

# Thesis Proposal

## Predictive Modeling with Imbalanced Data

Alyssa Forber

University of Colorado, Anschutz Medical Campus

December 2017

# Outline

Background

Problem

Objectives

Methods

Preliminary Results

Moving Forward

# Motivation

- Chronic opioid therapy has become an epidemic
- Over 2 million people had a prescription opioid use disorder (2015 National Survey of Drug Use and Health)
- Important to identify patients at high risk
- Allow for hospitals to make informative decisions about prescribing opioids

# Imbalanced Learning Problem

- Presents a problem of imbalanced data
- Poor sensitivity with rare outcomes
- Need to improve predictive performance

# The Data

- Design: Denver Health retrospective analysis electronic health record (EHR) data from 2008 to 2014.
- Patients: Hospitalized patients at an urban, safety-net hospital.
- Definition of Chronic Opioid Therapy (COT) one year following the index hospital discharge:

*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.*

# Patient Population

- 27,705 patients
- Majority had incomes <185% of the Federal Poverty Level
- 70% were ethnic minorities
- 5% with COT
- Excluded Patients:
  - ▶ <15 or >85 years old
  - ▶ Those in prison, jail, or police custody
  - ▶ Those who died within one year following their index hospitalization
  - ▶ Patients with <2 healthcare visits to Denver Health three years preceding their index hospitalization
  - ▶ Undocumented persons receiving emergent hemodialysis
  - ▶ Obstetric patients

# 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
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

# Aims

- Accurate predicting → improving sensitivity and specificity for imbalanced outcome
- Using and comparing methods of probability cutpoints and sampling



# Methods

- Create sampled datasets
- Run model on sampled data
- Get predicted probabilities on the test data
- Optimize probability cutoff for outcome

# Model

- Roughly 2/3 temporal split of data to get train and test set
- Cross validated lasso regression

- Lasso:

- ▶ Performs variable selection
- ▶ Shrinks estimates to 0

$$\hat{\beta}_{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$$

subject to  $\sum_{j=1}^p |\beta_j| \leq t$  where  $t$  is the tuning parameter.

# Cross Validation

- Cross validation:
  - ▶ Find the best “tuning measure” for model selection which determines amount of shrinkage of estimates
  - ▶ Split data into  $k$  parts and then train on each of those except one you validate against
  - ▶ Then pick the tuning measure that minimizes error

# Advantages and Disadvantages

- 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 (with pROC package):

- ROC curve plots sensitivity vs specificity
- Top left corner is ideal
- Youden Index is the furthest upper left corner or “max”

# ROC Curve

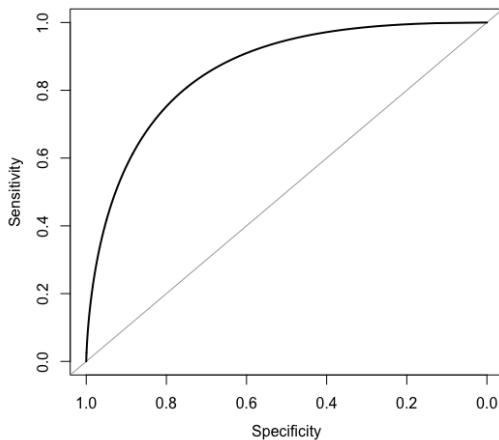


Figure 1: ROC Curve

# Confusion Matrix

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

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)



# First approach

## No Sampling, Optimize Cut-off:

- Use original unsampled data and get predictions off the lasso model
  - ▶ Predictions return probability between 0 and 1 for each observation
- Use 0.5 standard probability cutoff to compare
- Find “best” probability cutoff
  - ▶ Youden Index

# Second Approach

## Sampling:

- Create sampled data sets that are balanced
  - ▶ Down sample
  - ▶ Up sample
  - ▶ SMOTE
- Predict and use both standard 0.5 and Youden Index as cutoff

