps we should have boolean variables:

- Morning rush hour
- Mid-day
- Afternoon rush hour
- Evening
- Late night

and another variable, `Weekend`.

2.2 Number of Fatalities/Injuries

The number of fatalities or injuries is a function of how many people were in each vehicle, which (a) we don't know and (b) probably isn't correlated to any other data we have. Fatality and injury should be boolean variables, that there was a fatality or there was an injury, rather than a count of the number of fatalities or injuries.

3 Thoughts on our Data Set: Boolean and Metrics

3.1 Boolean Nature of our Data

Most of our data is boolean. Was alcohol involved? Did the car leave its lane? Was there a pedestrian? We have categorical variables, like type of vehicle which we represent as dummy (boolean) variables. We have some categories we could represent as numbers (like day of the week), and we could impose an order, (Monday comes before Tuesday), but the order isn't relevant in predicting injuries or fatalities, (Neither increases or decreases as the days "progress."), so we should represent them as categories, in dummy variables.

3.2 SMOTE: Synthetic Minority Oversampling TEchnique

Especially if we're doing fatalities, we have a terribly imbalanced data set. Ideally we'd like to have an equal number of fatal and nonfatal crashes to plug into our ML algorithm, but we have about 0.47% fatal and 99.53% nonfatal.

One solution is to randomly choose 681 nonfatal crashes to compare with our 681 fatal crashes, but that leaves behind a LOT of information.

Many of the papers I've read use SMOTE, which balances the data set by creating synthetic elements for the minority set (fatal crashes). It picks an element of the minority set, a, and picks one of its nearest neighbors, b, and creates a new synthetic element c. For each data category, D_i, in which they differ, SMOTE chooses $D_i(c)$ to be between $D_i(a)$ and $D_i(b)$. It randomly chooses a random number $r \in [0, 1]$, and makes $D_i(c) = D_i(a) + r(D_i(a) - D_i(b))$.

I get how that works for continuous variables. I get that it would work if $D_i(a)$ and $D_i(b)$ weren't very different.

How would that work for boolean variables? SMOTE would choose nearest neighbors a and b that agree on most variables, but for values of i where $D_i(a) = 0$ and $D_i(b) = 1$, it would randomly choose $D_i(c) \in \{0, 1\}$. There is no *between* for boolean variables. It doesn't seem to me that it would work as well.

[As I write this at my cabin in the woods, I don't have an internet connection and can't look it up.]

Update: Original SMOTE only works with continuous variables. There is something called SMOTE-NC that handles continuous and categorical, but it has to have some continuous variables to work on.

"Unlike SMOTE, SMOTE-NC for dataset containing numerical and categorical features. However, it is not designed to work with only categorical features."

`https://imbalanced-learn.org/dev/references/generated/imblearn.over_sampling.SMOTENC.html`

Since we have ≈ 200 times as many nonfatal crashes as fatal crashes, to balance the data set with SMOTE, we would have to make two hundred synthetic elements for each fatal crash. It seems to me that we would be making a mess of our data set.

3.3 Balanced Precision and Balanced f1 in the Penalty Function

Most ML algorithms work using a *penalty function* that measures how bad the current solution is, then iteratively improving the solution in the direction that minimizes the penalty. We should be able to write a custom penalty function.

Update: How-to instructions for changing the metrics in `scikit-learn`. The example is how to use recall instead of accuracy.

`https://stackoverflow.com/questions/54267745`

Recall only deals with the minority class, so the balance of the data set doesn't matter. *Precision*, on the other hand, takes results from both classes, so we can balance it by scaling the count of False Positive results, giving a *Balanced Precision* metric. From Recall and Balanced Precision we can get a *Balanced f1* metric.

If our penalty function uses balanced precision and balanced f1, it may not matter that our data set is imbalanced, and we can use all of, and only, the original data to build our model.

3.4 Balanced Precision in the Literature

Balanced Accuracy frequently appears in the literature. I have not found *balanced precision* in the literature. Two possible reasons. Either nobody has thought of it, or they did, found it not useful, and abandoned the idea.

Update: `imbalanced-learn` has more metrics than *scikit-learn*, but still no balanced precision. `https://imbalanced-learn.org/dev/metrics.html`

4 Thoughts on our Data Set: Trees

I suspect that a decision tree is the only realistic way to make a predict model for any aspect of crash data. If a pedestrian is involved, or it's a rural area, or alcohol is involved, the dynamics of the problem change. That there could be some linear (or nonlinear) function of all of the variables to fatality or injury is not reasonable to hope. If we think of it not as one big problem but as lots of little problems, like "What factors predict a fatality/injury in a crash involving a pedestrian in a rural area at night?" and, "What factors predict a fatality/injury in a crash where alcohol is involved at rush hour in an urban area?", we'll have much more likelihood of success.