Thesis Proposal

Predictive Modeling with Imbalanced Data?

Alyssa Forber

Kathryn Colborn, Elizabeth Juarez-Colunga, Susan Calcaterra

November 28, 2017

Abstract

Do I need an abstract?

Introduction

- Don't go into background problem too much? (for stats paper)
- Do I open on the imbalanced learning problem?
- How much to describe the dataset?

Predictive modeling with imbalanced data has been found to have report low sensitivity (reference?). To combat this issue of overlooking many true postitives,

To illustrate this issue, we are using electronic health record data from 2008 to 2014 of patients for patients with chronic opioid therapy (COT). Five percent of the 27,705 patients were reported with the outcome, which was defined as receipt of $\geq$ 90-day supply of opioids with $<$ 30-day gap in supply over a 180-day period or receipt of $\geq$ 10 opioid prescriptions over one year.

Methods

The analysis was done in RStudio version 1.1.383.

We used a roughly 2/3rd temporal split of the data to create training and testing datasets, where years 2008-2011 were used to train (65%), and 2012-2014 were used to test (35%).

The model used for this analysis was cross validated lasso regression. This was chosen as it has been found to perform better predictor selection than stepwise selection (reference on this?), and as we were not interested in having interpretable coefficients.

The predictors were first narrowed from ? to 35 (?) based on clinical relevance (Can I reference the paper that is under submission since that goes into more details? Is it even necessary).

We first evaluated the prediction performance of the dataset without sampling to see the effects of the imbalanced data on the accuracy, sensitivity, and specificity. This was to serve as a baseline to compare with the techniques available to mitigate the issue of poor sensitivity. The predicted probability cut-point used here was rounding at the standard 0.5 that would be appropriate in balanced datasets.

The first approach used to improve performance was to choose a more informed probability cut-point for the data. This was done using the Youden Index, which finds the maximum of the receiver operating characteristic (ROC) curve (reference here!) with the pROC package (do I need to say this?).

The second approach was through sampling the dataset. Three types of sampling methods were compared–down sampling, up sampling, and Synthetic Minority Over-sampling Technique (SMOTE). Down sampling takes a random sample from the majority class, in this case those who are not classified as having chronic opioid therapy, in order to match the size of the minority class (reference?). Up sampling does the reverse to take random samples of the minority class in order to match the majority (reference?). SMOTE combines sampling both from the majority and minority, but instead of taking identical copies of the minority it creates synthetic observations. For each of the three sampling techniques, the probability cut-point was optimized using the Youden Index as before.

Results

As expected, without using an optimized cut-point or sampling technique, the sensitivity of the model was extremely poor at 8%, with high specificity and accuracy (99% and 96%). Simply choosing a more informed probability cut-point to 0.043 instead of 0.5 improved the sensitivity to 85% and brought the specificity down to 73%. This cut-point is intuitive as the outcome is present at 5% in the dataset, which would be consistent with a 0.5 cutoff in a evenly split dataset. The up and down sampled datasets both showed the same improved sensitivity with probability cut-points at about 0.4, also with close specificities of 74 and 73%. SMOTE on the other hand, while still seeing improvements on sensitivity, had reversed values with 74% sensitivity and 84% specificity. However, there were improvements in accuracy for SMOTE at 86% as compared to the other three approaches, which had accuracies at 86-87%.

There was no change to the negative predicted value across the approaches, and a decrease in positive predicted value. In terms of the ROC analysis, the area under of the curve for each approach was about the same at 86-87%. See Table 2 for full results for the cut-point, sensitivity, specificity, accuracy, negative predicted value, positive predicted value, and area under the curve.

- Are we really comparing sampling to cut-points if we use both in the second approach?
- Doing both sampling and cut-points is overkill if just cut-points works, so it would maybe make sense to just compare them individually?
- Also I'm sure we could get matching results with SMOTE if we wanted to see them compare?

Discussion

Across the different samples, we saw similar improvements for sensitivity.

- Again are we really comparing cut-points and sampling if we do both in second approach? What do we say we're finding here

Conclusion

Acknowledgments

Do I just include committee members as authors and then acknowledge the University?

KL Colborn PhD

E Juarez-Colunga PhD

SL Calcaterra MD, MPH

References

ROC, Youden, SMOTE, LASSO, cross-validation Chronic Opioid Therapy

A Statistical Model for Prediction of Future Chronic Opioid Use among Hospitalized Patients

Appendix

Include full table 1?

Table 1:

Variable	Yes COT	No COT	p-value
	1,457 (5%)	26,248 (95%)	
Age 15-35	10%	22%	<.001
Age 45-55	35%	24%	<.001
Age 55-65	28%	21%	<.001
Discount payment or Medicaid	76%	61%	<.001
History of chronic pain	76%	53%	<.001
Discharge diagnosis chronic pain	50%	29%	<.001
Surgical patient	48%	39%	<.001
Past year:			
Benzodiazepine	16%	5%	<.001
Non-opioid analgesics	25%	9%	<.001
Number of opioid prescriptions:			
0	38%	80%	
1	17%	11%	
2	14%	4%	
3	9%	2%	
4-9	23%	3%	<.001
Receipt of opioid at discharge	56%	28%	<.001
MME per hospital day > 10	80%	52%	<.001

Table 2:

Data	Threshold	Specificity	Sensitivity	NPV	PPV	Accuracy	AUC
Unsampled 0.5	0.5	99	8	96	35	96	86
Unsampled	0.043	73	85	99	12	73	86
Down sampled	0.401	73	85	99	12	74	86
Up sampled	0.399	74	85	99	12	74	87
SMOTE	0.472	84	74	99	17	84	86

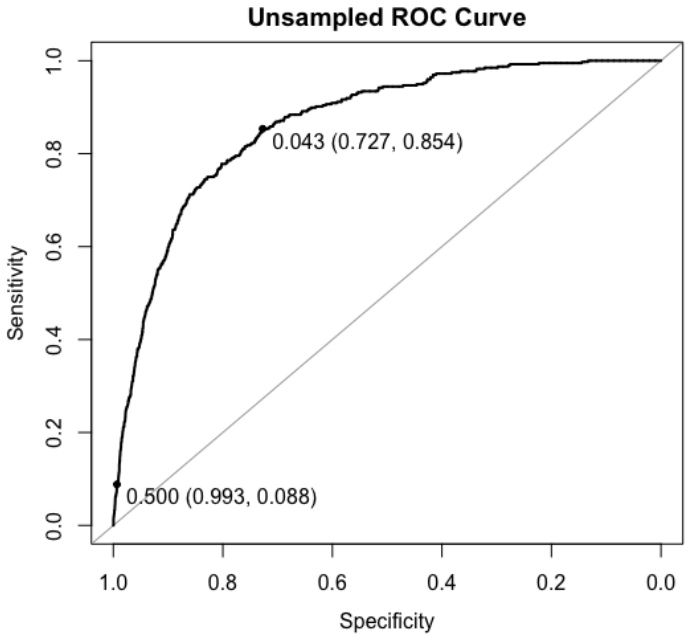

Figure 1: ROC for Original Data: Younden and 0.5 cutoffs