The Imbalanced Classification Add-In: Compare Sampling Techniques and Models

Michael Crotty, JMP Senior Statistical Writer
Marie Gaudard, Statistical Consultant
Colleen McKendry, JMP Technical Writer

Outline

- Add-In Purpose
- Background
  - What Is the Imbalanced Classification Problem
  - Obtaining a Classification Model
- Sampling Methods
  - Synthetic Minority Oversampling Technique (SMOTE)
  - Tomek Links
  - Smote Plus Tomek
- Imbalanced Classification Add-In Options
  - Evaluate Models Option
  - Three Sampling Options
- Obtaining the Add-In
- Example and Add-In Demo
- References
- Appendix (Additional Background)

Add-In Purpose

- The Imbalanced Classification add-in enables you to:
  - Apply a variety of sampling techniques designed for imbalanced data.
  - Compare the results of applying these techniques along with predictive models available in JMP Pro.
  - Compare models and sampling technique fits using Precision-Recall, ROC, and Gains curves, as well as other measures.
  - Choose a threshold for classification using these curves.
  - Apply the Tomek, SMOTE, and SMOTE plus Tomek sampling techniques directly to your data, enabling you to use existing JMP platforms and fine-tune the modeling options.
- Note: The Tomek, SMOTE, and SMOTE plus Tomek sampling techniques can be used with nominal and ordinal, as well as continuous, data.

Background

What is the Imbalanced Data Problem?

- The response variable is binary (could be multinomial)
  - # observations at one response level >> # observations at other response level
  - Call the response levels “majority” and “minority”
- The minority level is generally the level of interest
  - Examples include detection of fraud, disease, credit risk
- We want to predict class membership based on regression variables.
- We develop a predictive model that assigns probabilities of membership into the minority class.
- We choose a *threshold* value to optimize various criteria, such as the misclassification rate, the true positive rate, the false positive rate, precision, recall, etc.
- We classify an observation whose predicted probability of membership (or “score”) exceeds the threshold value into the minority class.

Background

Obtaining a Classification Model

- Some traditional measures of classification accuracy are not appropriate for imbalanced data.
  - For example, consider the case of a 2% minority class. You can achieve 98% accuracy simply by classifying all observations as majority cases!
- Precision-Recall (*PR*) curves are often used with imbalanced data. These plot the positive predictive value (*precision*) against the true positive rate (*recall*).
- Because precision takes majority instances into account, a PR curve is more sensitive to class imbalance than an ROC curve.
- As such, a PR curve is better able to highlight differences in models for imbalanced data.

Sampling Methods

- Sampling methods can be used to help modeling of the minority class.
- Sampling methods involve modifications to impose a more balanced distribution, or to better delineate the boundaries between majority and minority class observations.
- The Imbalanced Classification add-in implements seven sampling techniques:
  - No Weighting
  - Weighting
  - Random Undersampling
  - Random Oversampling
  - SMOTE
  - Tomek Links
  - SMOTE plus Tomek

Sampling Methods

- No Weighting
  - Can use for baseline comparison
- Weighting
  - Upweight each observation of the minority class by the same ratio
  - Define the ratio as # majority observations / # minority observations
- Random Undersampling
  - Randomly select a set of observations from the **majority** class
  - Remove this set from the data to decrease the total number of observations
- Random Oversampling
  - Randomly select (with replacement) a set of observations from the **minority** class
  - Add this set to the data to increase the total number of observations

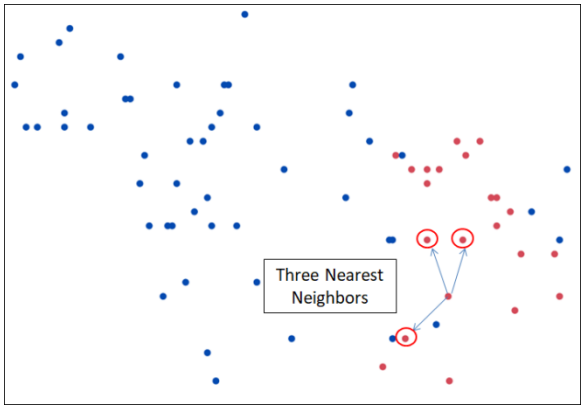
Note: For Undersampling and Oversampling, the resulting majority and minority class sets are equal in size.

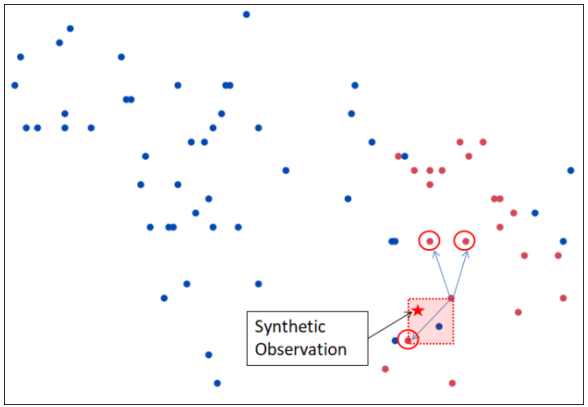
Sampling Methods

Synthetic Minority Oversampling Technique (SMOTE)

- A more sophisticated form of oversampling – adding more minority cases
- Generates new data observations that are similar to the existing minority class observations, rather than simply replicating them
- Using the Gower distance, perform K Nearest Neighbors on the minority class
- Observations are generated to fill in the space defined by the neighbors.
- We adapted the SMOTE algorithm to accommodate nominal and ordinal predictors.

Minority class in red





Sampling Methods

Tomek Links

- Attempts to better define the boundary between the minority and majority classes.
- Removes observations from the majority class that are "close" to minority class observations to better define cluster borders
- Using the Gower distance, finds Tomek links.
- A *Tomek link* is a pair of nearest neighbors that fall into different classes.
- To reduce overlapping of majority and minority instances, one or both members of the pair can be removed.
  - In the Evaluate Models option, we remove only the majority instance.
  - In the Tomek option, you can use either form of removal.

Sampling Methods

SMOTE plus Tomek

- Combines the two sampling methods
- First, apply the SMOTE algorithm to generate new minority observations.
- Then, with the newly generated observations included, apply the Tomek algorithm to find pairs of nearest neighbors that fall into different classes.
- In this sampling method, **both** observations in the pair are removed.

