Anomaly Detection and JMP® Pro

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

THE POWER TO KNOW.

Outline

- Project goal
- Background
- Imbalanced Data script: dialog, models and sampling techniques, evaluation report
- Real data examples
- Simulations
- Conclusions
- Future work
- References

Project Goals

- To highlight aspects of the imbalanced data problem in the context of classification into a minority and a majority class, where the minority class is under-represented relative to the majority class.
- To provide users with a tool that allows them to explore predictive models that are available in JMP Pro, in conjunction with sampling techniques that are useful in modeling imbalanced data.
- To show examples of the value of the Precision-Recall curve in imbalanced situations.
- To share conclusions about the relative performance of the prediction models and sampling techniques that we studied.
- To provide suggestions about when class imbalance may become an issue for typical modeling techniques.

Background

What is the Imbalanced Data Problem?

- Binary response variable
  - # observations at one response level >> # observations at other response level
  - Call the response levels “majority” and “minority”
- Minority level is generally the level of interest
  - Examples include: detection of fraud, disease, credit risk
- Want to predict class membership based on regression variables.
- Some traditional measures of classification accuracy are not appropriate for imbalanced data.

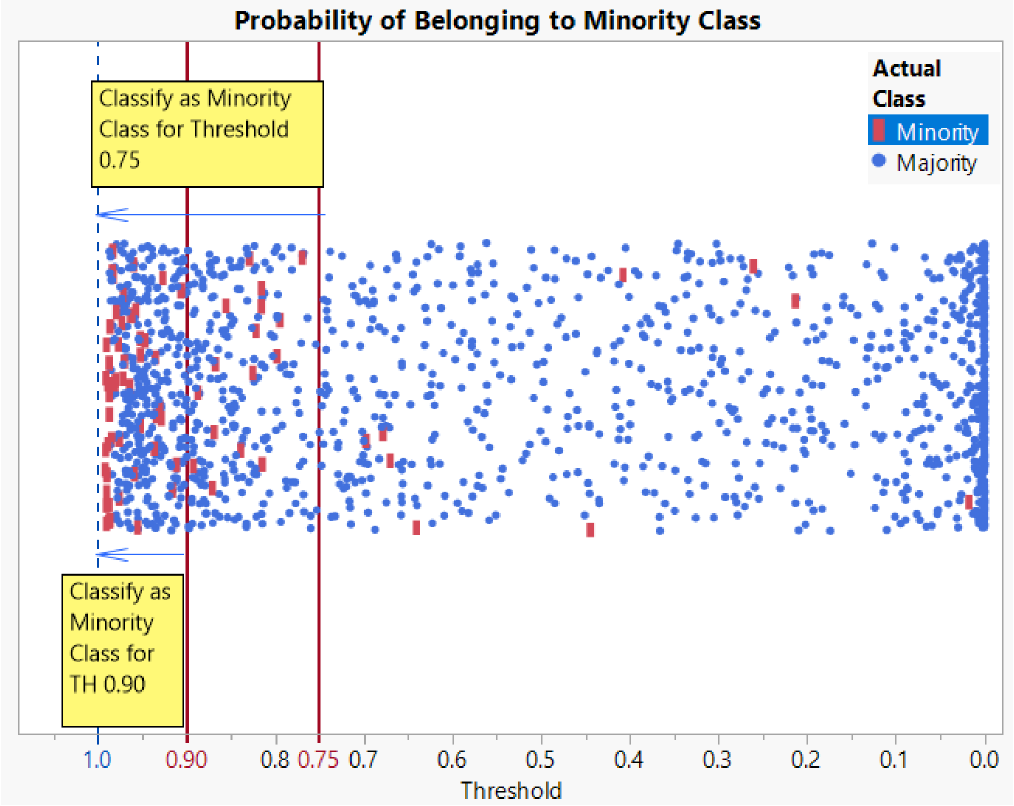
Background

Obtaining a Classification Model

- A predictive model that assigns probabilities of membership into the minority class is developed.
- Classification using the predictive model requires selection of a *threshold* value.
- An observation whose predicted probability of membership (or “score”) exceeds the threshold value is classified into the minority class.
- Thus, the threshold value defines the classification scheme.
- One tries to choose a threshold value to optimize various criteria, such as the misclassification rate, the true positive rate, the false positive rate, precision, recall, etc.

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

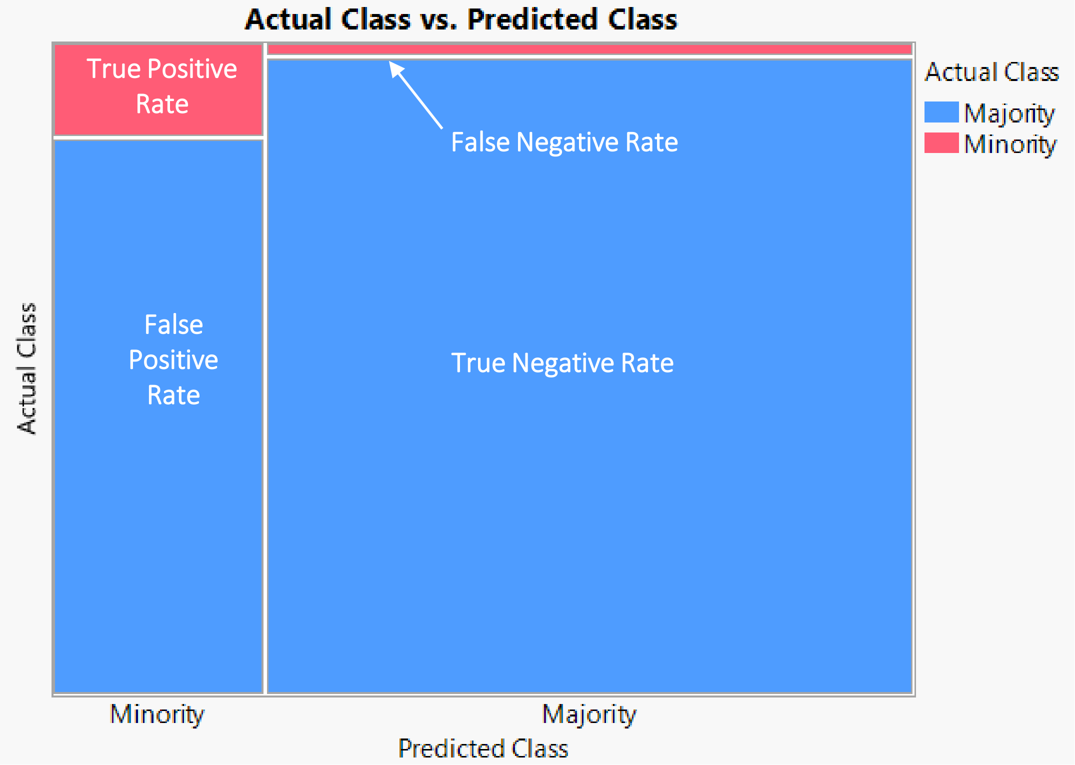
- For a binary response, one measure of accuracy is the *confusion matrix*.
- **It is based on selection of a given threshold**.
- The threshold in JMP is 0.5 by default, or you can set a threshold using the Profit Matrix column property.

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

Actual Class	Count / Row %	Minority	Majority	Total
	Count			
	Row %			
Minority		53 67.9%	25 32.1%	78
Majority		309 22.5%	1065 77.5%	1374
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

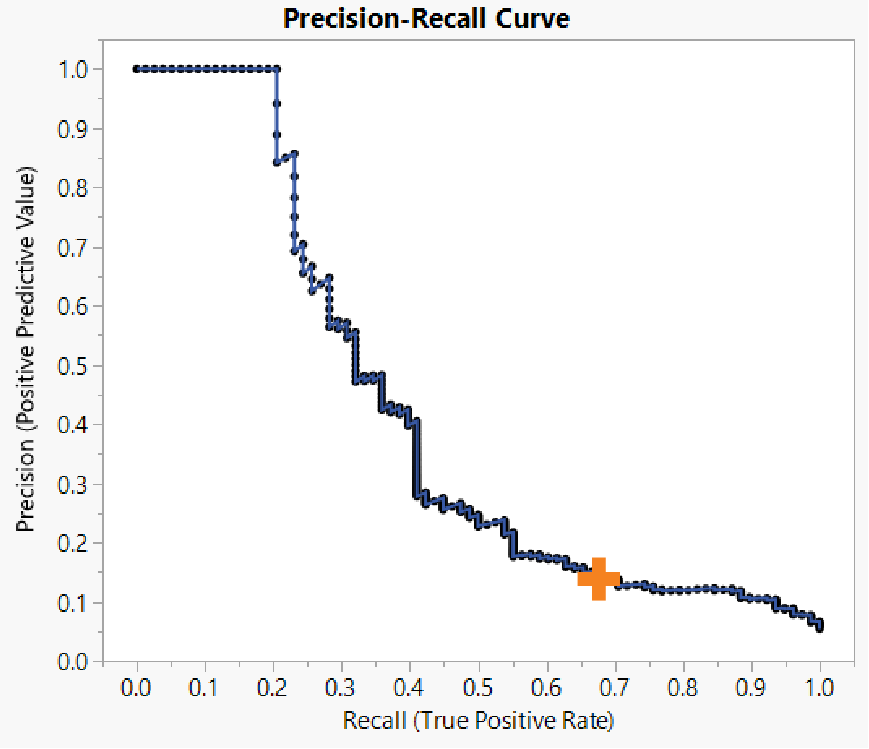
- The PR, ROC, and Cumulative Gains curves are related:

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%.

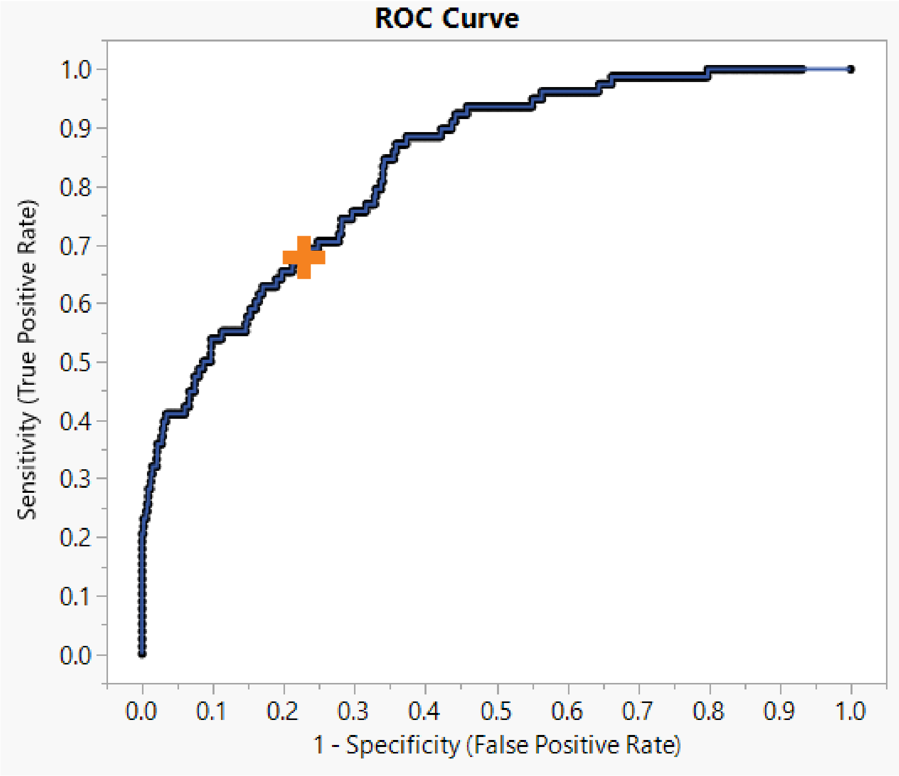
Background

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.

