



2016-02-22

DATA ANALYTICS

Assignment 3

Assignment 3: Validation and Testing

Using Software model: GeNIe

Using Data files: House Votes Manual.xdsl

House Votes Naive.xdsl

House Votes PC.xdsl

house-votes-84.txt

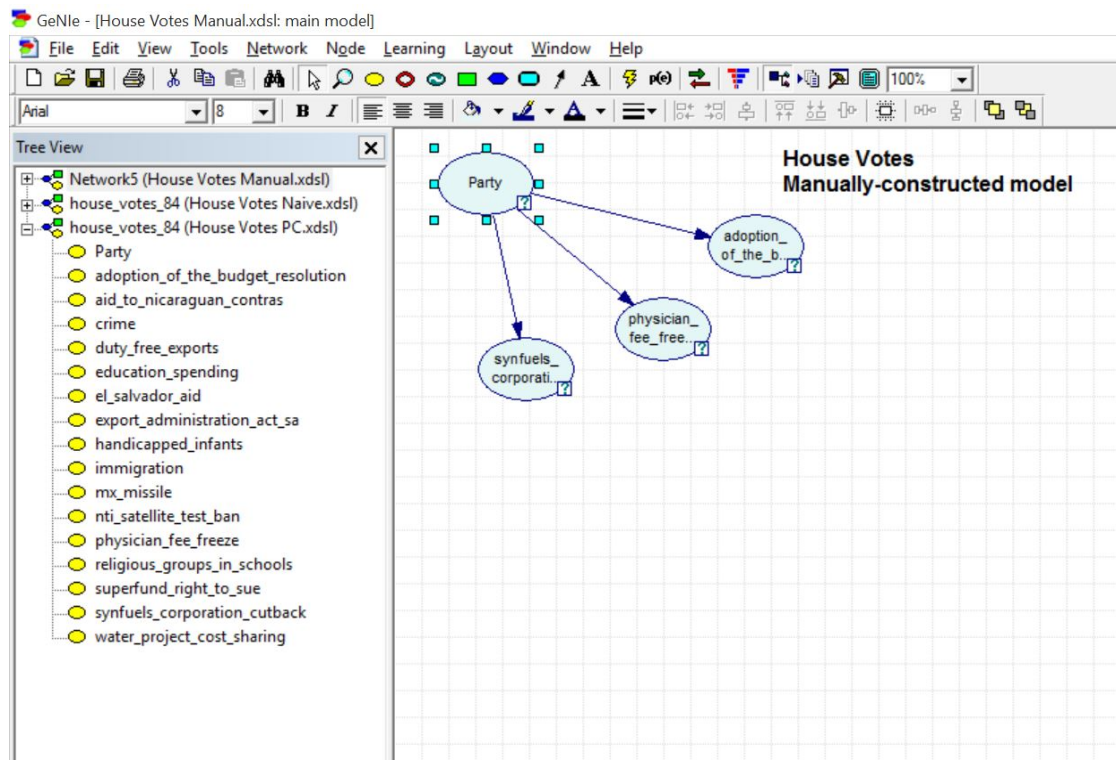
Group members: Nanxun Xie (nax4)

Yifan Zhao(yiz105)

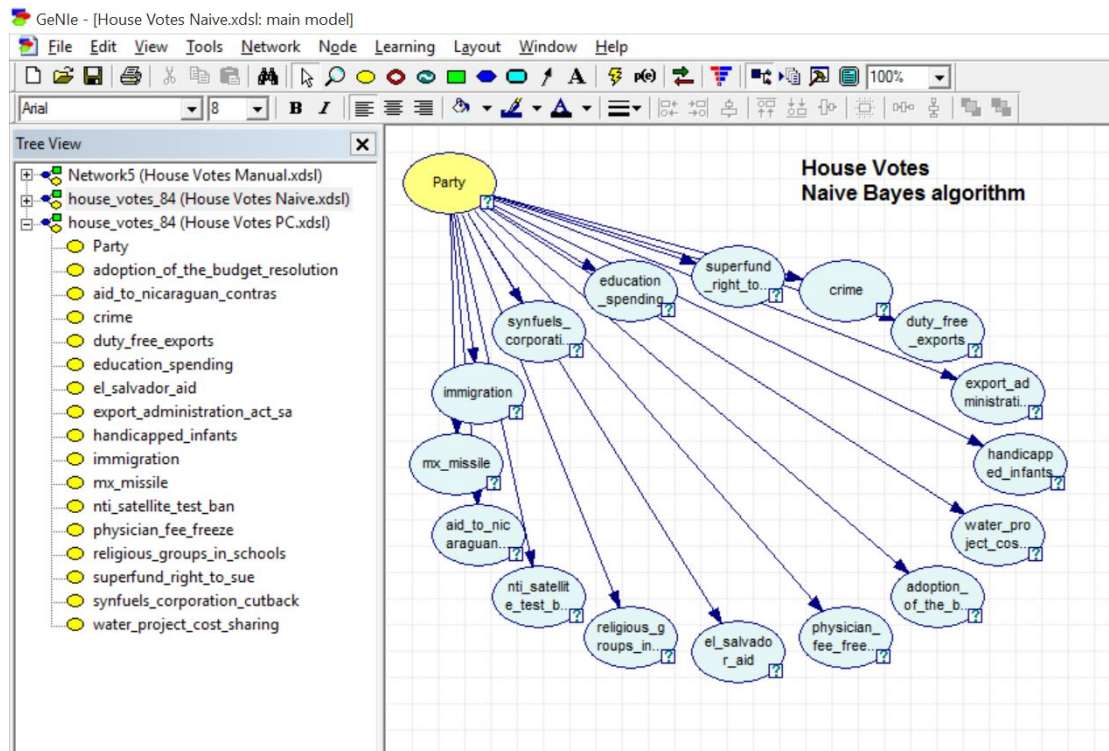
Import three models and data files into GeNIe:

