Thesis Proposal

Predictive Modeling with Unbalanced Data?

Alyssa Forber

University of Colorado, Anschutz Medical Campus

November 20, 2017

Abstract

Do I need an abstract?

Introduction

- Unbalanced learning problem (low sensitivity in prediction)
- Decribe chronic opioid therapy issue and dataset

Methods

- Roughly 2/3 temporal split of data to get train and test set
- splitting 08-11 and 12-14 to make 64.7% train, 35.3% test
- Cross validated lasso regression (and bagging?)
- Lasso:

    - Advantages:
        * Lower variance of the predicted values
        * More accurate predictions
        * Reduces the number of predictors
    - Disadvantages:
        * No interpretation of predictor coefficients
        * No standard errors out of the model
        * Biased coefficients

ROC curves and cutoff (with pROC package):

- Youden Index

1. No Sampling, Optimze Cut-off:

- Use 0.5 standard probability cutoff to compare
- Compare to Youden Index cutoff

2. Sampling:

- Create sampled data sets that are balanced
    - Down sample
        * under-sample majority to equal minority
    - Up sample
        * over-sample minority to equal majority
    - SMOTE
        * Synthetic Minority Over-sampling Technique
- Predict and use Youden Index as cutoff

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

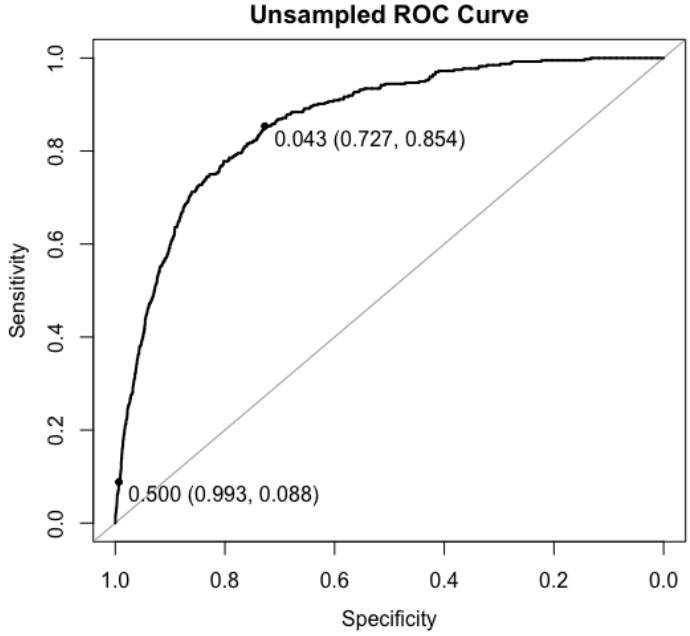

Figure 1: ROC for Original Data: Younden and 0.5 cutoffs