Imbalanced Classification Add-In Options

The Imbalanced Classification add-in consists of four options:

1. Evaluate Models

Fits a variety of models to data using a variety of sampling techniques.

2. Tomek Sampling
3. SMOTE Observations
4. SMOTE plus Tomek

Allows you to apply each of these sampling techniques to your data, so that you can select your own models and modeling options.

Imbalanced Classification Add-In Options
Evaluate Models

- Provides an Imbalanced Classification report that facilitates comparison of the model and sampling technique combinations.
- Shows Precision-Recall (PR) curves and ROC curves, and their AUC (*area under the curve*) values, as well as Gains curves.
  - PR curves use the Davis-Goadrich correction (Davis and Goadrich, 2006).
- Gives a plot of predicted probabilities by class.
- Also provides the Techniques and Thresholds table, which contains a script that allows you to reproduce the Imbalanced Classification report.
  - *Save this table in order to reproduce the report.*

Evaluate Models Option

The Evaluate Models option fits these models using these sampling techniques:

Models

- Naïve Bayes
- Neural Networks
  - NTanH(3) Model
  - Informative Missing
- Bootstrap Forest
  - Default options
  - Informative Missing
- Boosted Tree
  - Default options
  - Informative Missing
- Support Vector Machines
- Generalized Regression
  - Adaptive Lasso
  - Informative Missing
  - All two-way interactions (when number of predictors < 30)

Sampling Techniques

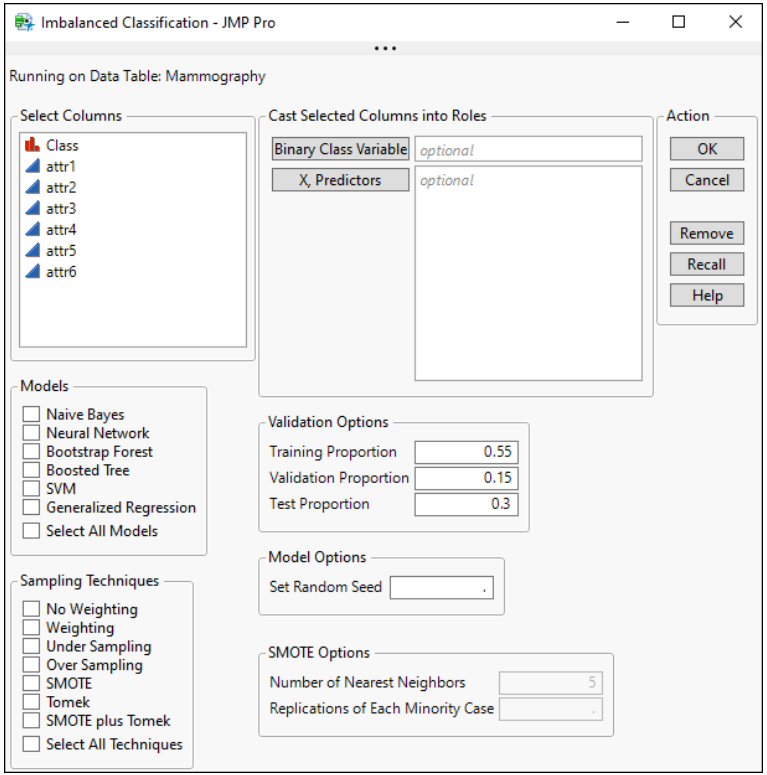
- No Weighting
- Weighting
- Random Undersampling
- Random Oversampling
- SMOTE
- Tomek Links
- SMOTE plus Tomek

Also available as stand-alone options

Imbalanced Classification Add-In Options

Evaluate Models Option

- Models, Sampling Techniques
  - Choose model and sampling combinations
- Validation Options
  - Used for all fitting options
- Model Options
  - Sets seed for sampling schemes as well as random validation within platforms
- SMOTE Options
  - Number of Nearest Neighbors – number of nearest neighbors selected for each minority observation
  - Replications of Each Minority Case – number of new observations generated for each minority observation, constrained between 1 and 10



Imbalanced Classification Add-In Options

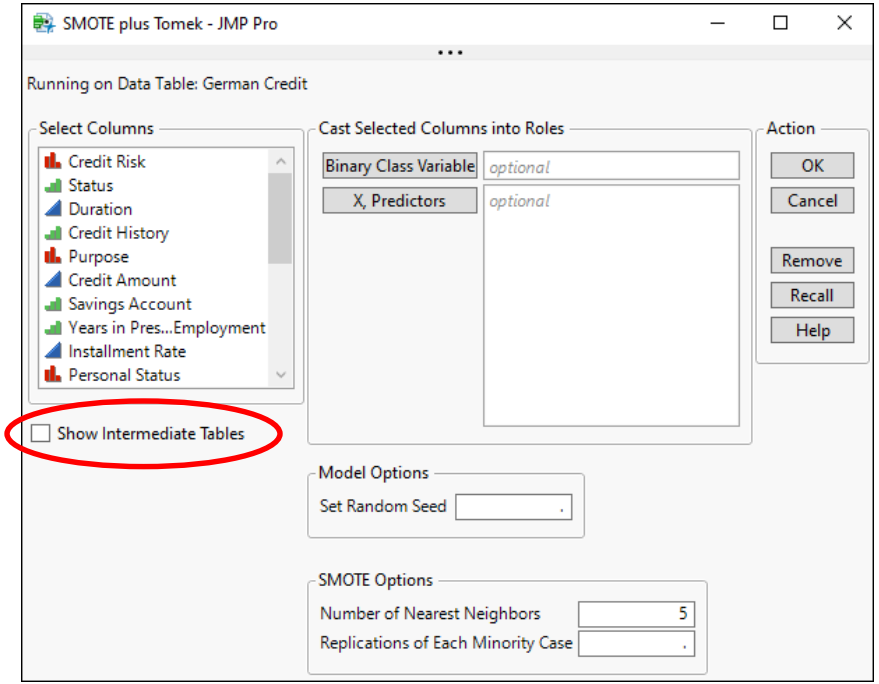
Three Sampling Options

- The Imbalanced Classification add-in also provides three other options that enable you to use sampling schemes:
  - Tomek Sampling
  - SMOTE Observations
  - SMOTE plus Tomek
- Tomek Sampling adds two columns to your data table that can be used as weights for predictive models.
  - One column removes only the majority nearest neighbor in a Tomek link, the other removes both members of a Tomek link.
- SMOTE Observations adds synthetic observations to your data table
- SMOTE plus Tomek adds synthetic observations and a weighting column that removes both members of a Tomek link.

