Data Mining and Modeling Methods for Site Inspection Selection

Elena Rantou, Ph.D.

Paul Schuette, Ph.D.

September 14, 2018

Disclaimer

This presentation reflects the views of the presenter and should not be construed to represent the United States Food and Drug Administration’s views or policies.

Outline

- Motivation
- Objectives and background
- Data sets and structures
- Challenges
- Methods and their performance
- Other considerations

Motivation

In a clinical trial setting, data reliability can be jeopardized by:

- Poorly Collected data
- Poorly Processed data
- Poorly Reported data
- Tampered or Fraudulent data

The number and complexity of clinical trials have risen dramatically making it difficult for regulators to choose clinical sites for inspection

Objectives

To determine whether

- supervised data mining methods can be used to predict site inspection results
- unsupervised statistical monitoring

can be used to identify ‘unusual’ clinical sites for inspection (*ongoing work*)

Objectives

Onsite inspections help ensure the integrity of the clinical trials via source data verification

Due to limited resources only less than 1% of the sites can be inspected annually. It is therefore crucial to select the appropriate clinical sites

Data sets

Site inspection results can be classified into:

- NAI (No Action Indicated)
- VAI (Voluntary Action Indicated)
- OAI (Official Action Indicated)

Data sets

Clinical trial data and the results from clinical site Inspections

Response

can be:

- Ordinal with three distinct classes (OAI, VAI, NAI)
- Binary: 2 of 3 ordinal classes are suppressed to 1 (VAI, OAI) vs. NAI

Challenges (ordinal response)

Missing data

Assumptions: missing values are MAR and can be predicted by observed values

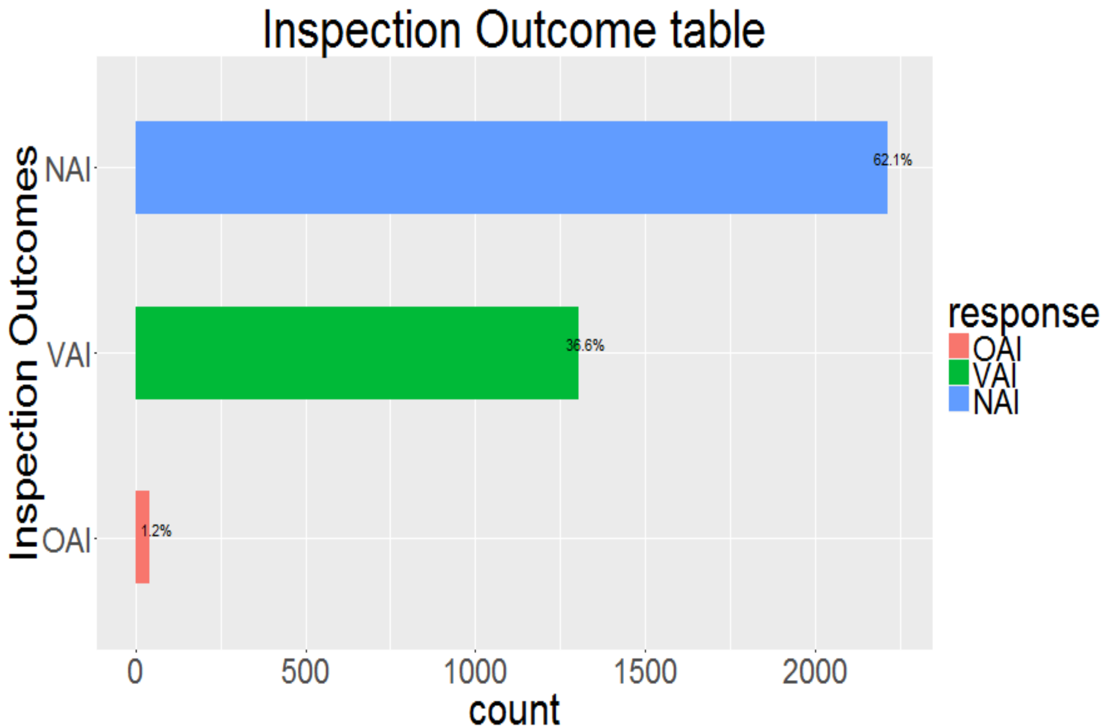
Random Forest (RF) imputation

- Replace missing values with sample median
- Use RF to compute proximity between missing and non-missing samples
- Repeat