# Down Sampling

- Under-sample majority to equal minority

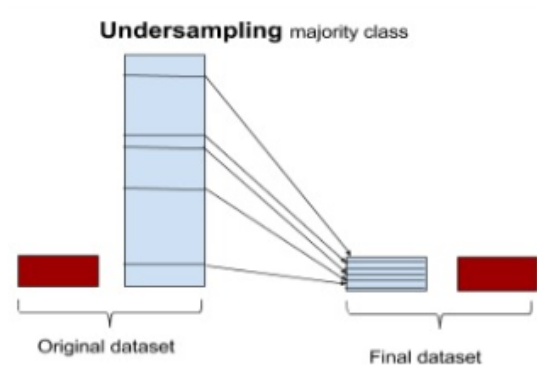


Figure 2: Down sample

# Up Sampling

- Over-sample minority to equal majority

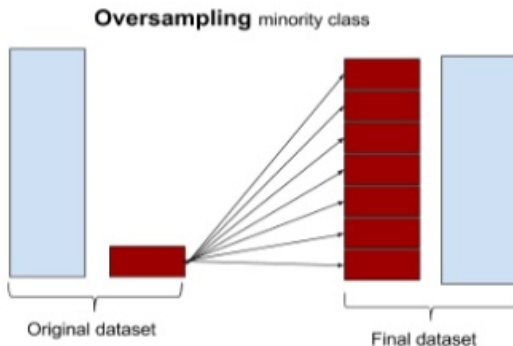


Figure 3: Up sample

# SMOTE

- Synthetic Minority Over-sampling Technique

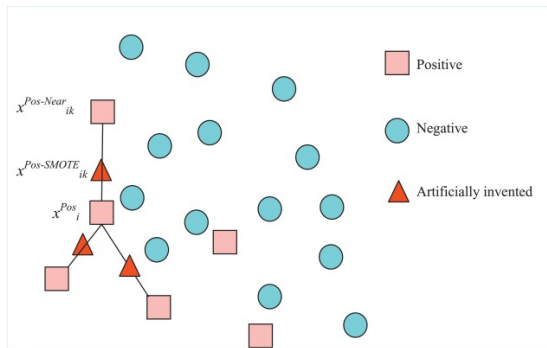


Figure 4: SMOTE

Table 1: Results

Data	Threshold	Specificity	Sensitivity	NPV	PPV	Accuracy	AUC	Covariates
Unsampled 0.5	0.5	99	8	96	35	96	86	31
Unsampled	0.043	73	85	99	12	73	86	31
Down sampled 0.5	0.5	81	75	99	15	81	86	34
Down sampled	0.401	73	85	99	12	74	86	34
Up sampled 0.5	0.5	82	75	99	15	82	87	34
Up sampled	0.399	74	85	99	12	74	87	34
SMOTE 0.5	0.5	86	71	99	17	85	86	33
SMOTE	0.472	84	74	99	17	84	86	33

# ROC Plot

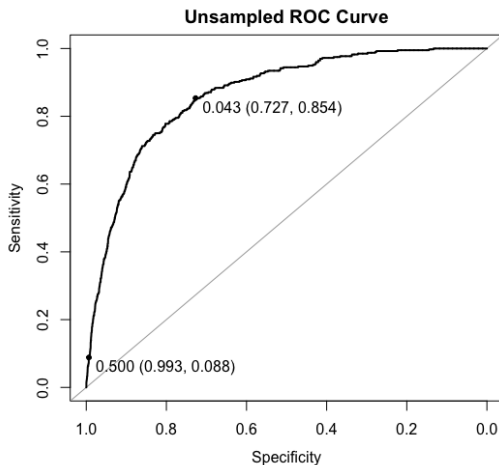


Figure 5: ROC for Original Data: Younden and 0.5 cutoffs

# Conclusions Thus Far

- Seeing similar results for both methods
- Depending on situation the clinician may like different sensitivity/specificity
- Some may want to be more conservative, others may not
  - ▶ Example: cancer patients in a significant pain



# Moving Forward

- Bagging (bootstrap aggregating)
- Simulation of different percentages for rare outcomes
  - ▶ explore method performance at 5%, 10%, 50% ect. of outcome
- Investigate other sampling techniques or cut-point methods

# Timeline

Defend in March

# Questions?

Questions or Suggestions?

# References

[https://rasbt.github.io/mlxtend/user\\_guide/evaluate/confusion\\_matrix/](https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/)

<https://svds.com/learning-imbalanced-classes/>