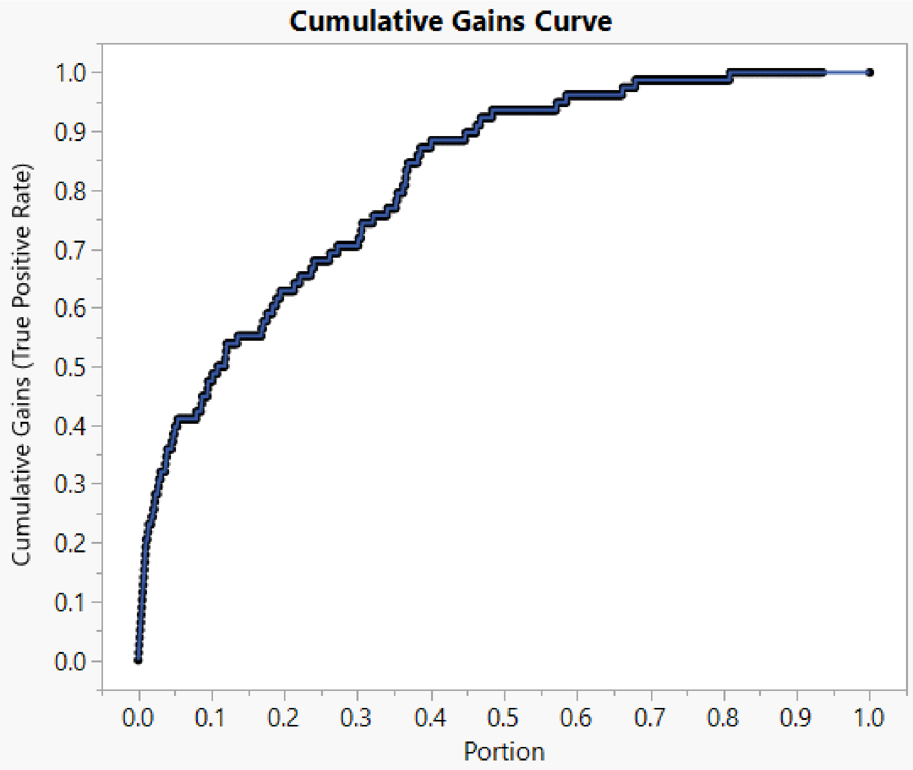
Background
ROC Curve



- 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

Background

Sampling Methods Approaches

- Sampling methods involve modifications to impose a more balanced distribution
  - Random oversampling and undersampling
  - Informed undersampling (*EasyEnsemble*, *BalanceCascade*)
  - Synthetic sampling with data generation (SMOTE)
  - Adaptive synthetic sampling (ADA-SYN)
  - Sampling with data cleaning (Tomek links)
  - Cluster-based sampling method
  - Integration of sampling and boosting

Imbalanced Data in JMP

- We want to address imbalanced data sets using JMP Pro.
- How can we implement sampling techniques and combine them with JMP Pro platforms to perform data analysis?
- Chose appropriate JMP Pro platforms.
- Chose a variety of sampling techniques.
- We created a script that enables users to fit and compare models for imbalanced data with a binary response.

Imbalanced Data in JMP

JMP Pro Platforms

- Naïve Bayes
- Neural Networks
  - NTanH(3) Model
- Bootstrap Forest
  - Default options
- Boosted Tree
  - Default options
- Logistic Regression
- Generalized Regression
  - Adaptive Lasso
  - All two-way interactions

Sampling Techniques

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

* *These techniques are implemented using R.*

Imbalanced Data in JMP

Sampling Methods

- No Weighting
  - Original data
  - Baseline comparison
- Weighting
  - Upweight each observation of the minority class by the same ratio
  - Define the ratio as # majority observations / # minority observations

Imbalanced Data in JMP

Sampling Methods

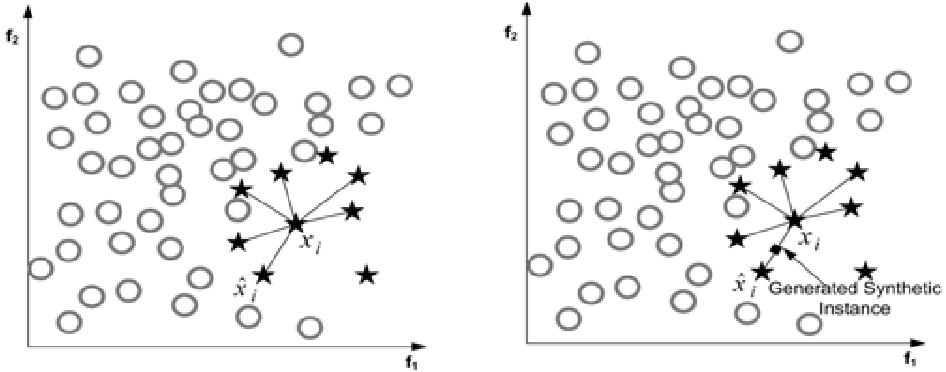
- Random Undersampling
  - Randomly select a set of observations from the **majority** class
  - Remove this set from the data to decrease the total number of observations
- Random Oversampling
  - Randomly select (with replacement) a set of observations from the **minority** class
  - Add this set to the data to increase the total number of observations

For both methods, the sets are selected such that the sizes of the minority and majority classes are equal.

Imbalanced Data in JMP

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
  - Perform K Nearest Neighbors on the minority class
  - $x_{new} = x_i + (\hat{x}_i - x_i) + \delta$
    - x_i minority class observation
    - $\hat{x}_i$ one of the nearest neighbors for x_i
    - δ random number in [0,1]



Figures from He and Garcia (2009; section 3.1)

Imbalanced Data in JMP

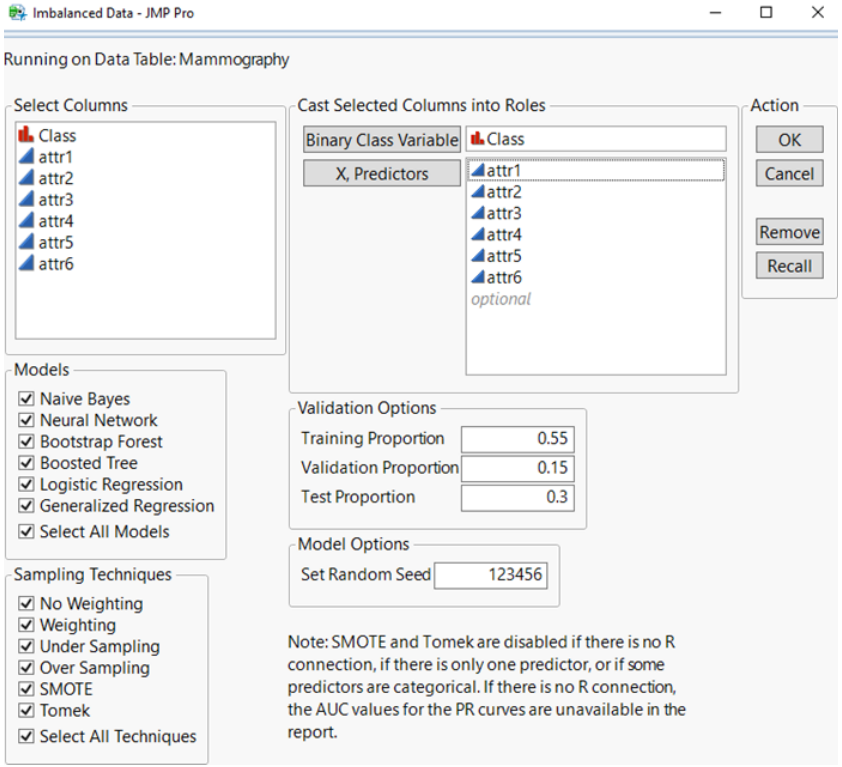
Sampling Methods

- Tomek Links
  - A more sophisticated form of undersampling – removing majority cases
  - Removes observations from the majority class that are "close" to minority class observations to better define cluster borders
  - Find pairs of nearest neighbors, (x_i, x_j), that fall into different classes to reduce overlapping of majority and minority instances.
    - x_i in minority class
    - x_j in majority class
    - Remove x_j from data

Imbalanced Data in JMP

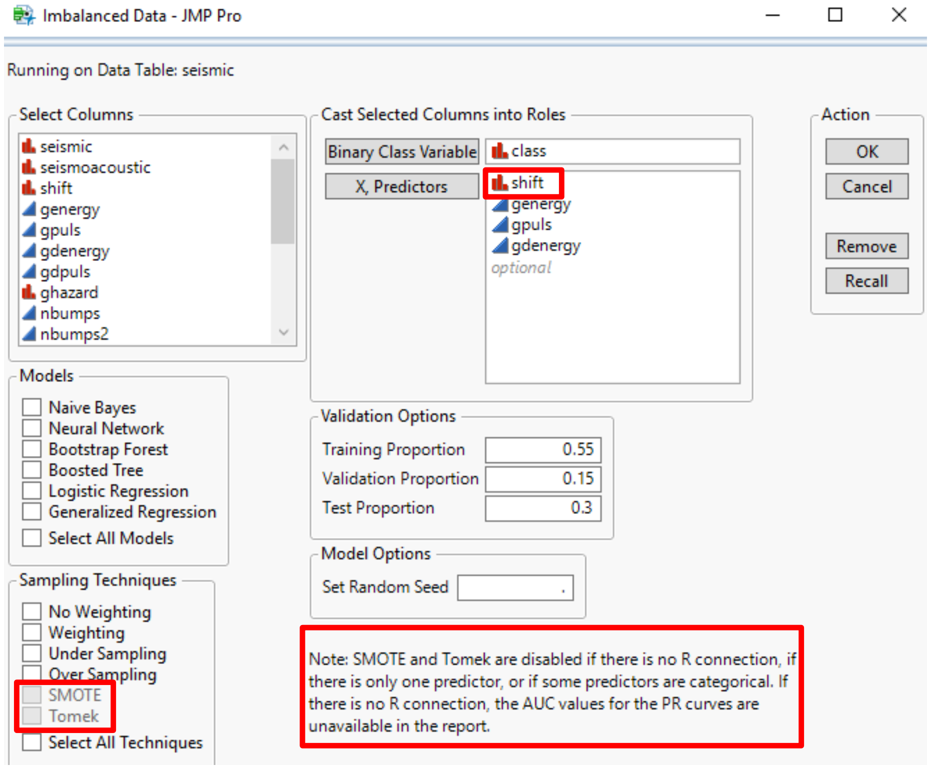
Dialog Window

- Choose model and sampling technique combinations
  - For use with SMOTE and Tomek, data are standardized
- Validation options
  - A validation set is used for all fitting options
- Random Seed
  - Sets seed for sampling schemes as well as random validation within platforms
  - Results not identical between JMP 14 and JMP 15 due to changes in random seeds
  - JMP 15 used in this presentation