Imbalanced Classification Add-In Options

Three Sampling Options

- The SMOTE plus Tomek dialog is shown to the right.
- Note the “Show Intermediate Tables” option, which is also available in the SMOTE Observations dialog.
- This option allows you to see the tables used in constructing the SMOTE observations.
- Also note that all sampling techniques, in particular SMOTE, Tomek, and SMOTE plus Tomek, are available for nominal and ordinal modeling types.

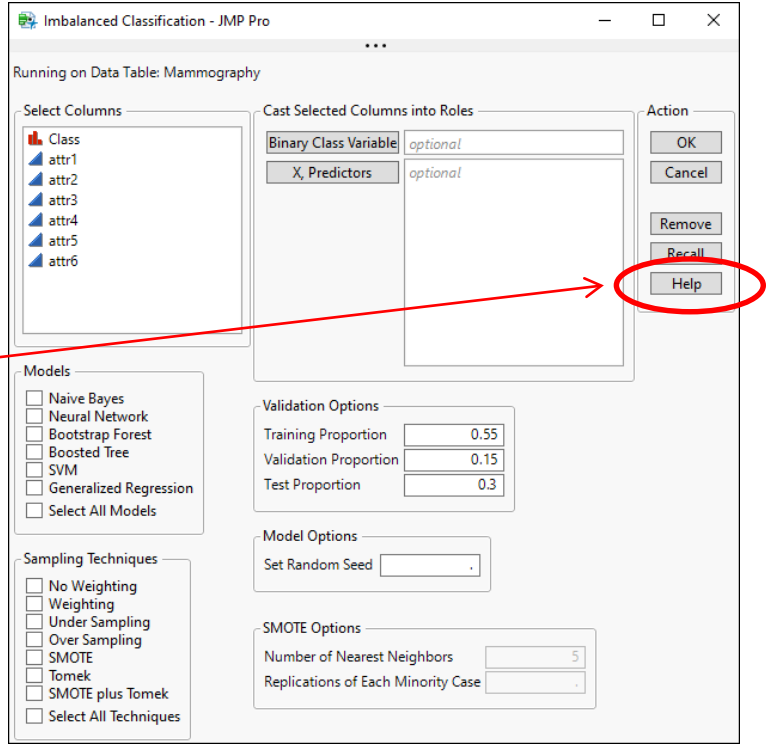


Obtaining the Add-In

- You can download the Imbalanced Classification add-in here:

  https://community.jmp.com/t5/Discovery-Summit-Americas-2020/The-Imbalanced-Classification-Add-In-Compare-Sampling-Techniques/ta-p/281551

- Documentation is available within the add-in.



Mammography Demo Data

- The Mammography data is based on a set of digitized film mammograms, used in a study of microcalcifications in mammographic images.
- Each record is classified as "1", representing calcification, or "0", representing no calcification.
- There are six continuous predictors and 11,183 observations.
- To reduce run time, the demo data set has 5,591 observations.
- 2.31% minority proportion

References

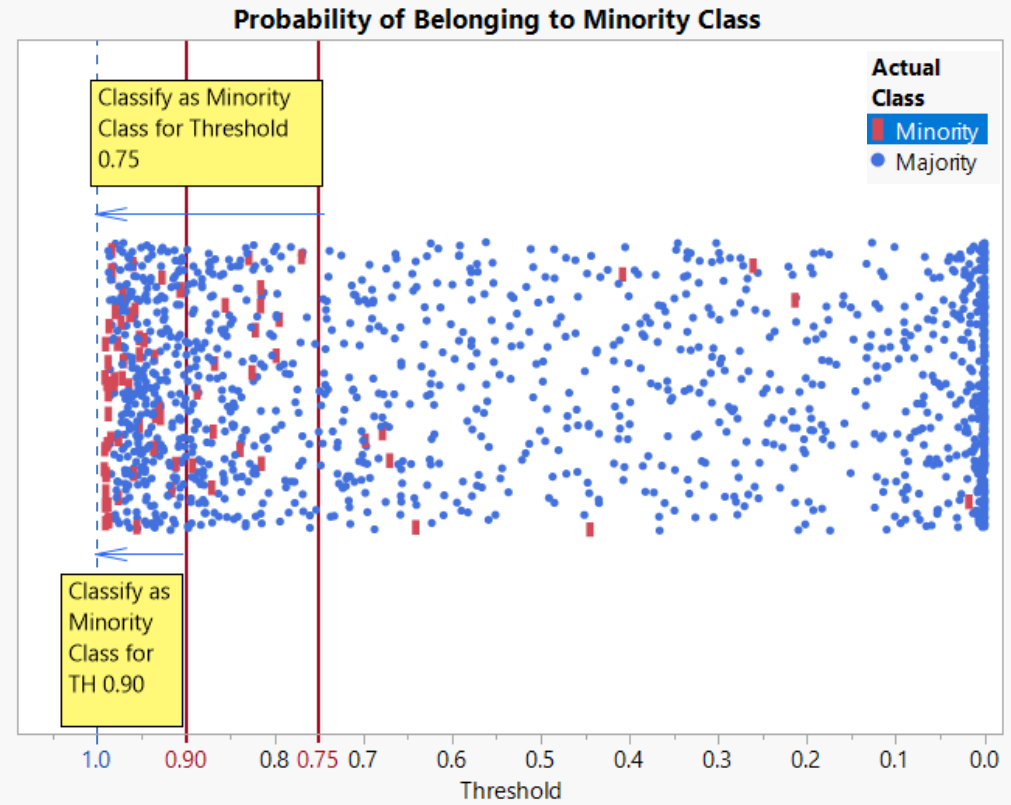
1. Chawla, N. V., et al. (2002). “SMOTE: Synthetic Minority Over-sampling Technique.” *Journal of Artificial Intelligence Research*, 16, pp. 321-357.
2. Davis, J., and Goadrich, M. (2006). “The Relationship between Precision-Recall and ROC Curves.” *Proceedings of the 23rd International Conference on Machine Learning*.
3. Flach, P. A., and Kull, M. (2015). “Precision-Recall-Gain curves: PR analysis done right.” *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems*, Vol. 1, pp 838-846.
4. He, H., and Garcia, E. A. (2009). “Learning from Imbalanced Data.” *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp. 1293-1284.
5. Kubat, M, and Matwin, S. (1997). “Addressing the Curse of Imbalance Training Sets: One-Sided Selection.” *Proceedings of the Fourteenth International Conference on Machine Learning*.
6. Longadge, R., Dongre, S. S., and Malik, L. (Feb. 2013). “Class Imbalance Problem in Data Mining: Review.” *International Journal of Computer Science and Network*, Vol. 2:1.
7. Saito T, and Rehmsmeier, M. (2015). “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE* 10(3).