Variable	Type	% missing
Enrollment	continuous	
Site Specific Efficacy	continuous	27.7%
Protocol deviation	continuous	
NS adverse event	continuous	
% subject death	continuous	
Enroll/Screen %	continuous	
Subject discontinuation	continuous	
Number of INDs	continuous	
Financial disclosure	continuous	29.9%
Complaint history	Binary	
Time since last inspection	continuous	4.32%
OAI history	Binary	

Challenges (ordinal response)

Imbalanced outcomes-OAI classification is a rare event with only 1% of sites being classified as OAI



Challenges (ordinal response)

Synthetic Minority Over-Sampling Technique-SMOTE

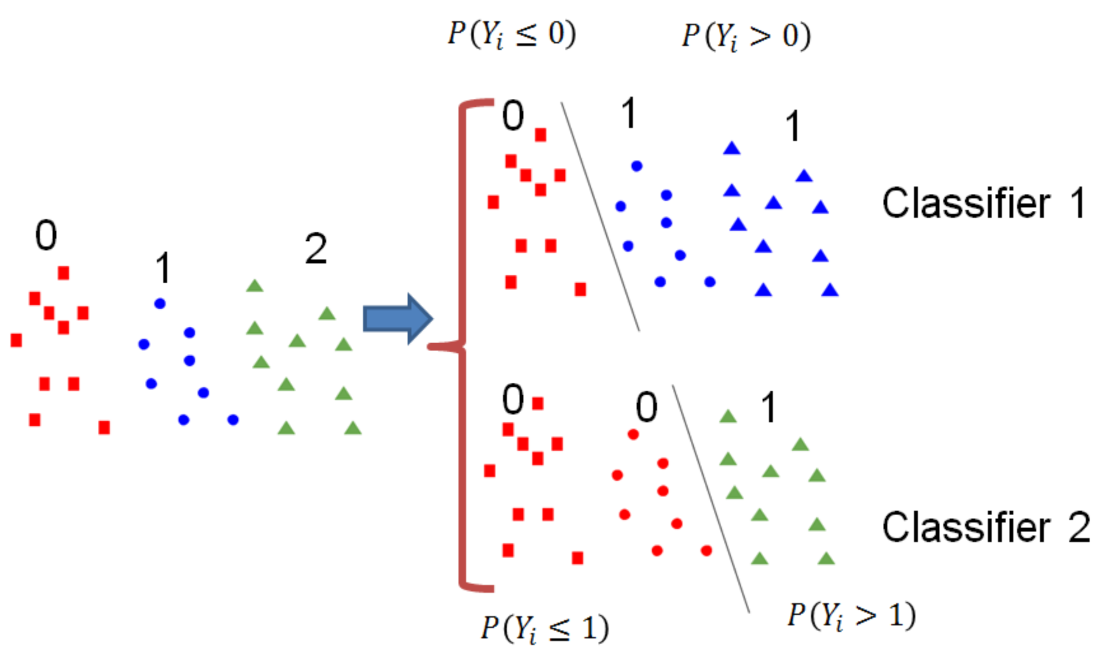
- Generate synthetic samples for the minority class
- Input the number of nearest neighbors, k, T minority class samples and size of SMOTE, N
- Output is the synthetic minority class samples

Statistical methods (ordinal response)