[illegible]

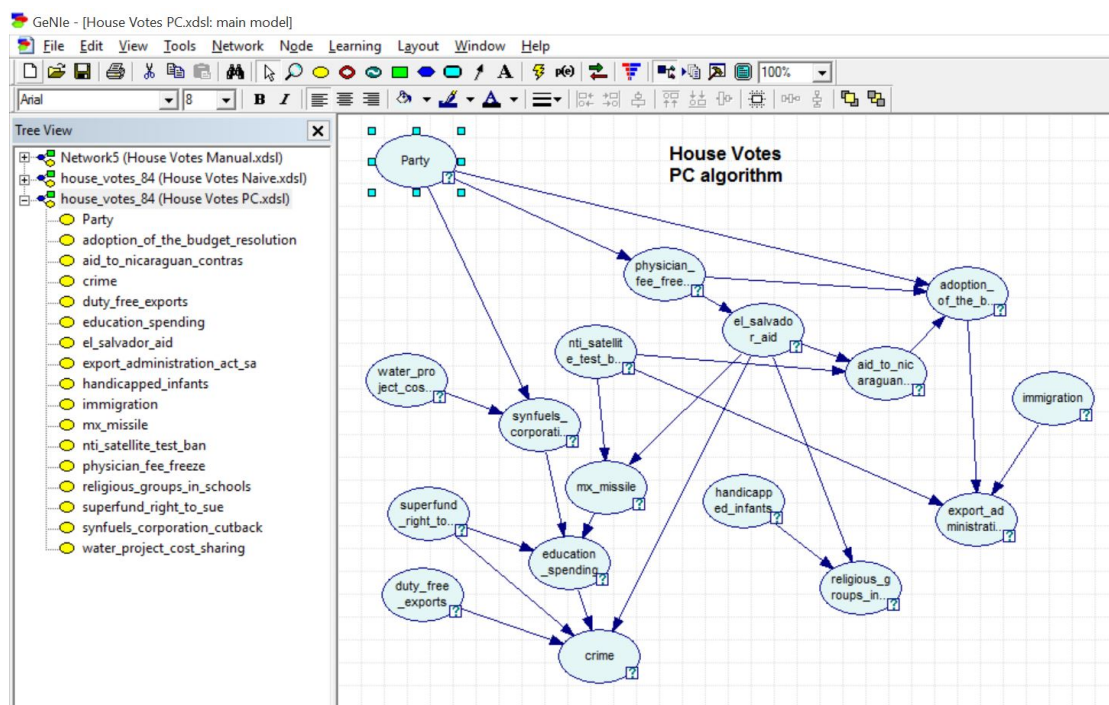
Import Data file



Import model 1



Import model 2



Import model 3

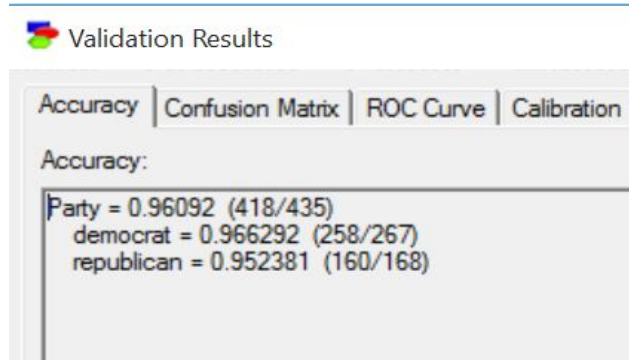
Answer Questions

Validation and testing(Using GeNIe):

(1) Overall classification accuracy:

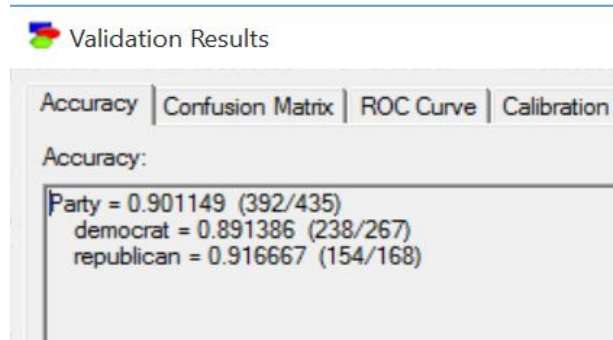
After importing the three models and data file into the GeNIe, for each model of three, click on the learning button above the GeNIe, click on validate, then OK, change the Validation method from original "Test-only" to "leave-one-out". Besides, as we've known in the question, our target is to guess the party affiliation of the representative, hence, we choose the party item of class nodes below the method to validate, then, click on OK without changing other options.

a) The validation result of 'House Votes Manual.xdsl' is shown below:



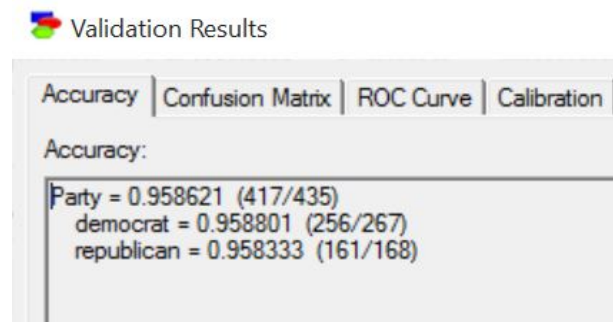
The classification accuracy of Manual model is 0.96092≈96%.

b) The validation result of 'House Votes Naive.xdsl' is shown below:



The classification accuracy of Naive model is 0.901149≈90%

c) The validation result of 'House Votes PC.xdsl' is shown below:



The classification accuracy of PC model is 0.958621≈96%

Observation: the classification accuracy of PC model and Manual model are almost the same and they are both bigger than the accuracy of Naive model. Therefore, the Manual model and PC model are more accurate.

(2) Sensitivity and specificity for each of the two parties:

To calculate the sensitivity and specificity of two parties, we should first derive the Confusion Matrix of class node: Party. As we output the validation results in GeNIe, we can find the confusion matrix panel on it.

Confusion matrix

The same thing as what we saw before but used with reference to the model's predictions and the true state of the World.

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = $TP / (TP + FP)$ = $20 / (20 + 180)$ = 10%
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = $TN / (FN + TN)$ = $1820 / (10 + 1820)$ ≈ 99.5%
		Sensitivity = $TP / (TP + FN)$ = $20 / (20 + 10)$ ≈ 67%	Specificity = $TN / (FP + TN)$ = $1820 / (180 + 1820)$ = 91%	

As we know, the test file has 2 parties: democrat & republican. Besides, due to the formula of sensitivity and specificity calculation as we known also, we can get the results below:

a). The confusion matrix of Manual model is:

Validation Results

Accuracy	Confusion Matrix	ROC Curve	Calibration
Class node: Party			
	democrat	republican	
democrat	258	9	
republican	8	160	

The sensitivity of the democrat in Manual model is:

$$\text{Sensitivity} = TP / (TP + FN) = 258 / (258 + 8) = 0.96992481 \approx 97\%$$

The specificity of the democrat in Manual model is:


$$\text{Specificity} = TN / (FP + TN) = 160 / (9 + 160) = 0.94674556 \approx 95\%$$

The sensitivity of the republican in Manual model is:

$$\text{Sensitivity} = TP / (TP + FN) = 160 / (160 + 9) = 0.94674556 \approx 95\%$$

The specificity of the republican in Manual model is:
 $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = 258 / (258 + 8) = 0.96992481 \approx 97\%$


b). The confusion matrix of Naive model is:

 Validation Results

Validation Results		
Accuracy Confusion Matrix ROC Curve Calibration		
Class node: Party		
	democrat	republican
democrat	238	29
republican	14	154

The sensitivity of the democrat in Naive model is:
 $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 238 / (238 + 14) = 0.94444444 \approx 94\%$
The specificity of the democrat in Naive model is:
 $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = 154 / (29 + 154) = 0.84153005 \approx 84\%$
The sensitivity of the republican in Naive model is:
 $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 154 / (154 + 29) = 0.84153005 \approx 84\%$
The specificity of the republican in Naive model is:
 $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = 238 / (238 + 14) = 0.94444444 \approx 94\%$

c). The confusion matrix of PC model is:

 Validation Results

Validation Results		
Accuracy Confusion Matrix ROC Curve Calibration		
Class node: Party		
	democrat	republican
democrat	256	11
republican	7	161

The sensitivity of the democrat in PC model is:
 $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 256 / (256 + 7) = 0.97338403 \approx 97\%$
The specificity of the democrat in PC model is:
 $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = 161 / (11 + 161) = 0.93604651 \approx 94\%$
The sensitivity of the republican in PC model is:
 $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) = 161 / (161 + 11) = 0.93604651 \approx 94\%$
The specificity of the republican in PC model is:
 $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) = 256 / (256 + 7) = 0.97338403 \approx 97\%$

Observation: Just like what we see above, the sensitivity of positive value in model exists some kinds of connection with the specificity of negative value in model, the value is the same after the calculation. Besides, the sensitivity and specificity of Manual or PC model is larger than it of Naïve model.