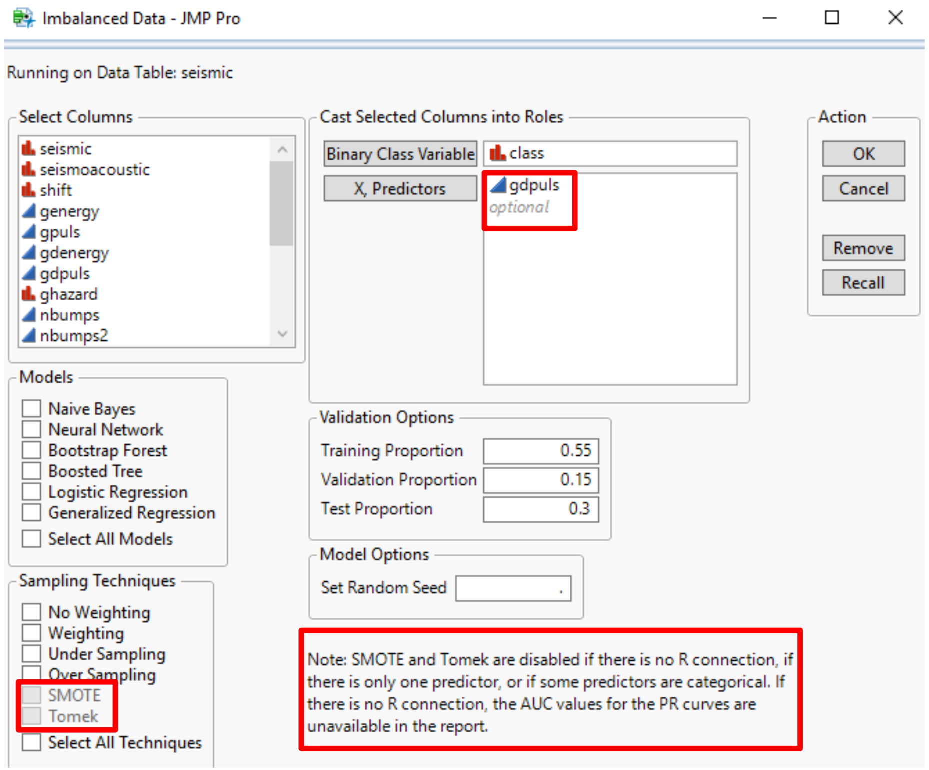
Imbalanced Data in JMP

Dialog Window



Imbalanced Data in JMP

Dialog Window



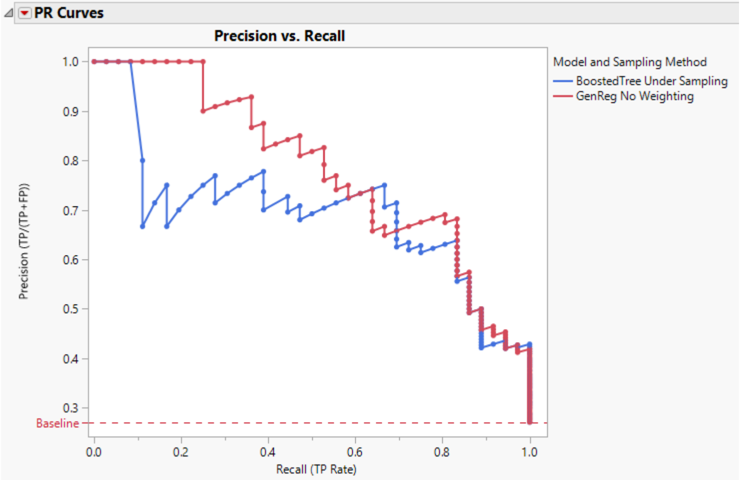
Data Sets Studied

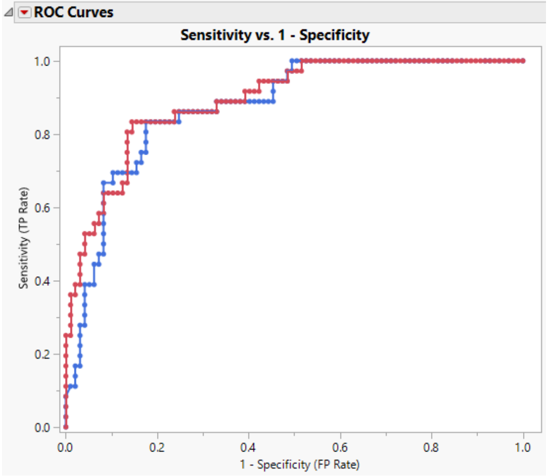
- Considered nine data sets.
- Minority class representation runs from 35.90% to 0.17%

	Data Set	N Predictors	N Continuous	N Nominal	N	N Minority Class	Minority %
1	Ionosphere	34	32	2	351	126	35.90%
2	Pima Indians	8	8	0	768	268	34.90%
3	Diabetes Modified	10	9	1	442	121	27.38%
4	Ecoli	7	7	0	336	77	22.92%
5	New Thyroid	5	5	0	215	35	16.28%
6	Seismic	18	14	4	2584	170	6.58%
7	Wilt Data	5	5	0	4839	261	5.39%
8	Mammography	6	6	0	11183	260	2.32%
9	Credit Card Fraud	30	30	0	284807	492	0.17%

Data Sets Studied

- The three data sets with the highest minority class percentage showed less difference in terms of classification models and sampling methods than did the other data sets.
- However, even for Diabetes Modified.jmp, with a 27.4% minority proportion, the PR curves differentiate between models, while the ROC curves are similar.





Data Sets Studied

Mammography

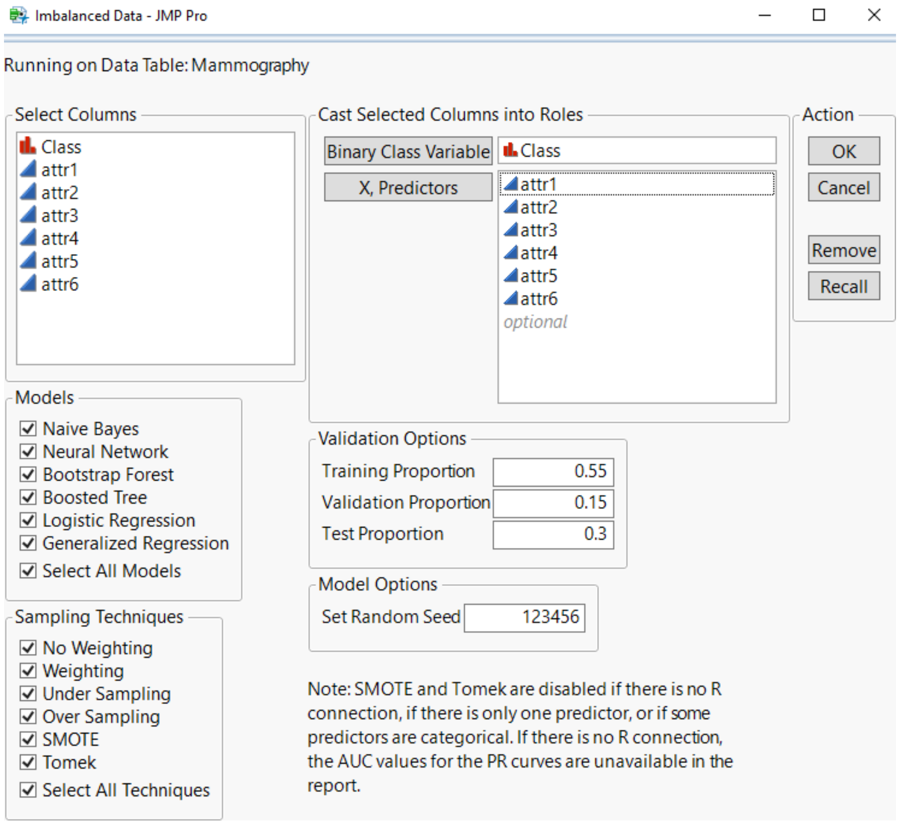
- The data table Mammography.jmp is based on a set of digitized film mammograms, used in a study of microcalcifications in mammographic images.
- There are six continuous predictors and 11,183 observations.
- Each record is classified as "1", representing calcification, or "0", representing no calcification.
- How might one use the Imbalanced Data script, and the Evaluation Report, to select a model?
- Details are given in the following slides, which are for your reference.

Go to Assessment of Differences

Data Sets Studied

Mammography

- Run the Imbalanced Data script with your data table of interest as the active data table.
- The script opens the dialog to the right.
- Make appropriate selections and run the script.



Data Sets Studied

- When you run the Imbalanced Data script, the following are provided:
  - The Evaluation Report, called “Imbalanced Data for <current data table>”
  - The Techniques and Thresholds data table, which contains scripts for the Evaluation Report and the Summary Table.
  - The Summary Table
  - The Training Set – these are the observations used to fit the models, and they include the validation set selected using the specifications in the dialog
  - The Test Set – this is the independent set of observations used to produce the Techniques and Thresholds data table and the Evaluation Report.
- The Techniques and Thresholds table contains the detailed data used to produce the Evaluation Report.
- The Summary Table links to the Techniques and Thresholds table, and thus to the Evaluation Report.

Data Sets Studied

Mammography

Techniques and Threshol...
- Source
- Summary Table
- Evaluation Report

Columns (13/0)
- Model Type
- Sampling Method
- Model and S...pling Method
- Probs
- Class
- Precision
- Recall
- Sensitivity
- 1 - Specificity
- Cumulative Gains
- Portion
- ROC Curve AUC
- PR Curve AUC

Rows	
All rows	120,816
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Model Type	Sampling Method	Model and Sampling Method	Probs	Class	Precision	Recall
1	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	•	0.68571	0.000000
2	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
3	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
4	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
5	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
6	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
7	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
8	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
9	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	0	0.68571	0.307692
10	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
11	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
12	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
13	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
14	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	0	0.68571	0.307692
15	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
16	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
17	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
18	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	0	0.68571	0.307692
19	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
20	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	1	0.68571	0.307692
21	Naive Bayes	No Weighting	Naive No Weighting	1.0000000	0	0.68571	0.307692

- The Techniques and Thresholds data table shows each selected modeling and sampling technique, its probability thresholds, and the computed values that are plotted on the curves.
- Note the Summary Table and Evaluation Report table scripts.

Data Sets Studied

Mammography

- The Summary table gives AUC values for each selected method.
- It also provides an easy way to select curves for methods in the Evaluation report, or rows in the Techniques and Thresholds table.