Appendix

Additional Background Information

Background

Threshold for Prediction



- A data set consists of 1,452 observations, with only 78 in the minority class.
- The plot shows predictive probabilities of membership in the minority class (thresholds) based on a given model.
- Two thresholds are shown: 0.90 and 0.75.
- Each defines a classification rule.
- As the threshold decreases, more minority instances are identified. But the false positive rate also increases.

Background

Misclassification Measures

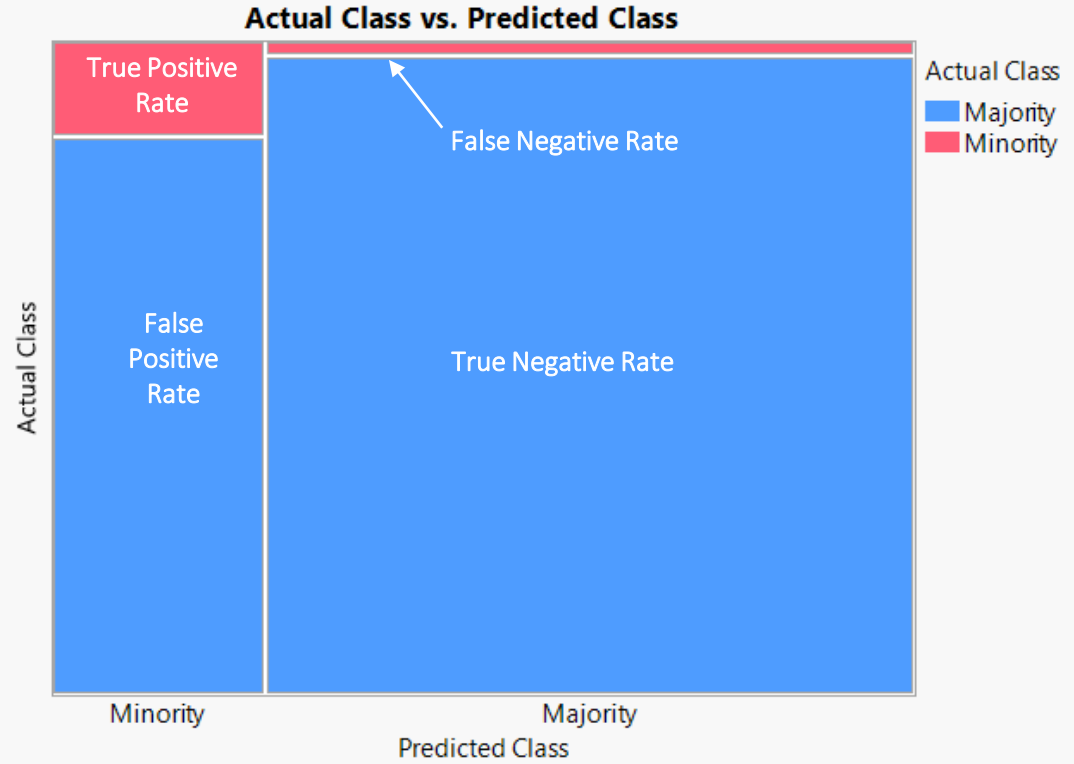
- For a binary response, one measure of accuracy is the *confusion matrix*.
- **It is based on selection of a given threshold**.
- The threshold in JMP is 0.5 by default, or you can set a threshold using the Profit Matrix column property.

Confusion Matrix	Predicted Yes	Predicted No
Actual Yes	True Positive	False Negative
Actual No	False Positive	True Negative

- A related summary measure: **Accuracy** = (TP + TN) / (TP + FP + TN + FN)
- JMP reports: **Misclassification Rate** = 1 - Accuracy

Background

Misclassification Measures



- Here is a confusion diagram and matrix for threshold 0.90.

		Predicted Class		
	Count	Minority	Majority	Total
	Row %			
Actual Class	Minority	53	25	78
		67.9%	32.1%	
	Majority	309	1065	1374
		22.5%	77.5%	
	Total	362	1090	1452

Background

Misclassification Measures

- Misclassification rate breaks down with severe imbalance
- Consider the case of a 2% minority class:
  - You can achieve 98% accuracy simply by predicting all majority cases!
  - This would be a bad classifier, however.