(3) Positive and negative predictive value for each of the two parties:

Confusion matrix

The same thing as what we saw before but used with reference to the model's predictions and the true state of the World.

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = $TP / (TP + FP)$ = $20 / (20 + 180)$ = 10%
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = $TN / (FN + TN)$ = $1820 / (10 + 1820)$ ≈ 99.5%
		Sensitivity = $TP / (TP + FN)$ = $20 / (20 + 10)$ ≈ 67%	Specificity = $TN / (FP + TN)$ = $1820 / (180 + 1820)$ = 91%	

As the pictures we got before,

Validation Results

Accuracy	Confusion Matrix	ROC Curve	Calibration
Class node: Party			
	democrat	republican	
democrat	258	9	
republican	8	160	

a) The predictive value of two parties in Manual model is:

The positive predictive value of democrat is:

$$\text{Positive predictive value} = TP / (TP + FP) = 258 / (258 + 9) = 0.96629213 \approx 97\%$$

The negative predictive value of democrat is:

$$\text{Negative predictive value} = TN / (FN + TN) = 160 / (8 + 160) = 0.95238095 \approx 95\%$$

The positive predictive value of republican is:

$$\text{Positive predictive value} = TP / (TP + FP) = 160 / (8 + 160) = 0.95238095 \approx 95\%$$

The negative predictive value of republican is:

$$\text{Negative predictive value} = TN / (FN + TN) = 258 / (258 + 9) = 0.96629213 \approx 97\%$$

Validation Results

Accuracy Confusion Matrix ROC Curve Calibration		
Class node: Party		
	democrat	republican
democrat	238	29
republican	14	154

a) The predictive value of two parties in Naive model is:

The positive predictive value of democrat is:

$$\text{Positive predictive value} = \text{TP} / (\text{TP} + \text{FP}) = 238 / (238 + 29) = 0.89138577 \approx 89\%$$

The negative predictive value of democrat is:

$$\text{Negative predictive value} = \text{TN} / (\text{FN} + \text{TN}) = 154 / (14 + 154) = 0.91666667 \approx 92\%$$

The positive predictive value of republican is:

$$\text{Positive predictive value} = \text{TP} / (\text{TP} + \text{FP}) = 154 / (14 + 154) = 0.91666667 \approx 92\%$$

The negative predictive value of republican is:

$$\text{Negative predictive value} = \text{TN} / (\text{FN} + \text{TN}) = 238 / (238 + 29) = 0.89138577 \approx 89\%$$

Validation Results

Accuracy Confusion Matrix ROC Curve Calibration		
Class node: Party		
	democrat	republican
democrat	256	11
republican	7	161

c) The predictive value of two parties in PC model is:

The positive predictive value of democrat is:

$$\text{Positive predictive value} = \text{TP} / (\text{TP} + \text{FP}) = 256 / (256 + 11) = 0.9588015 \approx 96\%$$

The negative predictive value of democrat is:

$$\text{Negative predictive value} = \text{TN} / (\text{FN} + \text{TN}) = 161 / (7 + 161) = 0.95833333 \approx 96\%$$

The positive predictive value of republican is:

$$\text{Positive predictive value} = \text{TP} / (\text{TP} + \text{FP}) = 161 / (7 + 161) = 0.95833333 \approx 96\%$$

The negative predictive value of republican is:

$$\text{Negative predictive value} = \text{TN} / (\text{FN} + \text{TN}) = 256 / (256 + 11) = 0.9588015 \approx 96\%$$