Summary Table for Mammography
Source
Bar Graph Comparison

Columns (6/0)
Model Type
Sampling Method
Model and Sampling Method
N Non-Missing Probabilities
PR Curve AUC
ROC Curve AUC

Rows

All rows	36
Selected	0
Excluded	0
Hidden	0
Labelled	0

	Model Type	Sampling Method	Model and Sampling Method	N Non-Missing Probabilities	PR Curve AUC	ROC Curve AUC
1	Boosted Tree	No Weighting	BoostedTree No Weighting	3356	0.6002	0.9341
2	Boosted Tree	Over Sampling	BoostedTree Over Sampling	3356	0.5163	0.9491
3	Boosted Tree	SMOTE	BoostedTree SMOTE	3356	0.5699	0.9437
4	Boosted Tree	Tomek	BoostedTree Tomek	3356	0.5889	0.9344
5	Boosted Tree	Under Sampling	BoostedTree Under Sampling	3356	0.5086	0.9462
6	Boosted Tree	Weighting	BoostedTree Weighting	3356	0.5221	0.9499
7	Bootstrap Forest	No Weighting	BootstrapForest No Weighting	3356	0.5739	0.9461
8	Bootstrap Forest	Over Sampling	BootstrapForest Over Sampling	3356	0.5402	0.9438
9	Bootstrap Forest	SMOTE	BootstrapForest SMOTE	3356	0.4791	0.9432
10	Bootstrap Forest	Tomek	BootstrapForest Tomek	3356	0.5360	0.9421
11	Bootstrap Forest	Under Sampling	BootstrapForest Under Sampl...	3356	0.1348	0.8373
12	Bootstrap Forest	Weighting	BootstrapForest Weighting	3356	0.4908	0.9435
13	Generalized Regr...	No Weighting	GenReg No Weighting	3356	0.6022	0.9351
14	Generalized Regr...	Over Sampling	GenReg Over Sampling	3356	0.6304	0.9460
15	Generalized Regr...	SMOTE	GenReg SMOTE	3356	0.6397	0.9455
16	Generalized Regr...	Tomek	GenReg Tomek	3356	0.6027	0.9350
17	Generalized Regr...	Under Sampling	GenReg Under Sampling	3356	0.4339	0.9507

Data Sets Studied

Mammography

Summary

Report Details

Model Specifications

Binary Class Variable	Predictors
Class	attr1
	attr2
	attr3
	attr4
	attr5
	attr6

Models	Sampling Techniques	Validation Options	Reproducibility
Naive Bayes	No Weighting	Training Proportion: 0.55	123456
Neural Network	Weighting	Validation Proportion: 0.15	
Bootstrap Forest	Under Sampling	Test Proportion: 0.3	
Boosted Tree	Over Sampling		
Logistic Regression	SMOTE		
Generalized Regression	Tomek		

Binary Class Distributions in Training and Test Sets

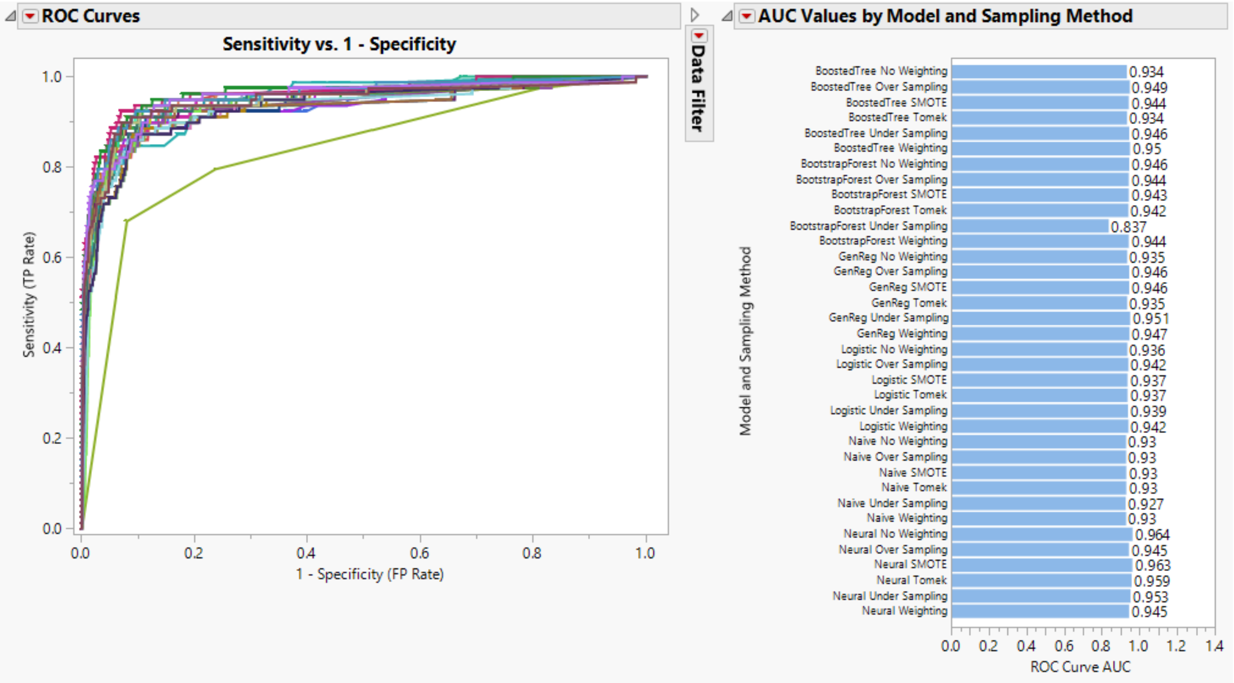
Training Binary Class Variable	Number	Percent
0	7646	97.7%
1	182	2.3%
Total	7828	100.0%

Test Binary Class Variable	Number	Percent
0	3277	97.7%
1	78	2.3%
Total	3355	100.0%

- Run the Evaluation Report script in the Techniques and Thresholds data table to obtain the Evaluation report.
- The Summary outline provides details about the report and information about the analysis that generated the report.
- This outline is followed by the Precision-Reliability Curves, ROC Curves, and Cumulative Gains Curves outlines.

Data Sets Studied

Mammography

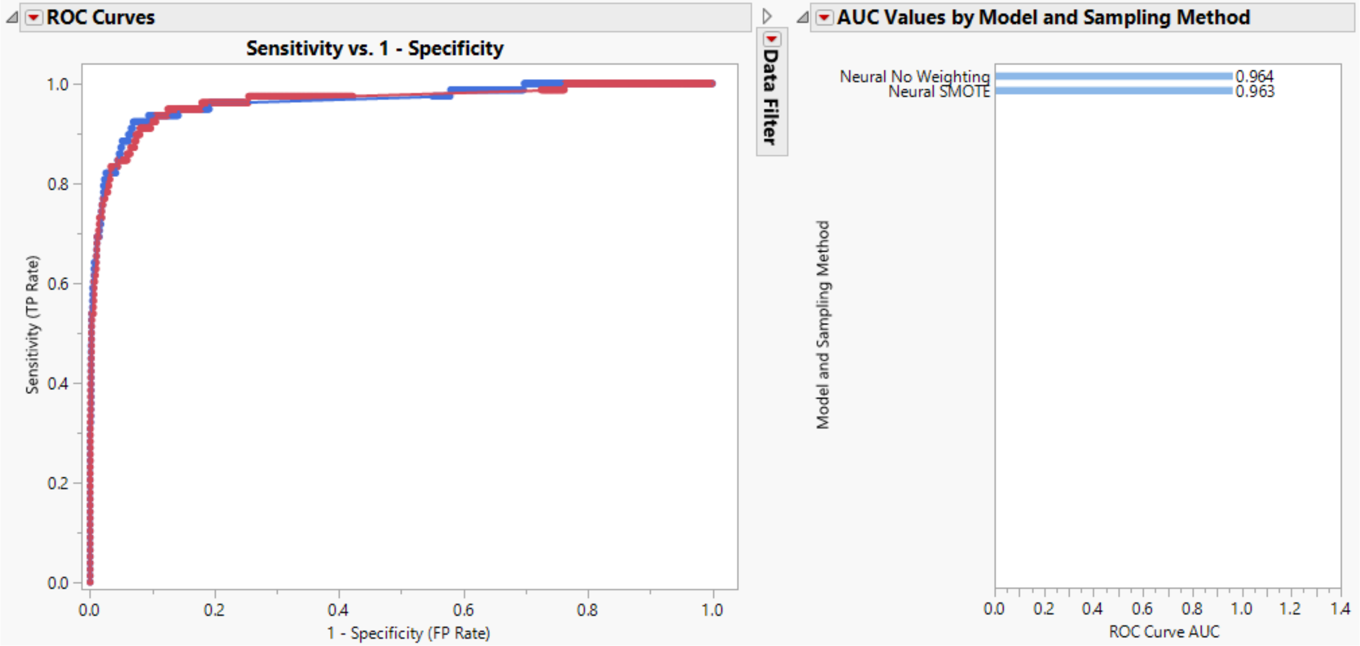


- For the methods and sampling techniques considered, the ROC curves are similar and have high AUC values.
- It is tempting to select Neural No Weighting, or perhaps Neural SMOTE, as the best techniques, as these have the highest AUC values.

Data Sets Studied

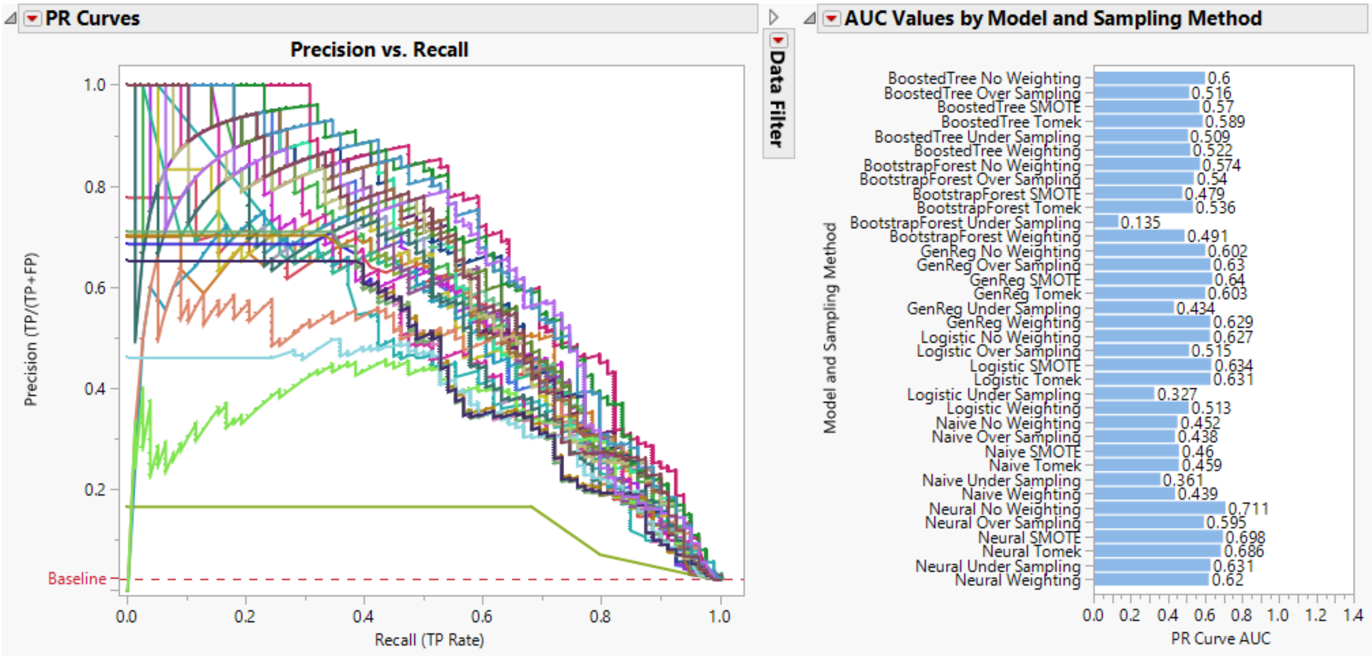
Mammography

- But the ROC curves for Neural No Weighting and Neural SMOTE are very similar. How do you choose between them?