- Each threshold value defines a classification scheme and confusion matrix
- Consider curves that plot classification behavior across all thresholds:
  - Precision-Recall Curves
  - Receiver Operating Characteristic (ROC) Curves
  - Gains Curves

Background

Misclassification Measures

- For a given threshold:

Predicted Class				
Actual Class	Count	Minority	Majority	Row Total
	Minority	TP	FN	TP + FN = P
	Majority	FP	TN	FP + TN = N
	Col Total	TP + FP	TN + FN	

- Sensitivity = True Positive Rate = TP / P
- Specificity = True Negative Rate = TN / N
- 1 – Specificity = False Positive Rate = FP / N
- Precision = Positive Predictive Value = TP / (TP + FP)
- Recall = Sensitivity = TP / P

Background

Comparison of Curves

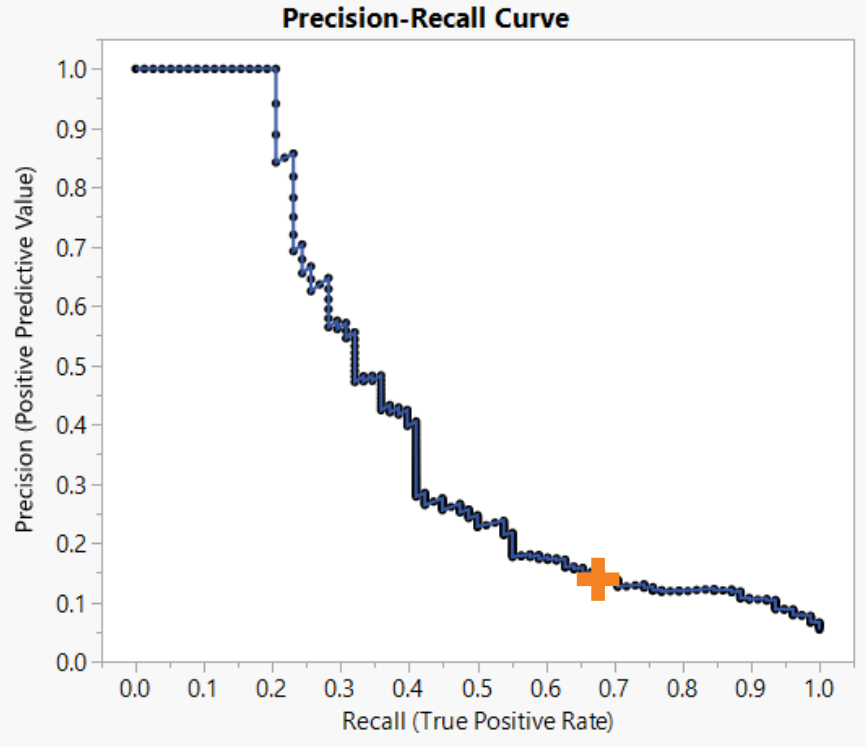
- The PR, ROC, and Cumulative Gains curves are related:

Plot	Y Axis		X Axis	
PR Curve	Precision	True Positives/ (True + False Positives)	Recall	*True Positive Rate*
ROC Curve	Sensitivity	*True Positive Rate*	1 - Specificity	False Positive Rate
Cumulative Gains Curve	Cumulative Gains	*True Positive Rate*	Portion	Proportion of Top-Ranked Observations

- The ideal curve has the Y axis quantity equal to 100%.

Background

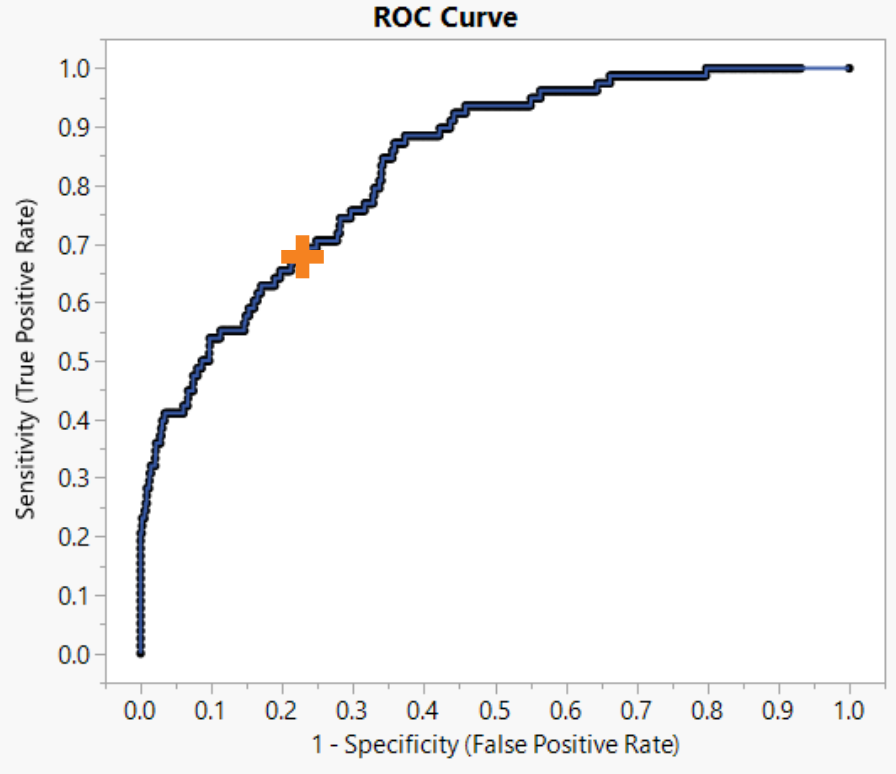
Precision-Recall Curve



- Precision-Recall (PR) Curve
  - Plots precision versus recall
  - Precision = TP / (TP + FP)
  - Recall = TP / P
- Precision is the Positive Predictive Value
- Recall is the True Positive Rate (Sensitivity)
- The PR curve is preferred for imbalanced data.