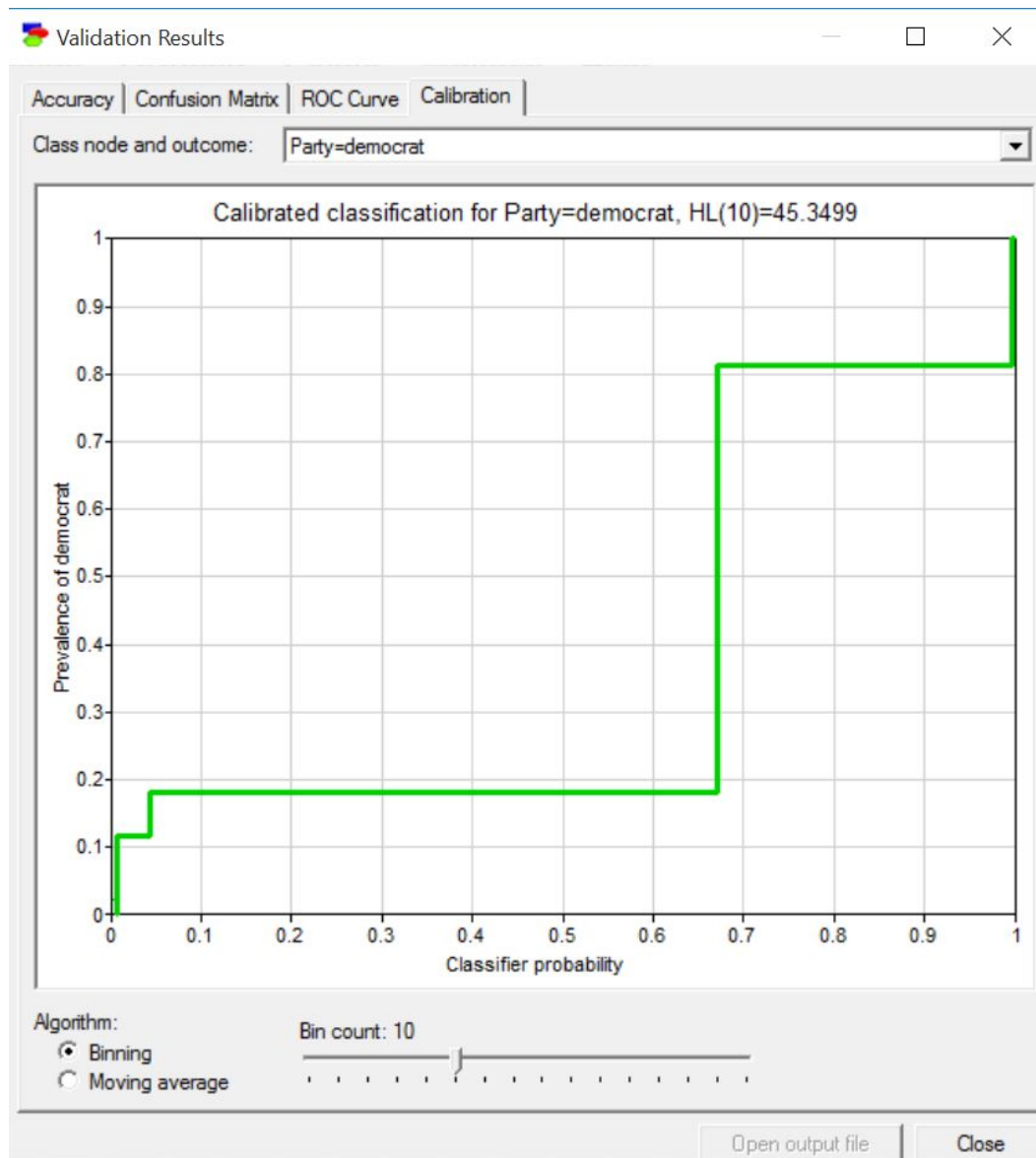
Observation: the positive predictive value of democrat is the same as the negative predictive value of republican, besides, the two predictive values of Manual model and PC model are larger than the value of Naive model.

(4) Calibration curve for a selected bin count or window size:

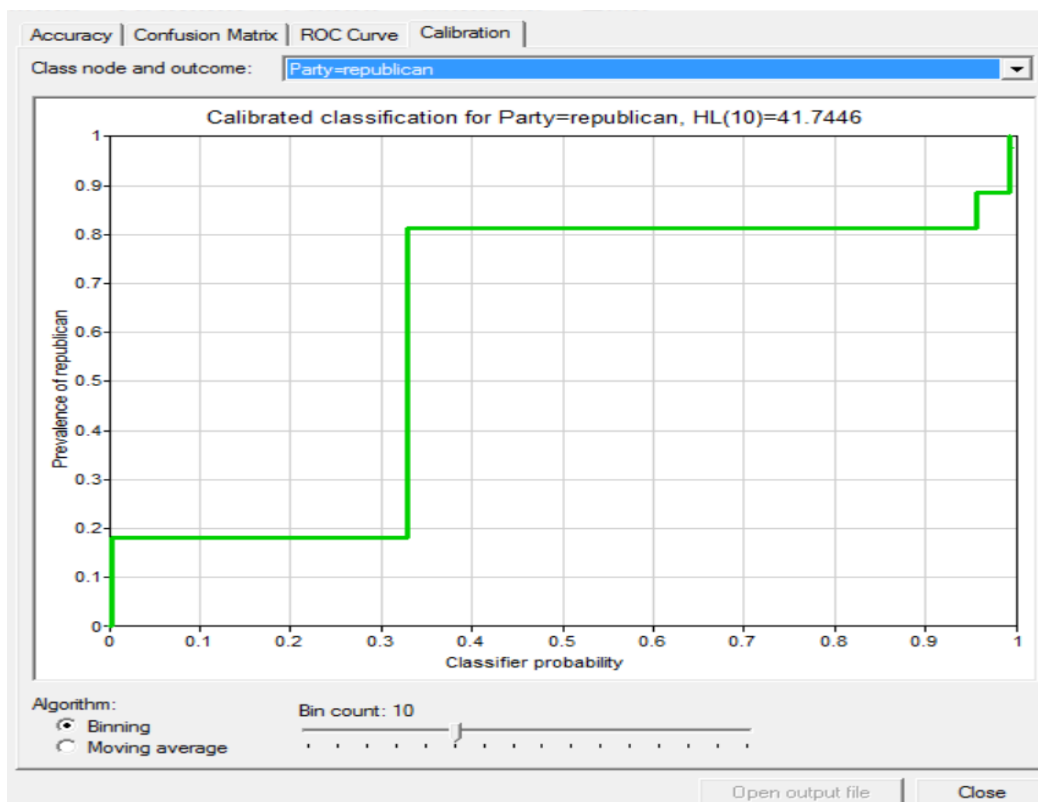
Click on the calibration panel on the validation results.