Data Sets Studied

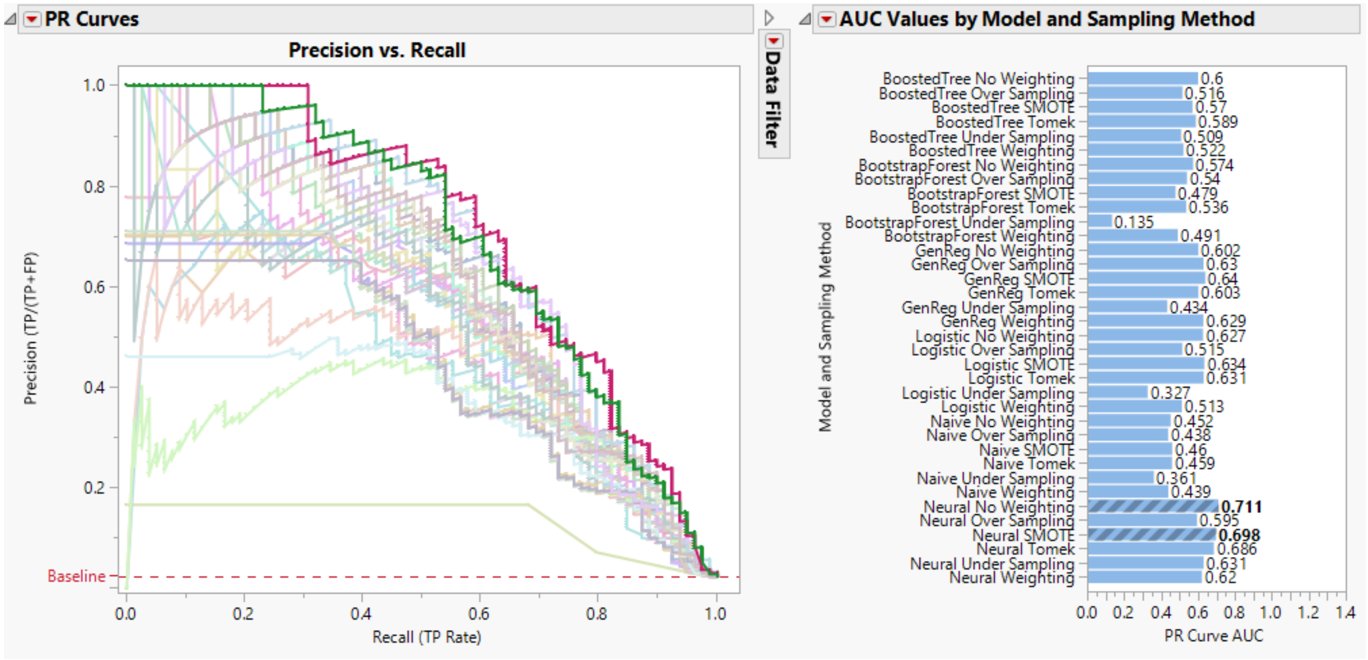
Mammography



- The PR curves differ substantially for the models considered.

Data Sets Studied

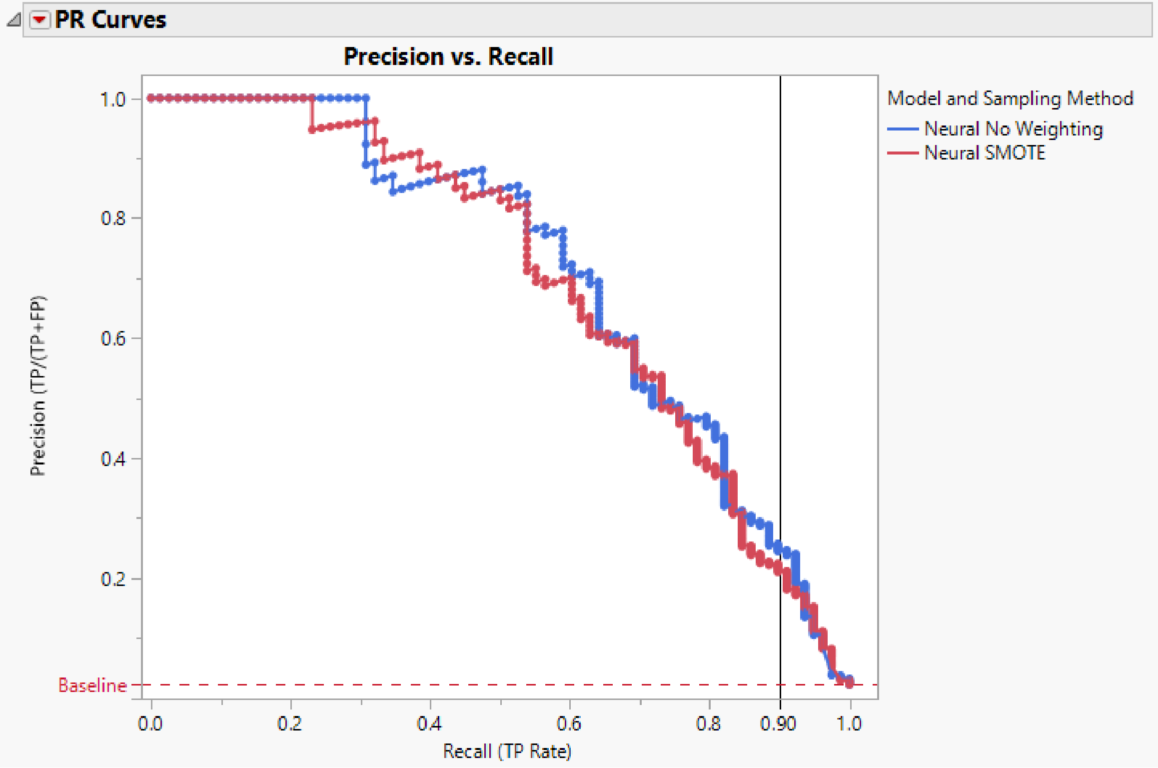
Mammography



- In particular, the PR curves for Neural No Weighting and Neural SMOTE differ.
- Neural No Weighting has the higher AUC value.

Data Sets Studied

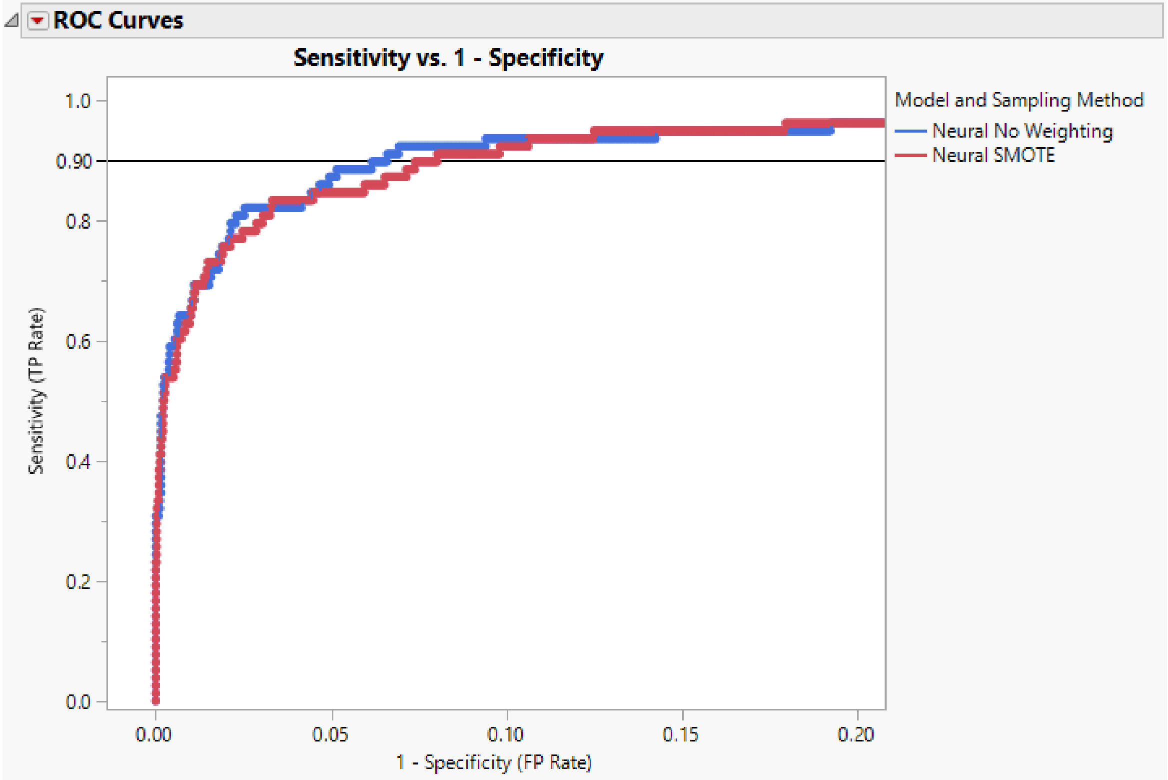
Mammography



- Suppose you are considering a threshold that gives sensitivity (or recall) around 0.90.
- The Neural No Weighting method gives greater precision than the Neural SMOTE method.

Data Sets Studied

Mammography



- To see this difference on the ROC curve, you would have to expand the horizontal scale.

Data Sets Studied

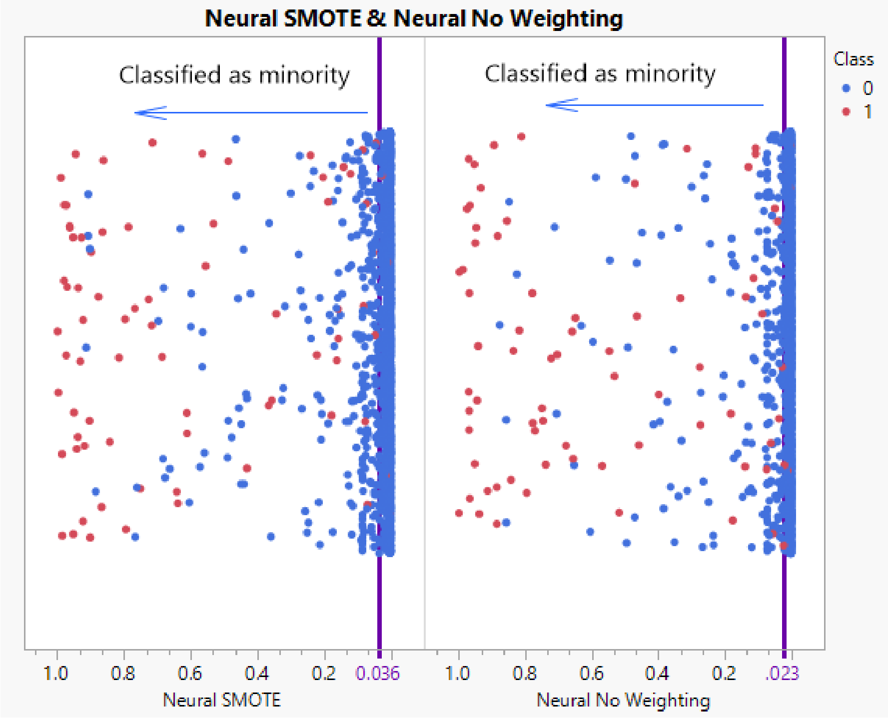
Mammography

- From the Techniques and Thresholds table, we see that Neural No Weighting is more precise at sensitivity 0.897 than Neural SMOTE.
- For Neural No Weighting, of the 8.5% of cases tested, 24.5% are positive.
- For Neural SMOTE, of the 9.9% of cases tested, 21.0% are positive.
- Neural No Weighting gives higher precision with fewer tests than does Neural SMOTE. It follows that Neural No Weighting has a lower false positive rate (1 – Specificity).