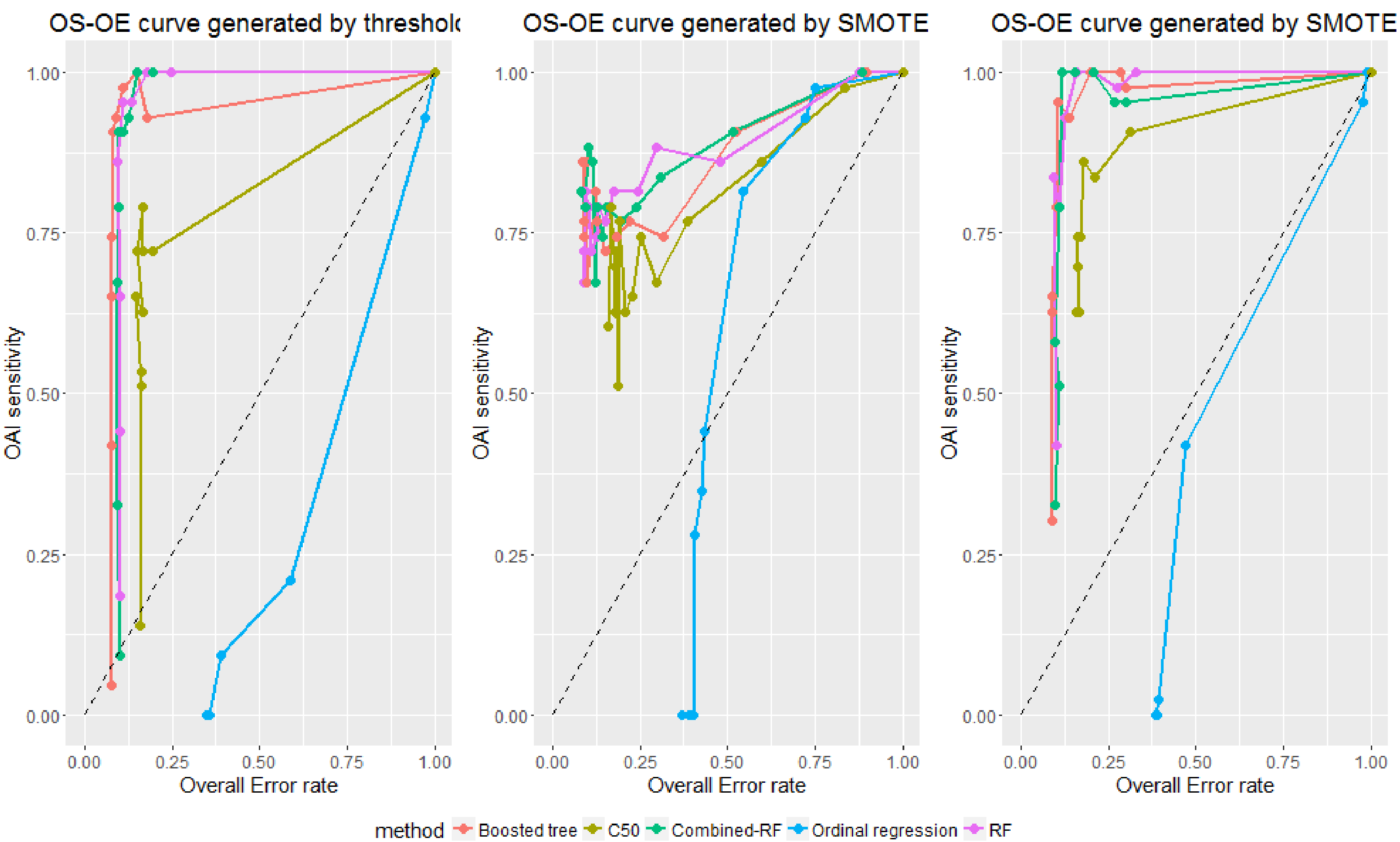
- Ordinal regression
- Combined binary classifiers
- Random forests
- Boosted trees

Combined binary classifier

Convert an ordinal regression problem into nested binary classification problems by splitting the data into groups $Y_i \leq j$ and $Y_i > j$ and a binary probability classifier to estimate the probabilities $P(Y_i \leq j)$ and $\mathrm{P}(Y_i > j)$



Classifier performance



Statistical methods (binary response)