Background

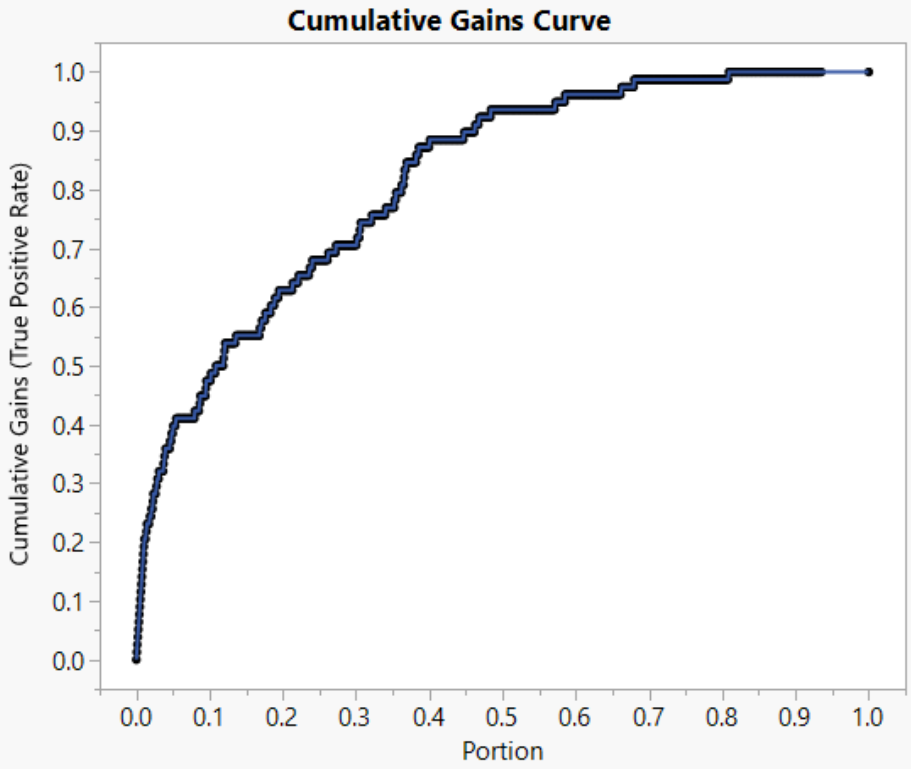
ROC Curve



- ROC Curve
  - Plots sensitivity vs. 1 - specificity
  - Sensitivity = TP / P
  - 1 - Specificity = FP / N
- Sensitivity is the True Positive Rate (Recall)
- 1 - Specificity is the False Positive Rate

Background

Cumulative Gains Curve



- Cumulative Gains Curve
  - Plots cumulative gains vs. portion of the data
  - Cumulative Gains = TP / P (Sensitivity)
  - Portion = proportion of the observations ranked by their probability of membership in the minority class

Background

Solutions for Imbalanced Data Problems

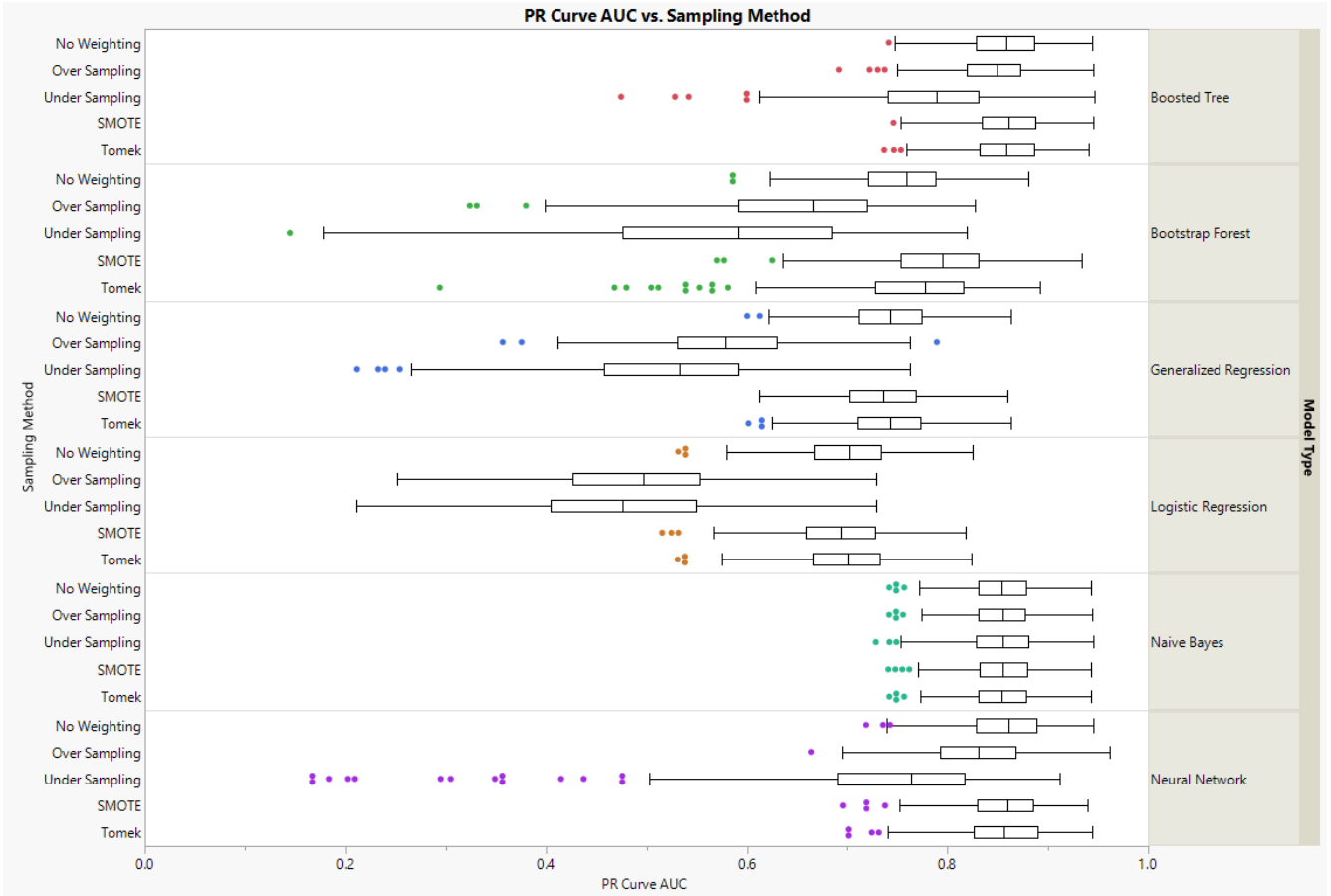
- Sampling methods
  - Make modifications to impose a more balanced distribution
- Cost-sensitive methods
  - Use cost to differentiate misclassification consequences or to combine models in an ensemble
  - Incorporate cost information into the classification scheme
- Kernel-based methods
  - Support vector machines (SVMs); can also be integrated with sampling methods