Model and Sampling Method	Probs	Class	Precision	Recall	Sensitivity	1 - Specificity	Cumulative Gains	Portion
Neural No Weighting	0.0227600	0	0.24476	0.897436	0.897436	0.065914	0.89744	0.085246
Neural SMOTE	0.0363622	0	0.21021	0.897436	0.897436	0.080256	0.89744	0.099255

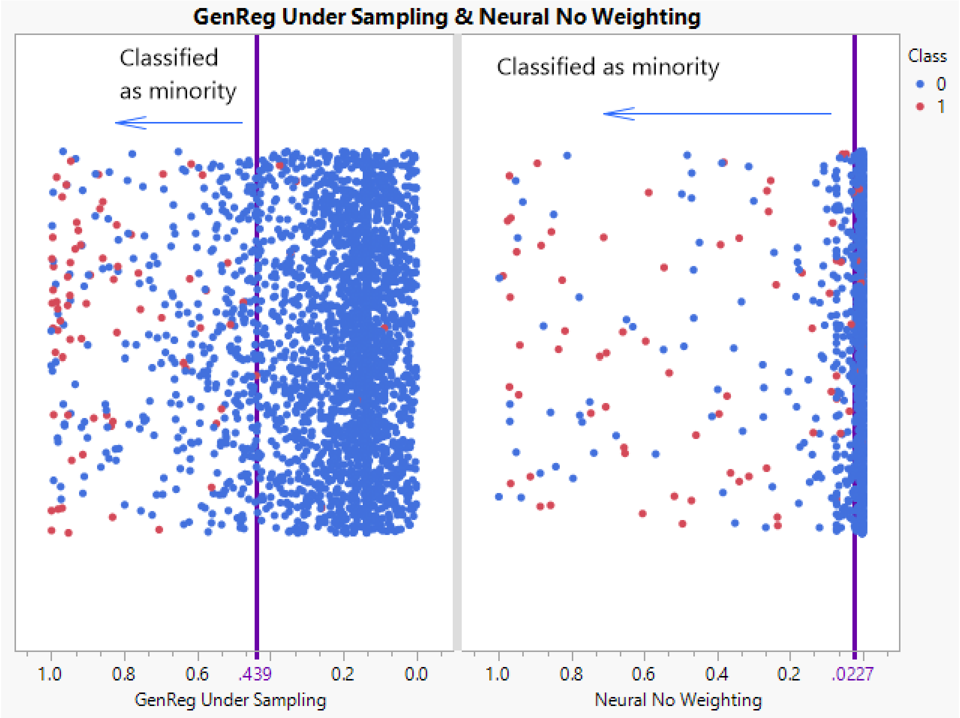
Data Sets Studied

Mammography



Data Sets Studied

Mammography



- The probabilities of class membership, which define the thresholds, have quite different distributions for the two techniques.
- However, this is not of interest.
- Only the ranking of the thresholds is relevant.

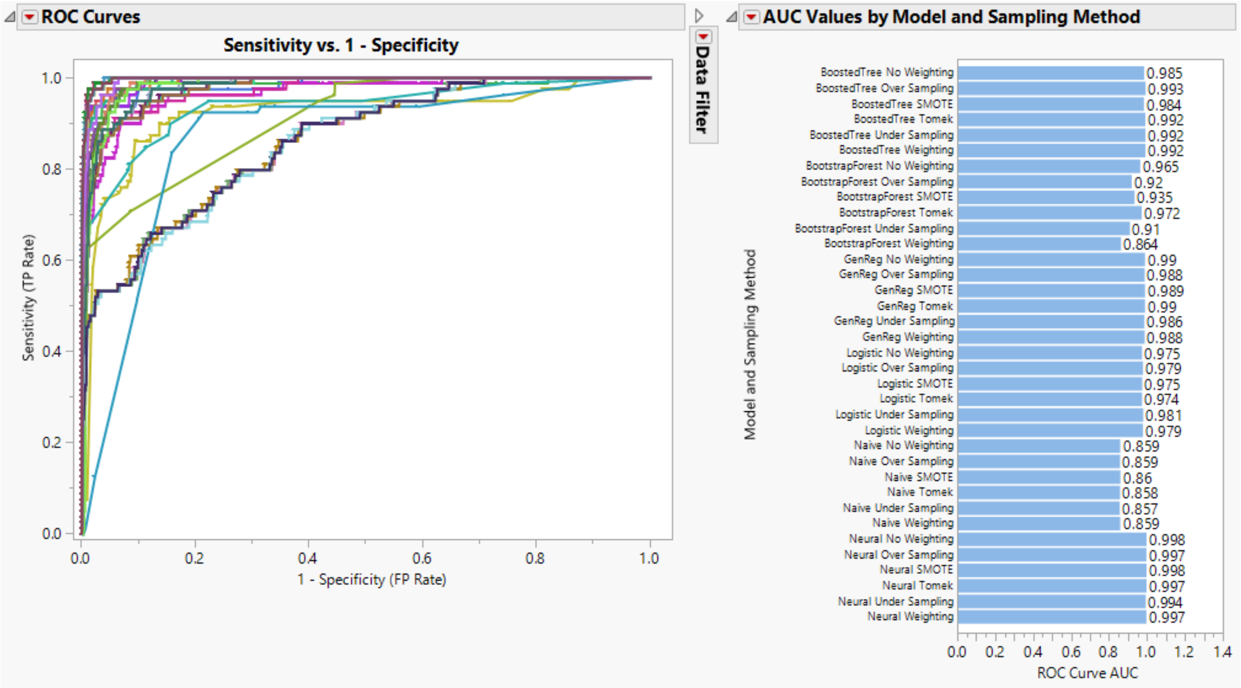
Data Sets Studied

Wilt

- Wilt.jmp contains data from a remote sensing study.
- The study involved detecting diseased trees using Quickbird satellite imagery.
- The data set consists of five continuous variable measuring various aspects of image segments.
- The binary response categorizes each image as containing diseased trees or not.
- There are 4,839 images.

Data Sets Studied

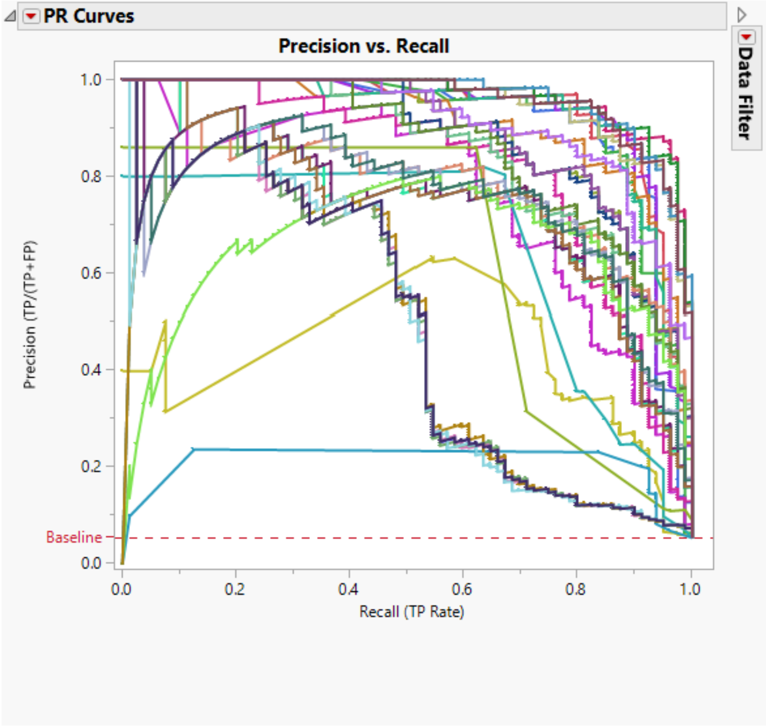
Wilt

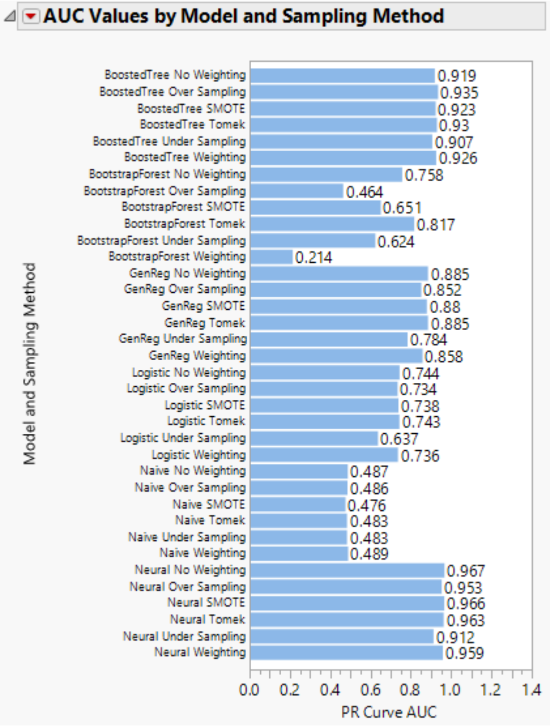


- The model accounts for differences in ROC curves and AUC values, with Naïve Bayes and Bootstrap Forest not performing as well as other models.
- Neural models appear to perform the best.
- Sampling technique has little effect, except for Bootstrap Forest.

Data Sets Studied

Wilt

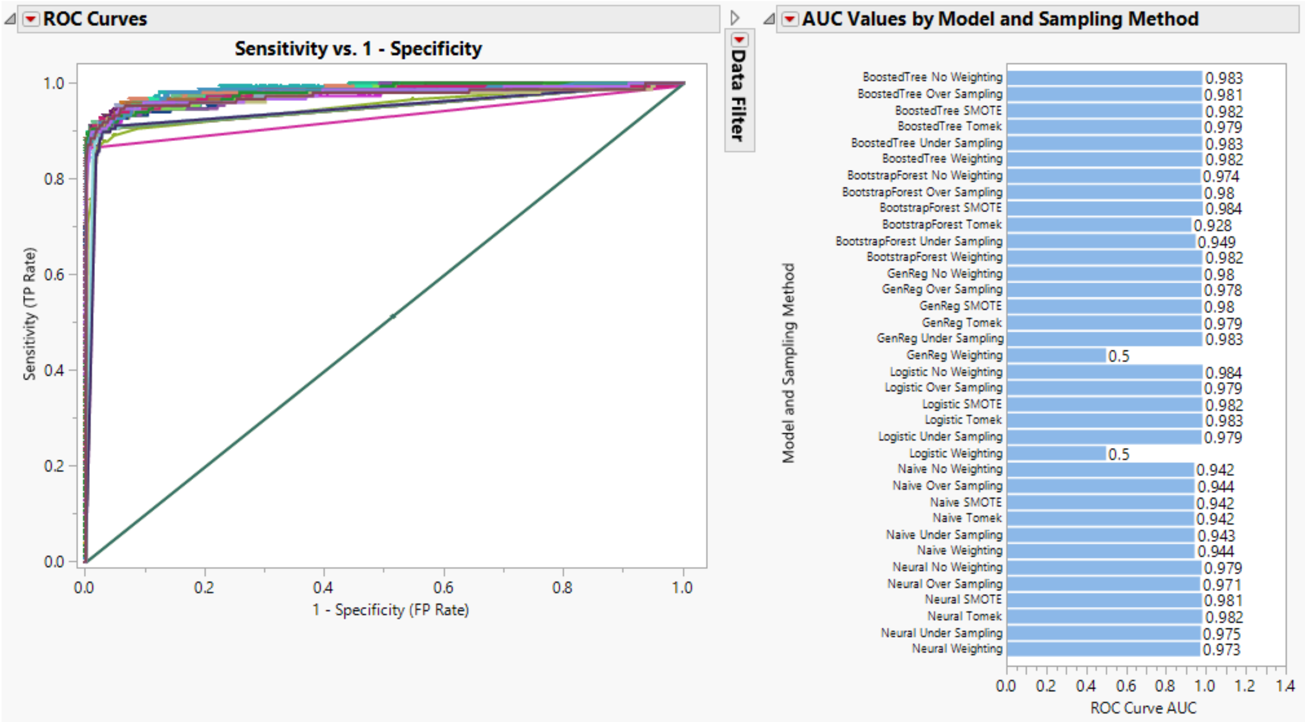




- Differences are more apparent for PR curves and their AUC values.
- Although model seems to have the largest impact, sampling technique has an effect as well.