a) the calibration curve of Manual model:

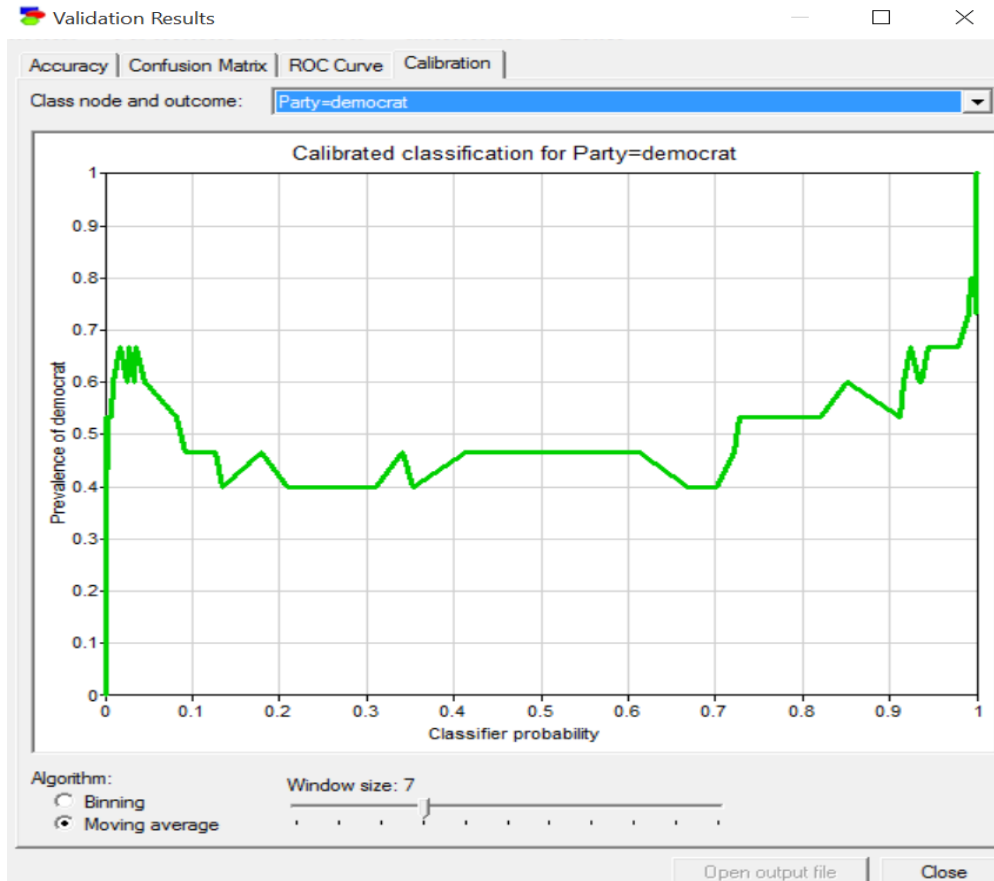
Bin count=10, party=democrat:



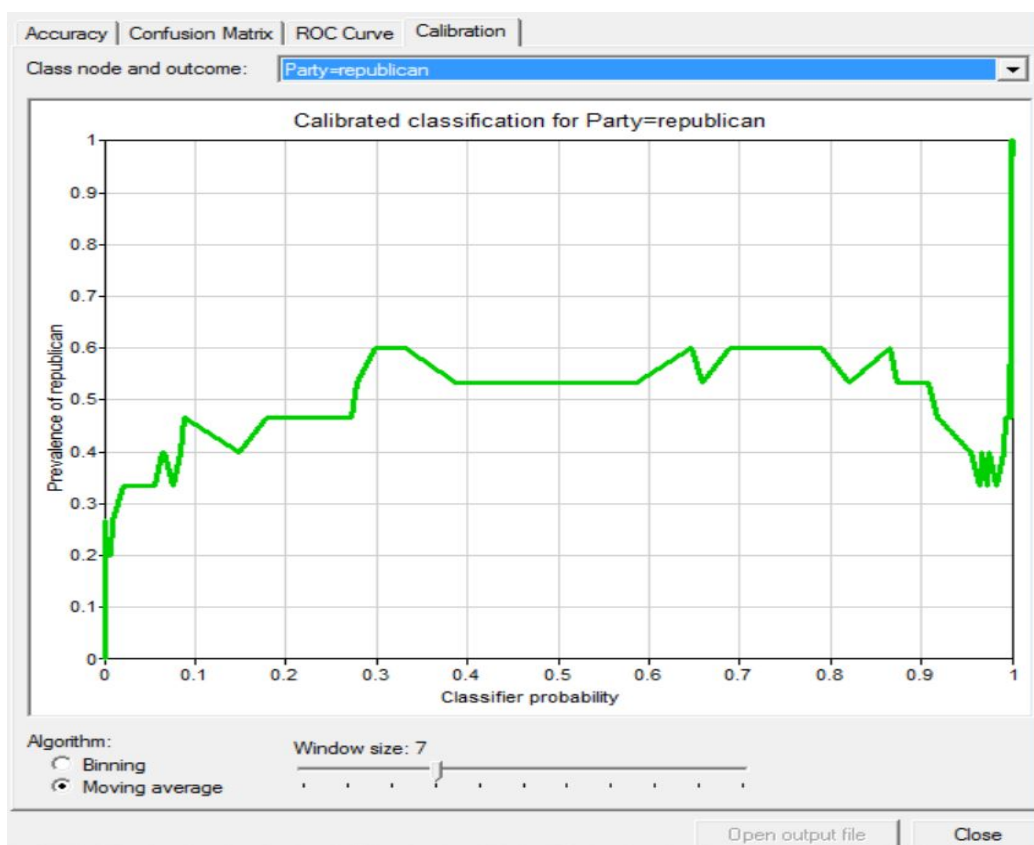
Bin count=10, party=republican:



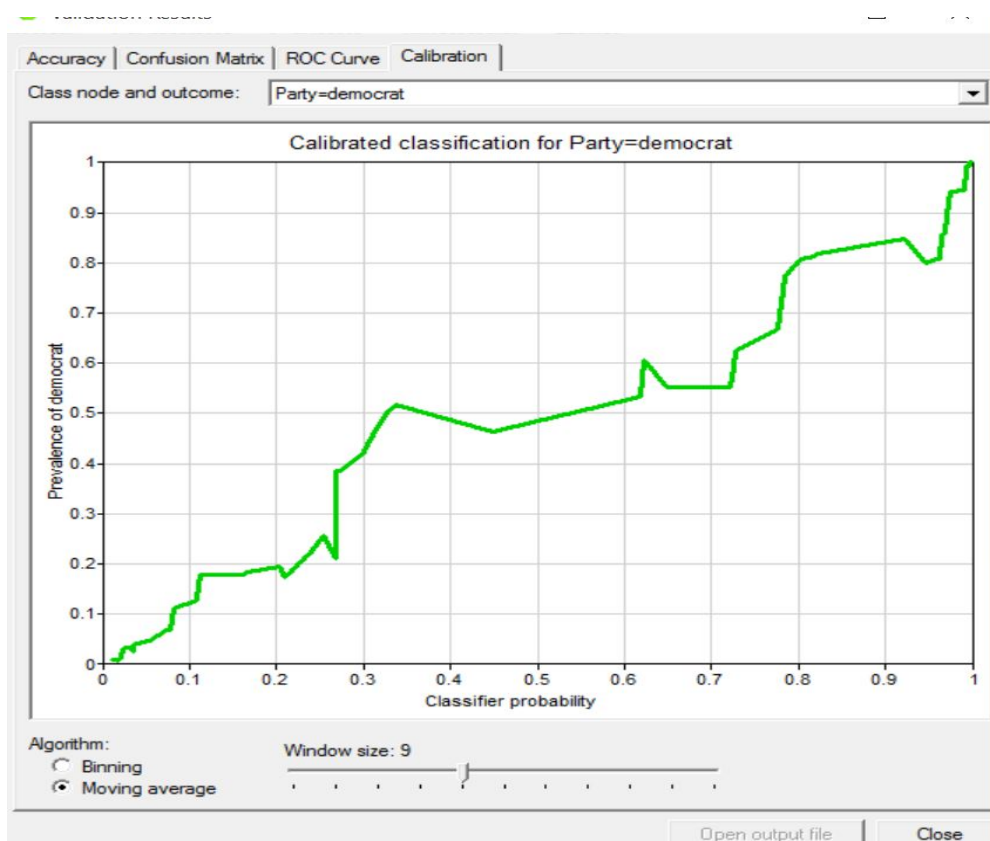
b) the calibration curve of Naive model:
window size=7, party=democrat:



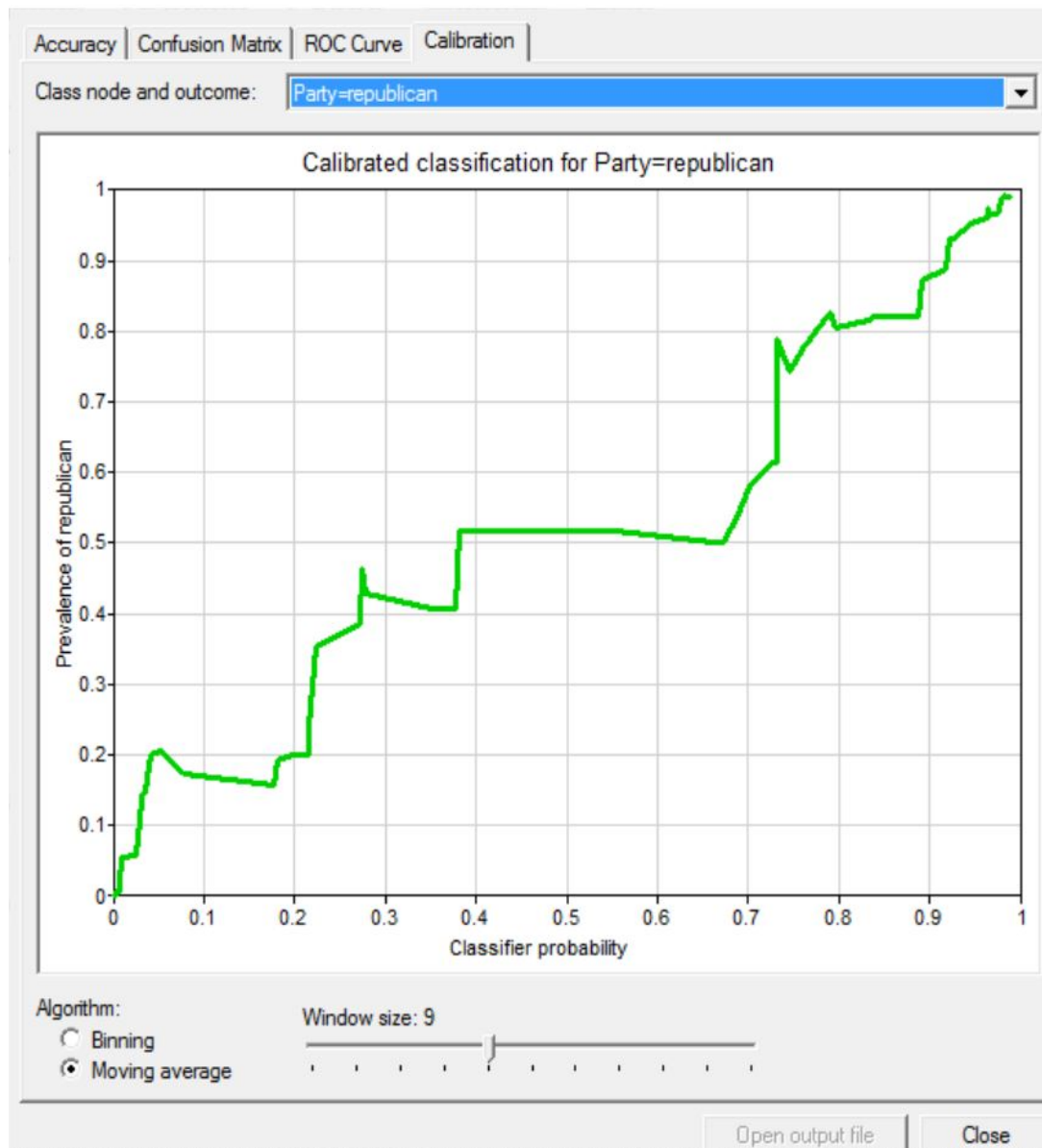
Window size=7, party=republican:



c) the calibration curve of PC model:
window size=9, party=democrat:



window size=9, party=republican:



Observation:

- (1) Firstly, when we use the binning algorithm, the curve is more flat to some extent.
- (2) What's more, as the classifier probability increasing, the prevalence is changing substantially.
- (3) With changing the window size or bin count, the curve is also changing and by a large margin.
- (4) Last, the trends of curves in two different parties are almost similar.