Thesis Proposal

Predictive Modeling with Imbalanced Data?

Alyssa Forber

University of Colorado, Anschutz Medical Campus

November 26, 2017

Abstract

Do I need an abstract?

Introduction

- Unbalanced learning problem (low sensitivity in prediction)
- Decribe chronic opioid therapy issue and dataset
- Don't go into background problem too much? (for stats paper)
- Focus more on the imbalanced learning probelm
- But do I open on the opioid problem or the learning problem?**
- How much to describe the dataset?

Methods

The analysis was done in RStuido 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, sensitivy, 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 cutpoint 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 probabilty cutpoint 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 classfied as having chronic opioid therapy, in order to match the size of the minority class (referenc?). 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

cutpoint was optmized using the Youden Index as before.

Results

See table 2 for threshold (cutoff), sensitivity, specificity, accurary, npv, ppv, and AUC

Discussion

Conclusion

Acknowledgments

Do I just include committee as authors and then acknowledge the University? Or whoever is funding me?

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 1,457 (5%)	No COT 26,248 (95%)	p-value
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

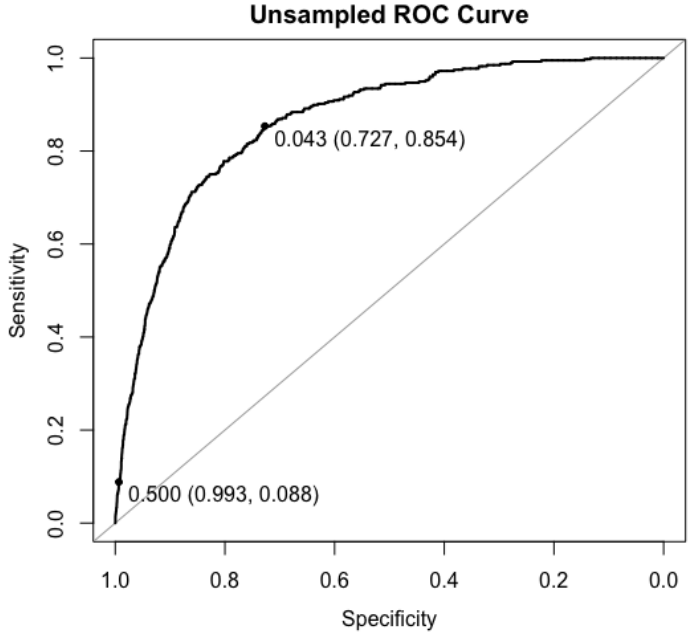

Figure 1: ROC for Original Data: Younden and 0.5 cutoffs