Data Sets Studied

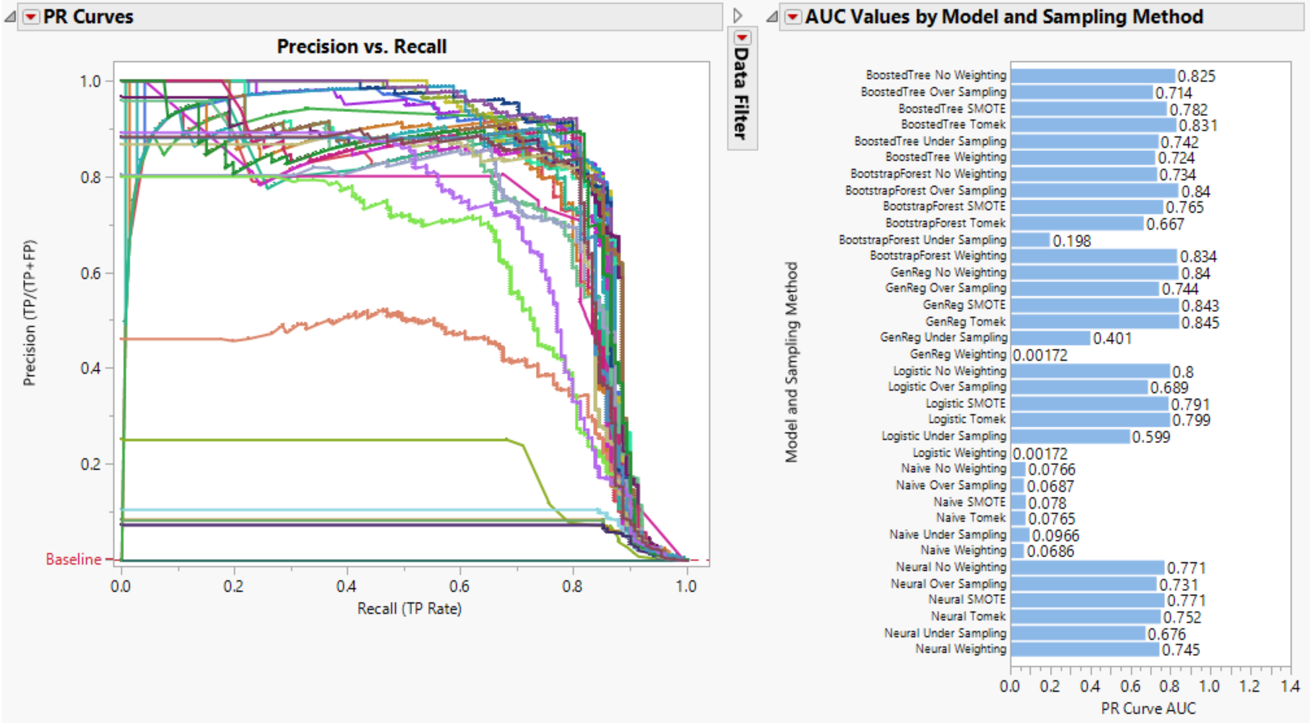
Credit Card Fraud



- The ROC curves and their AUC values show little difference among models, other than for Naïve Bayes.
- The curves and AUC values show virtually no differences among sampling technique, other than for Weighting.

Data Sets Studied

Credit Card Fraud



- The PR curves and their AUC values show major differences both among models and sampling techniques.
- Some models and sampling techniques identify the top-scored 85% or so of minority observations with much higher precision than others.

Data Sets Studied

Assessment of Differences

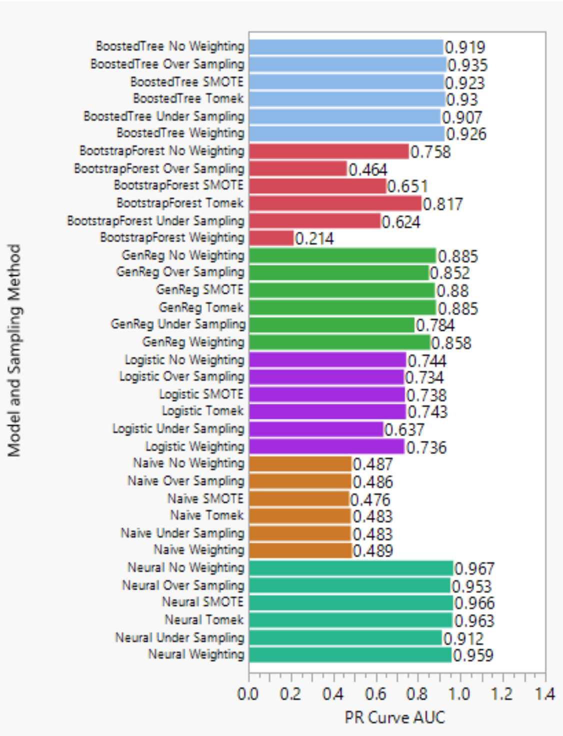
- As expected, differences between PR and ROC curves are most evident for data sets with a small minority representation.
- For such data sets, PR curves are more informative than ROC curves.

	Data Set	N Predictors	N Continuous	N Nominal	N	Minority %	Informal Assessment of ROC vs PR Curve Differences
1	Ionosphere	34	32	2	351	35.90%	Minor
2	Pima Indians	8	8	0	768	34.90%	Some
3	Diabetes Modified	10	9	1	442	27.38%	Some
4	Ecoli	7	7	0	336	22.92%	Major
5	New Thyroid	5	5	0	215	16.28%	Minor
6	Seismic	18	14	4	2584	6.58%	Some, but no models perform well
7	Wilt Data	5	5	0	4839	5.39%	Major
8	Mammography	6	6	0	11183	2.32%	Major
9	Credit Card Fraud	30	30	0	284807	0.17%	Major

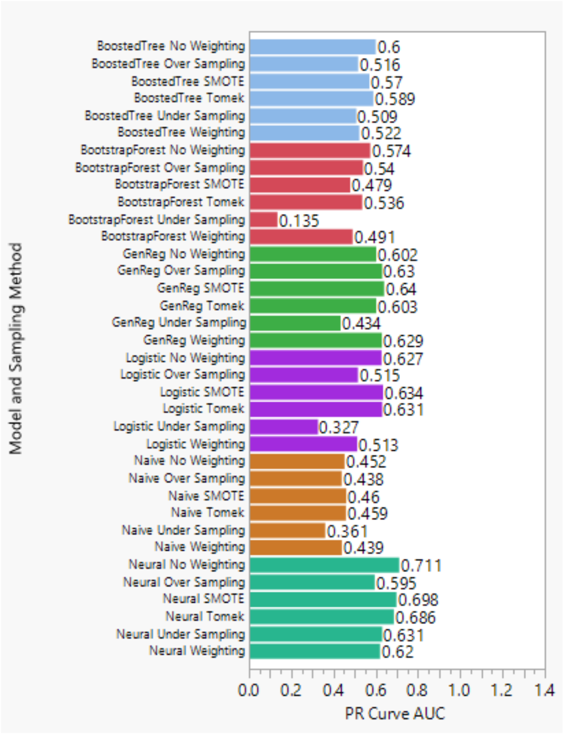
Data Sets Studied

Minority Proportion < ~0.05

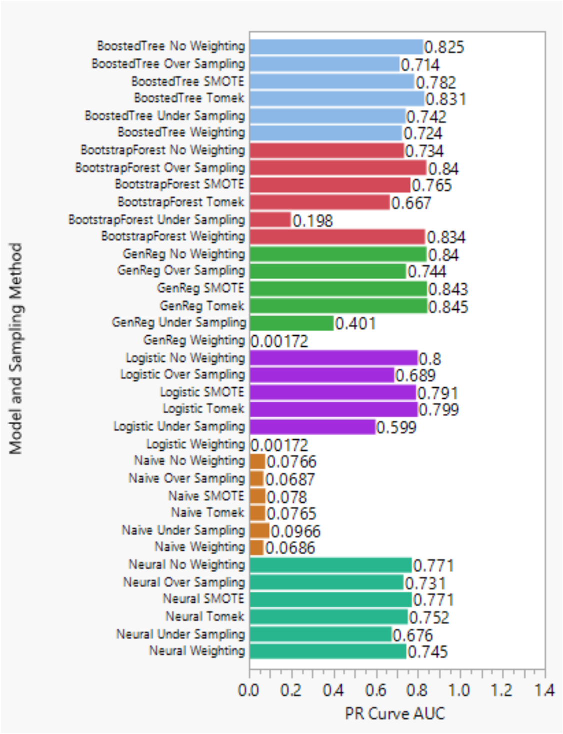
Wilt (5.39%)



Mammography (2.32%)



Credit Card Fraud (0.17%)

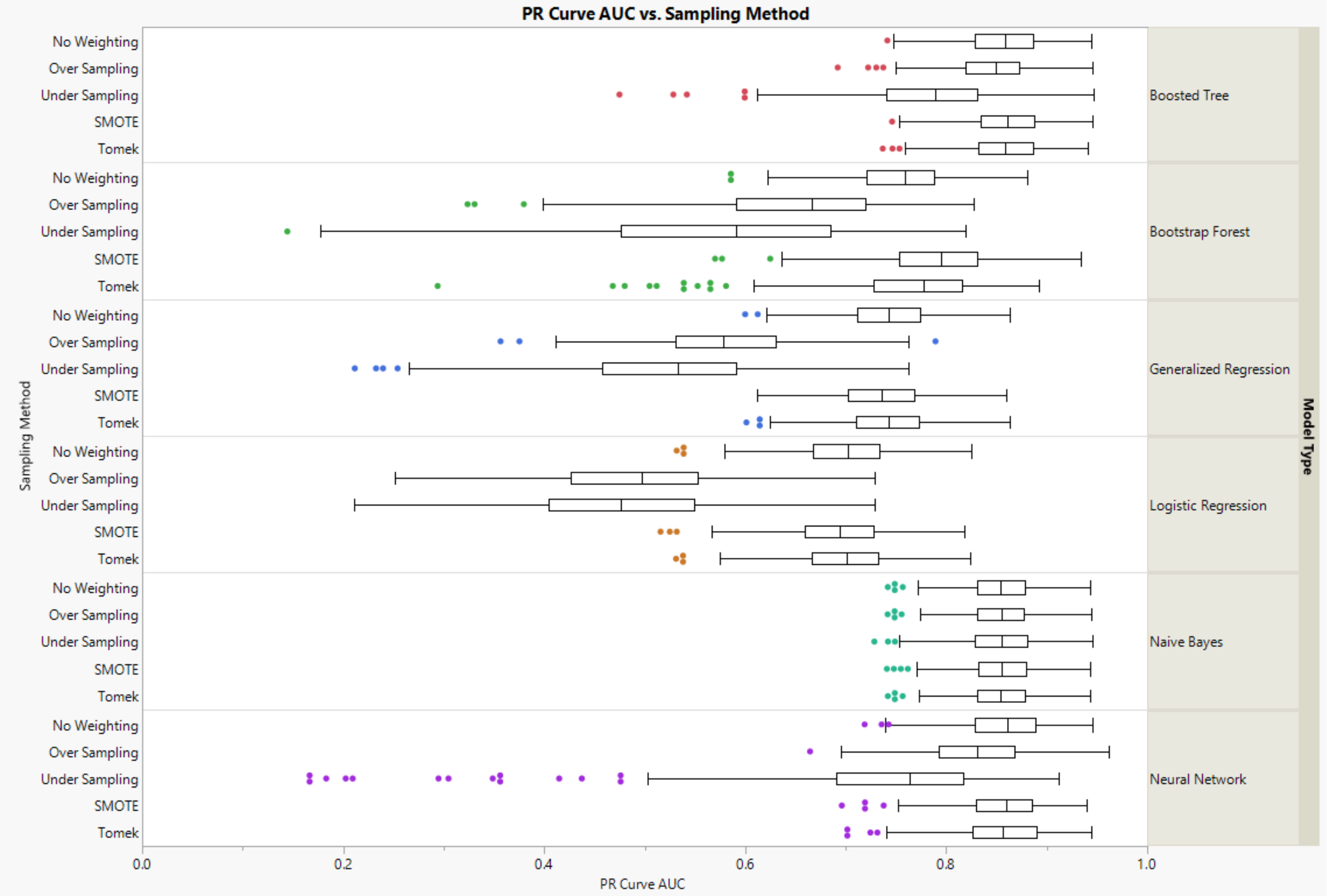


Data Driven Simulations

Structure

- Simulations based on two of the studied 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

