# Thesis Proposal

## Predictive Modeling with Imbalanced Data

Alyssa Forber

University of Colorado, Anschutz Medical Campus

December 2017

# Outline

# Imbalanced Learning Problem

- Predictive models learn poorly when datasets are imbalanced
- Over learning the majority and under learning the minority
- Results in low sensitivity and high specificity
- Need to improve predictive performance
- Many examples in medicine, diseases or adverse reactions taking place in small percent of the population

# Aims

- Improve predictive performance for imbalanced dataset
- Utilizing measures of sensitivity, specificity, accuracy, and ROC analysis to evaluate performance
- Use and compare two methods to handle imbalance
  - Informed robability cutpoints for predicted probabilites
  - Sampling techniques to balance datasets
- Evaluate methods to recommend approaches in medicine

# Methods

- Split data into training and test sets
- Create balanced datasets through sampling on test set
- Run predictive model on sampled data
- Get predicted probabilities on the hold-out test data
- Optimize probability cutoff for outcome

# Model

Cross-validated lasso regression - Least Absolute Shrinkage and Selection Operator

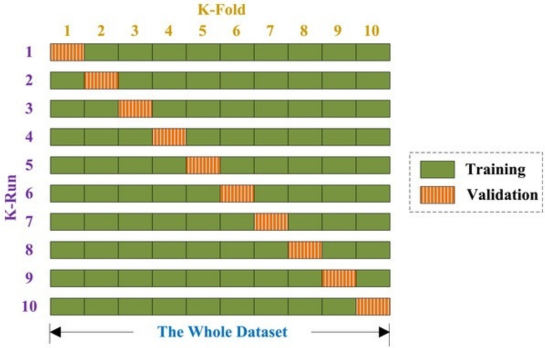
- Lasso:
  - Shrinks estimates by penalizing size of coefficients
  - Performs variable selection by shrinking some to 0

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

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:
  - Biased coefficients, inference not same as logistic regression
  - No standard errors or p-values out of the model

# ROC & Cutoff Probabilities

ROC (with pROC package):

- Receiver Operating Characteristics
- ROC curve plots sensitivity vs specificity
- Each point on curve corresponds to a decision cutoff
- Youden's Index calculated the furthest upper left corner or "max" on curve
- Area under the curve (AUC) should be maximized

# ROC Curve

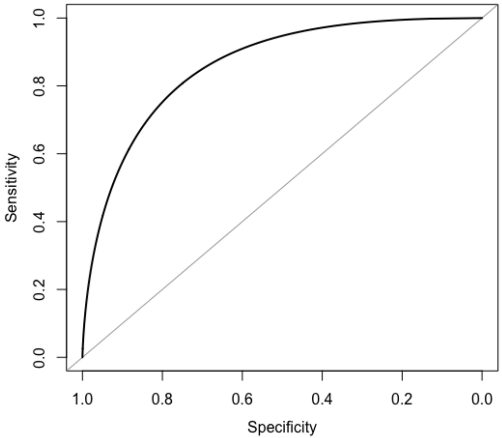



Figure 2: ROC Curve

# Confusion Matrix

**Correctly identify those w/ outcome:**

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

**Correctly identify those w/o outcome:**

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

**Correctly identify either group:**

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

| | | **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 from 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's Index

# Second Approach

Sampling:

- Create sampled data sets that are balanced
  - Down sample
  - Up sample
  - SMOTE
- Predict and use Youden's Index as cutoff