Data Driven Simulations

Structure

- Simulations studies were conducted using two data sets
  - Mammography and Wilt
- Use the sample size of the data set
  - N = 11,183 in Mammography
  - N = 4,839 in Wilt
- Use the covariance structure of the data set
- Vary the mean vector of the minority class
  - The original mean vector from the data
  - Mean vector that is half the original distance from the majority mean vector
  - Mean vector that is twice the original distance from the majority mean vector
- Vary the proportion of minority class observations
  - Proportion vector (.002, .005, .01, .02, .04, .06, .1, .15, .25, .5)
- Evaluation based on AUC from ROC and PR curves
- 250 iterations for each combination

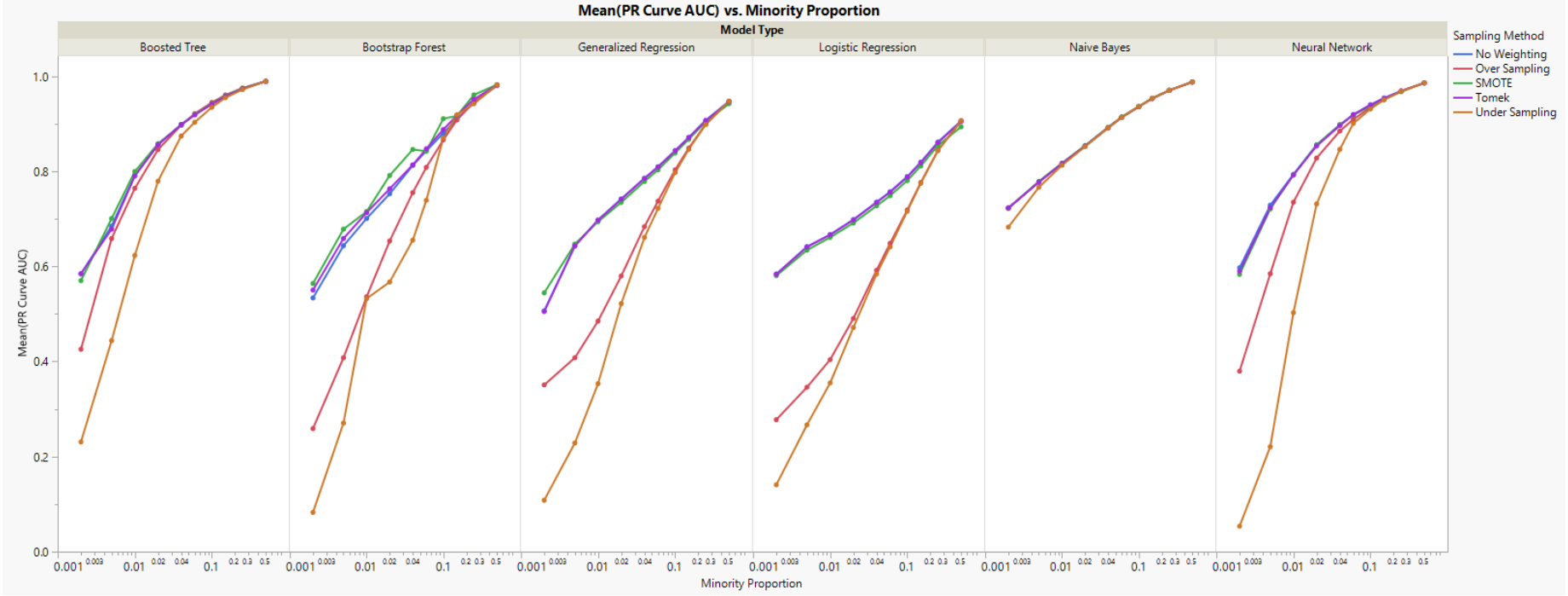
Simulations Based on Mammography Data



2% minority proportion and original mean vector

Simulations Based on Mammography Data

Original mean vector

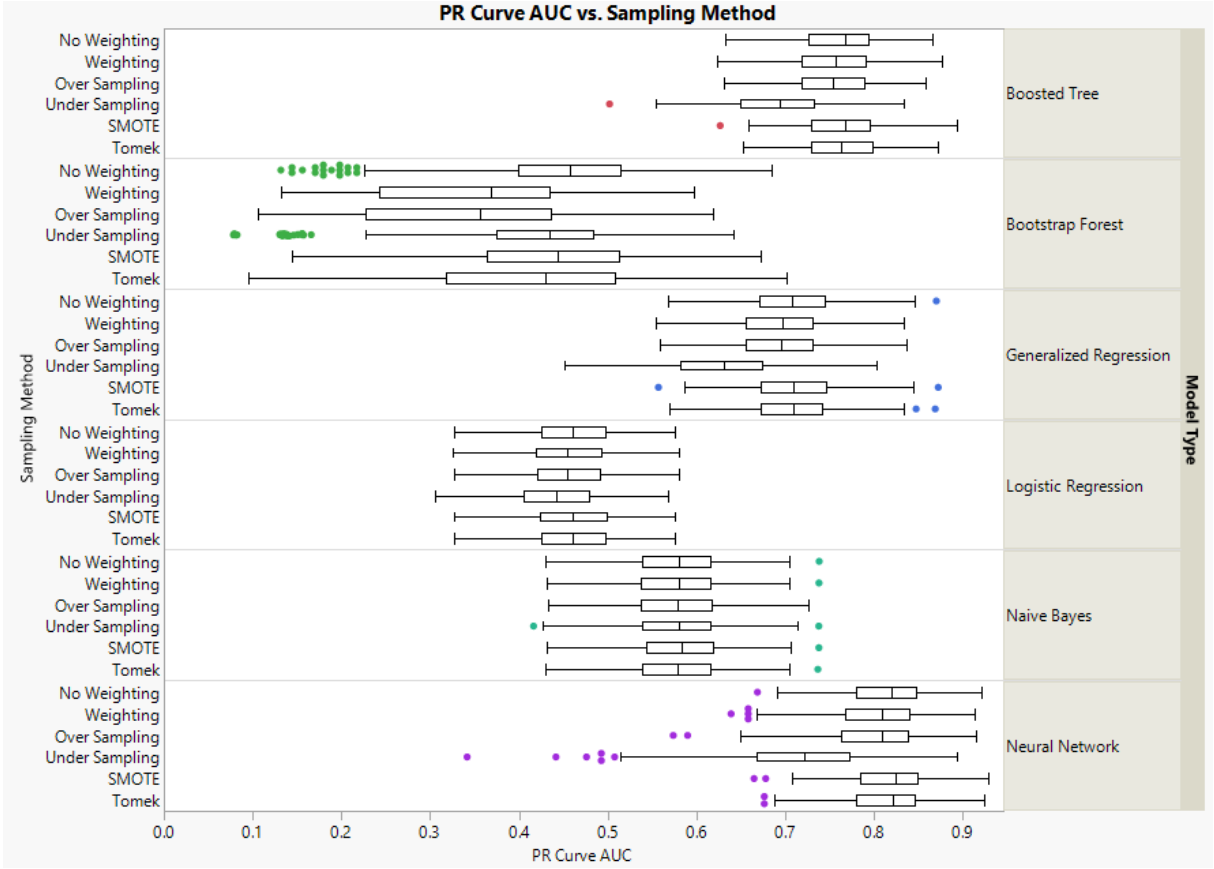


Simulations Based on Mammography Data

Conclusions

- The Boosted Tree, Neural Network, and Naïve Bayes models perform well.
- Undersampling performs poorly for almost all models up to about 10% minority proportion.
- Sometimes no weighting performs better than some of the simpler sampling techniques (weighting, oversampling, and undersampling).
- SMOTE and Tomek consistently perform as well as or better than no weighting.
- There is variation in sampling technique performance for all models except Naïve Bayes.

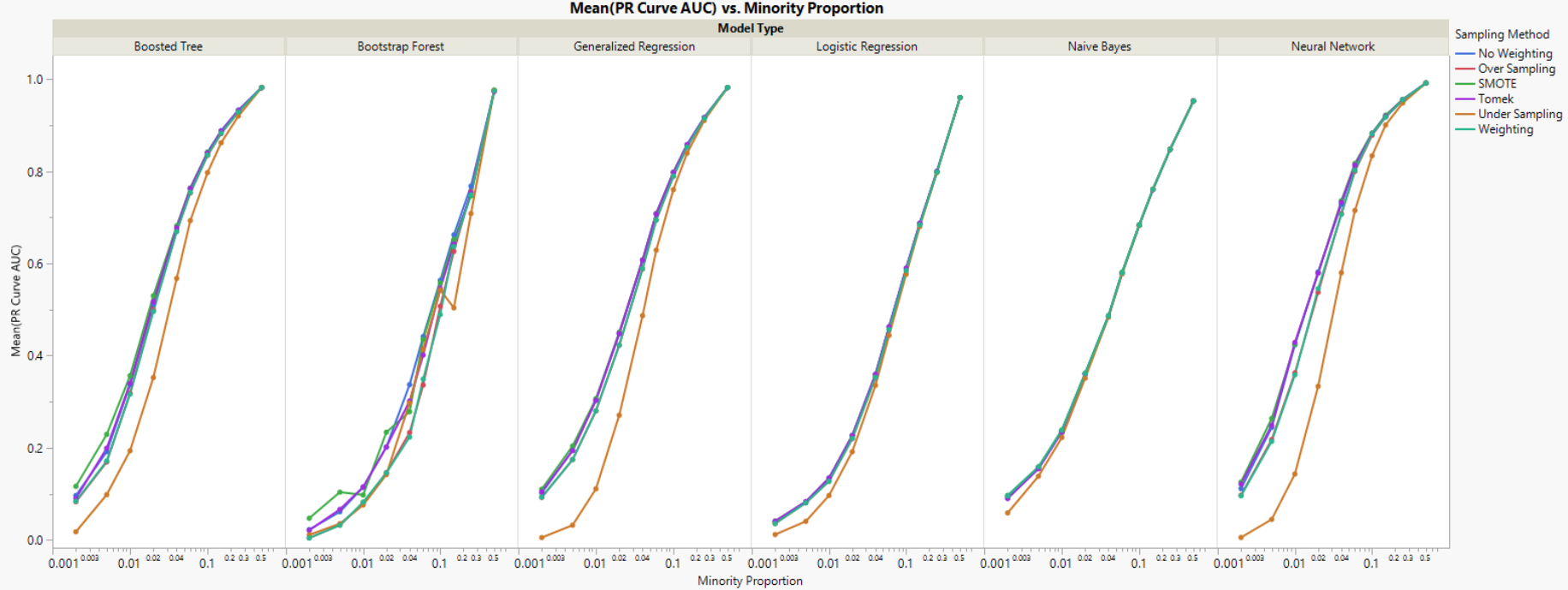
Simulations Based on Wilt Data



6% minority proportion and original mean vector

Simulations Based on Wilt Data

Original mean vector



Simulations Based on Wilt Data

Conclusions

- Insights obtained from exploring the data indicate that the minority/majority class overlap in the Wilt data is greater than in the Mammography data.