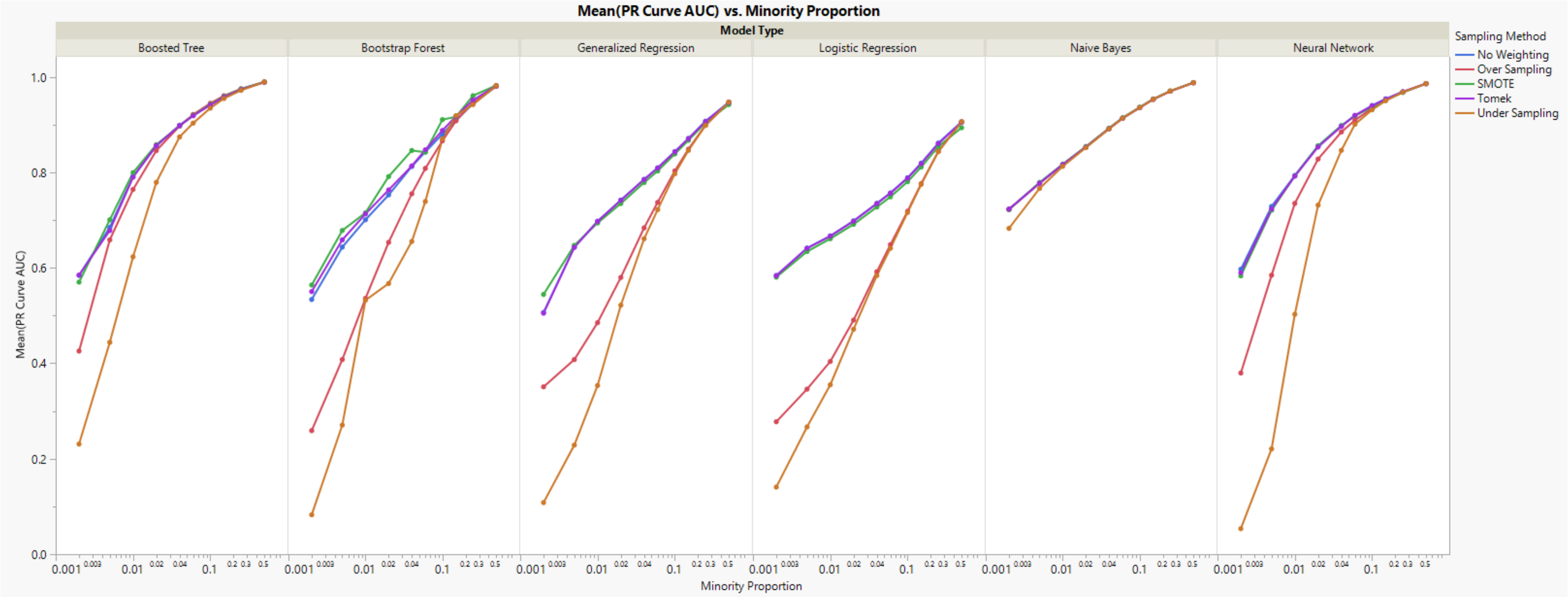
Simulations Based on Mammography Data



2% minority proportion and original mean vector

Simulations Based on Mammography Data

Original mean vector

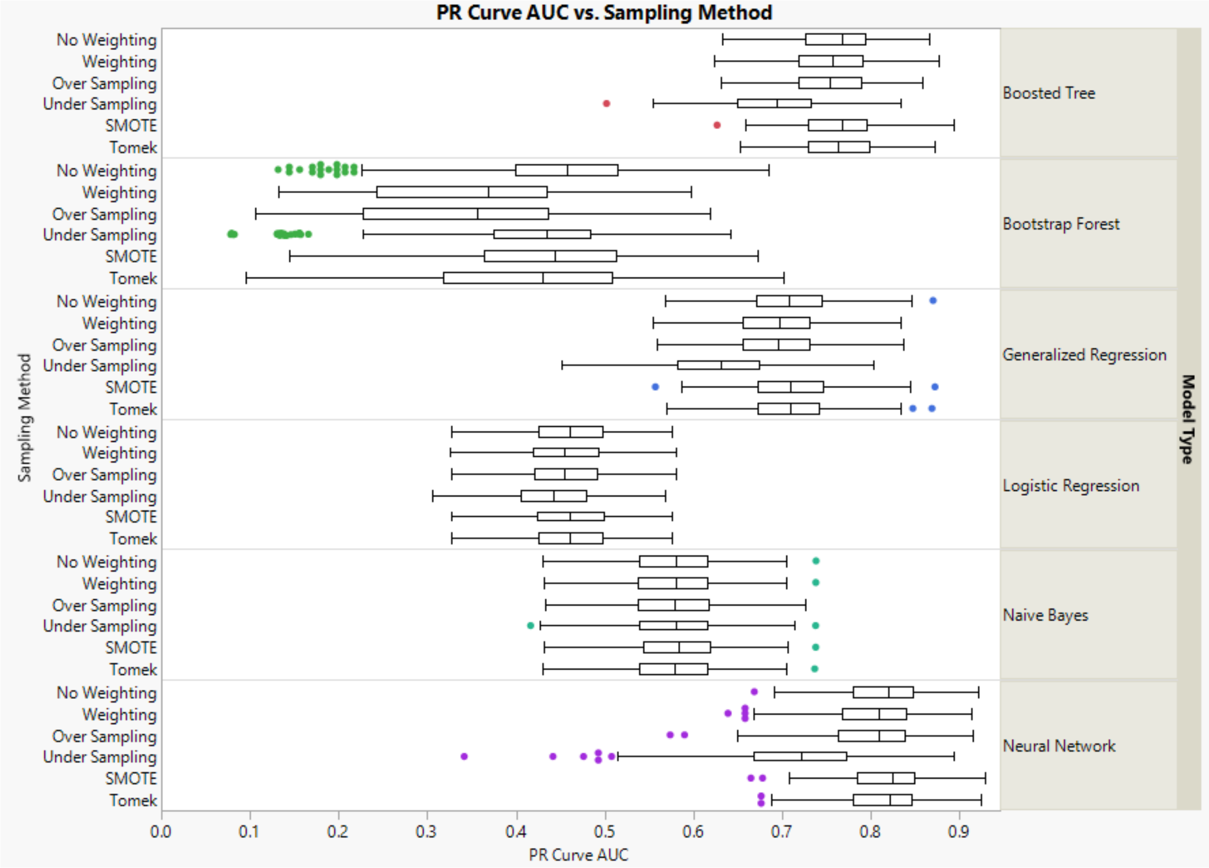


Simulations Based on Mammography Data

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.

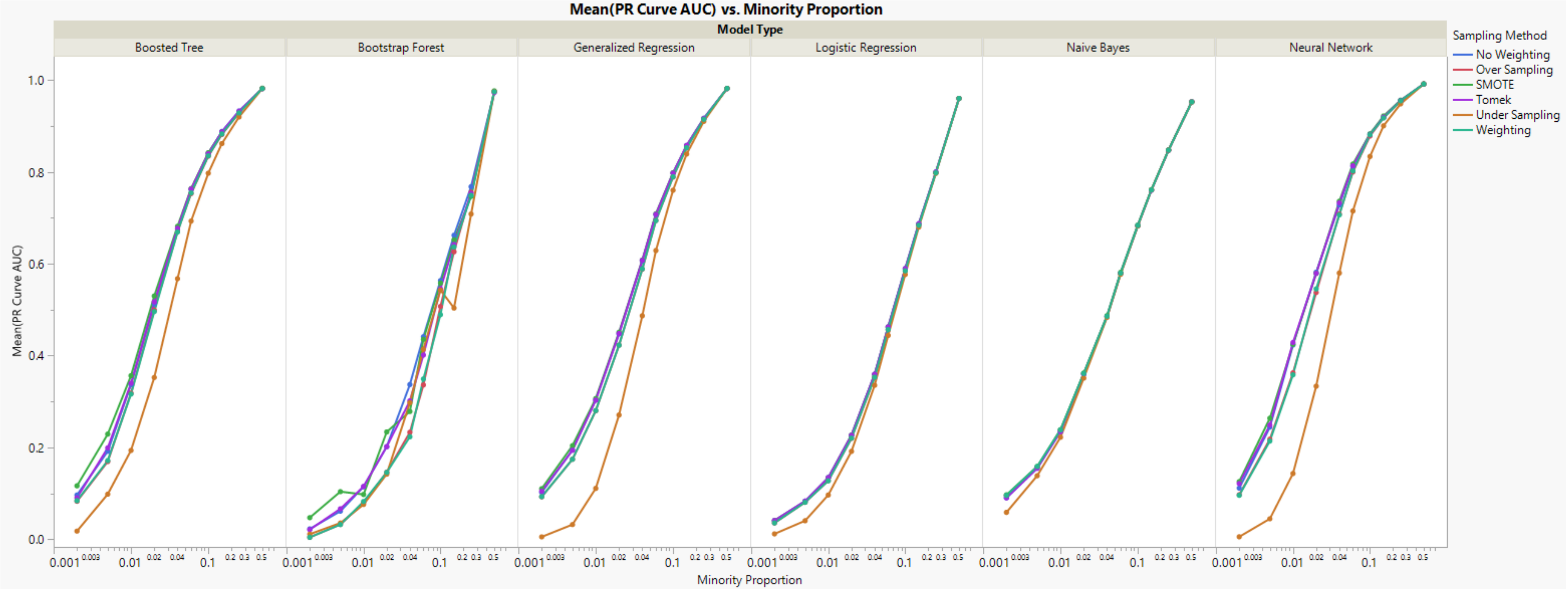
Simulations Based on Wilt Data



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.

Future Work

- Extend the Imbalanced Data script:
  - Add new models: SVM
  - Add new sampling methods: combined SMOTE/Tomek
  - Allow categorical predictors for SMOTE, Tomek, and SMOTE/Tomek sampling methods.
  - Add model specification options
    - Generalized Regression: validation and estimation methods
    - Tree models: tree and resampling specification options
    - Neural nets: multiple hidden layers, boosting
- Study cases where there are more predictors than observations (n < p)

Possible Simulation Study Extensions

- Use different covariance structures.
- Standardize the distances between means.
- Explore the impact of dimensionality.
- Explore model specifications and model options for a specific class of models, perhaps Gen Reg.

Be able to better answer the question: “At what point are my data so imbalanced that I need to worry about the imbalance?”

References

- Chawla, N. V., et al. (2002). “SMOTE: Synthetic Minority Over-sampling Technique.” *Journal of Artificial Intelligence Research*, 16, pp. 321-357.
- Davis, J., and Goadrich, M. (2006). “The Relationship between Precision-Recall and ROC Curves.” *Proceedings of the 23rd International Conference on Machine Learning*.
- 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.
- 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.
- 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*.
- 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.
- 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).

Thanks!

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

sas.com