# Down Sampling

- Under-sample majority to equal minority

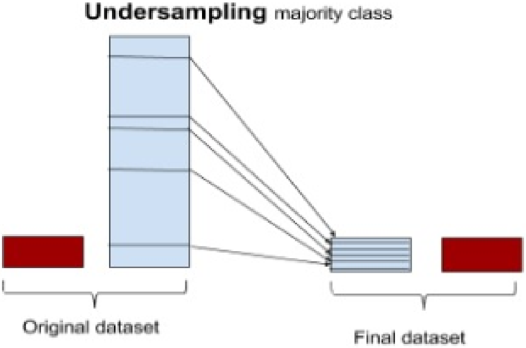



Figure 3: Down sample

# Up Sampling

- Over-sample minority to equal majority

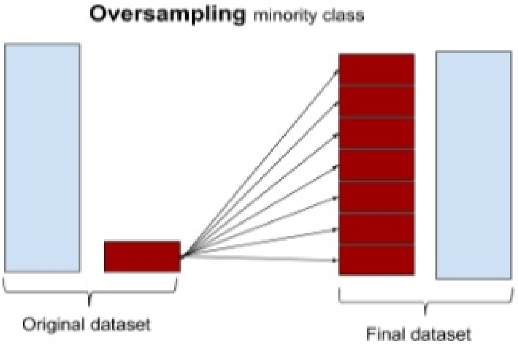



Figure 4: Up sample

# SMOTE

- Synthetic Minority Over-sampling Technique

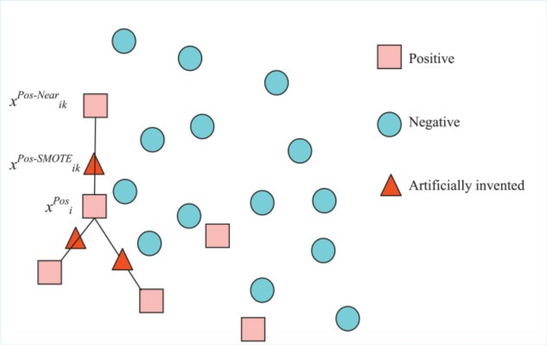



Figure 5: SMOTE

# Models

- Model 1: No Sampling
  - 0.5 cutoff
  - Youden's Index
- Model 2: Down Sampling
  - Youden's Index
- Model 3: Up Sampling
  - Youden's Index
- Model 4: SMOTE
  - Youden's Index

# Case Studies

We used these methods on two case studies to see how results may differ between two real world examples with rare outcomes.

- Case Study 1: Predicting chronic opioid therapy in hospitalized patients (5%)
- Case Study 2: Predicting surgical site infections in hospitalized patients (3.4%)

# Case Study 1

- 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.*

# Case Study 1: 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

# Case Study 1: Results

Table 1: Results

| Model | Threshold | Sensitivity | Specificity | NPV | PPV | Accuracy | AUC | Covariates |
|---|---|---|---|---|---|---|---|---|
| **Unsampled 0.5** | 0.5 | 8 | 99 | 96 | 35 | 96 | 86 | 31 |
| **Unsampled** | 0.043 | 85 | 73 | 99 | 12 | 73 | 86 | 31 |
| **Down sampled** | 0.401 | 85 | 73 | 99 | 12 | 74 | 86 | 34 |
| **Up sampled** | 0.399 | 85 | 74 | 99 | 12 | 74 | 87 | 34 |
| **SMOTE** | 0.472 | 74 | 84 | 99 | 17 | 84 | 86 | 33 |

# ROC Plot

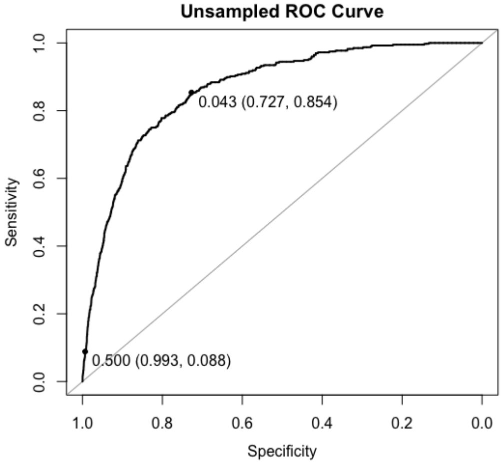



Figure 6: ROC for Original Data: Younden and 0.5 cutoffs

# Case Study 2

# Case Study 2:

# Case STudy 2: Results

# Simulation

# Simulation Results

# Conclusions

# Questions?

Questions