- The Boosted Tree and Neural Network models perform best.
- There is not much variation in the sampling techniques, except when the distance between means is doubled.

Simulation Study Conclusions

- Undersampling performs poorly compared to other sampling techniques.
  - In simulations based on the Mammography data, it performs poorly for almost all models up to about 10% minority proportion.
  - In simulations based on the Wilt data, it performs poorly for almost all models when the distance between the means is doubled.
- The Boosted Tree and Neural Network models perform the best.
  - Naïve Bayes performs better in simulations based on the Mammography data.
  - Generalized regression performs better in simulations based on the Wilt data.
- There appears to be an interaction between model type and distance between means in their impact on performance.
  - When classes are well separated, logistic and generalized regression perform well, but perform very poorly for classes that overlap.
- Bootstrap Forest has the most variability.

Conclusions

- PR curves highlight differences in sampling methodologies whereas ROC curves tend to mask these differences.
- For highly imbalanced data, PR curves give insight on how to choose a “better” modeling technique – one that gives greater precision for a given true positive rate, thus resulting in fewer false positives.
- The separation between means and the minority proportion have an impact on which models and sampling techniques perform well.
  - We suggest using the Imbalanced Data script whenever the minority proportion is less than 10%.
- The Imbalanced Data script is useful in evaluating and selecting models, whether or not the binary class is imbalanced.

Thanks!

Michael.Crotty@jmp.com
Colleen.McKendry@jmp.com

sas.com