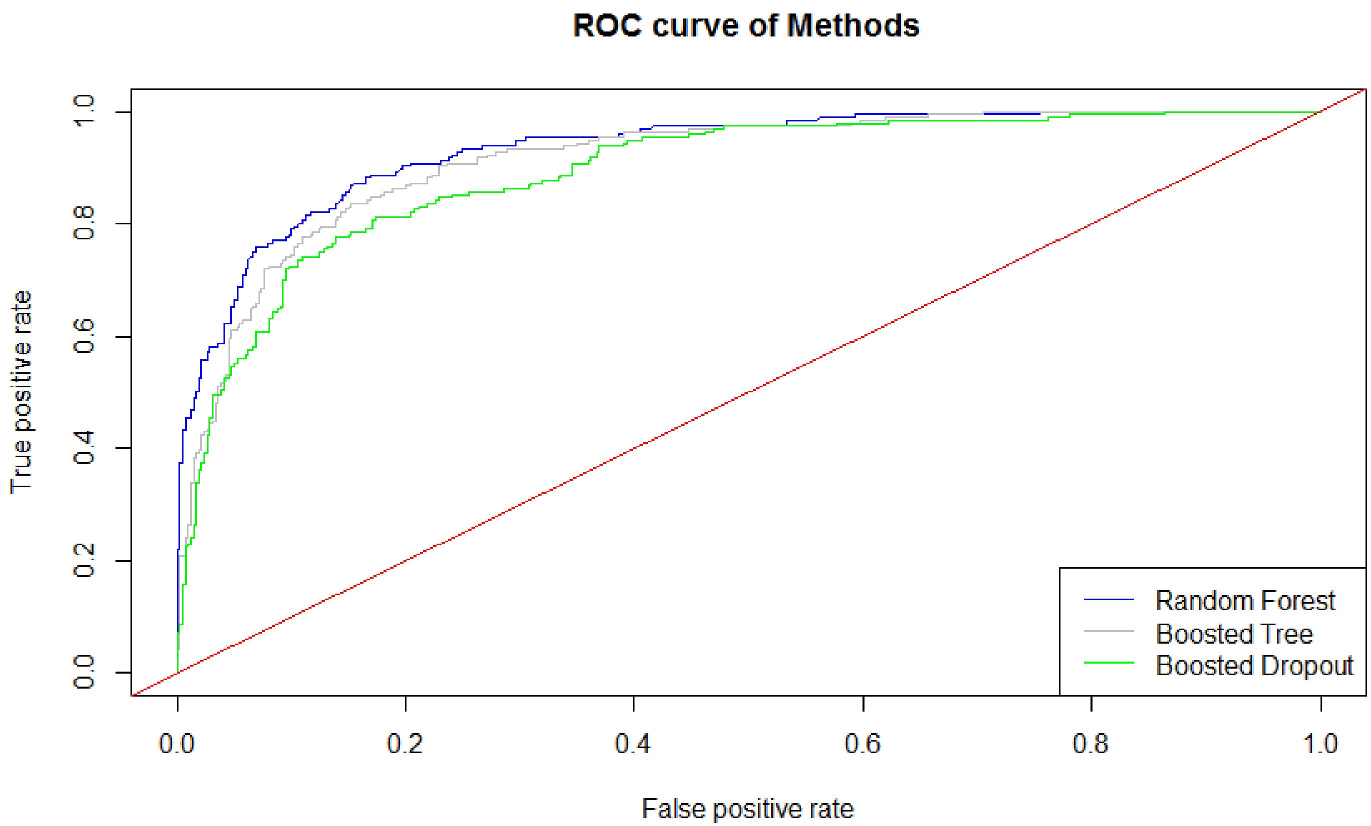
- Random Forest
- Boosted Tree
- Boosted Dropout

  (As boosting is susceptible to overfitting-high bias, low variance)

Challenges (binary response)

- Studying the sensitivity of each variable to predict the outcome
- Using the EM-algorithm to impute missing data
- Using 5-fold cross-validation to assess model performance

Classifier Performance



Model performance

Method	CV error	Misclassification
RF	13.5%	14.0%
Boosted Tree	15.9%	14.9%
Boosted Tree with Dropout	16.9%	16.4%

Outcome

R-Shiny application that uses the supervised learning methods and

- Predicts the potentially fraudulent cases from different clinical sites
- Validates the parameter that gives the best fit
- Detects the covariates that are most predictive of the outcomes

CRADA

Cooperative Research and Development Agreement with CluePoints

The main objective is to detect atypical sites in a multicenter study

Method tests the distribution of data in one center with data in other centers and produces a p-value demonstrating how unlikely the outcomes from one clinical center are (unsupervised approach)

Approaching the end of 2nd year is a 3 year agreement

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Rashmi, K. V., & Gilad-Bachrach, R. (2015, May). Dart: Dropouts meet multiple additive regression trees. In *International Conference on Artificial Intelligence and Statistics*(pp. 489-497).

Acknowledgements

Mingwei Tang

Chetkar Jha

Nicholas Hein

Thank you!

U.S. FOOD & DRUG ADMINISTRATION