INFSCI 2725 Data Analytics

Assignment 3 - Validation and Testing

Student name: Tong Wei

PITT ID: TOW6

Executive Summary

Evaluation is an essential part in data analysis task as it is of importance to validate your results or justify your conclusion. This report aims to utilize "leave-one-out" cross validation to evaluate the performance of three classification models including Manually-constructed model, Naïve Bayes algorithm and PC Algorithm, all of which are applied to analyze house votes dataset. This dataset contains 435 records of United States representatives where each record shows how the representative voted on each of 16 different issues. In implementing classification models, party affiliation of the representative (Democratic and Republican) is the response variable while the rest 16 different issues are predictors. Figure 1, Figure 2 and Figure 3 are the tree views for three models. At first, accuracy, sensitivity and specificity are used to evaluate these three models respectively, followed by the presentation of positive and negative predictive value for each of the two parties. Finally, the calibration curve for different models are illustrated and compared.

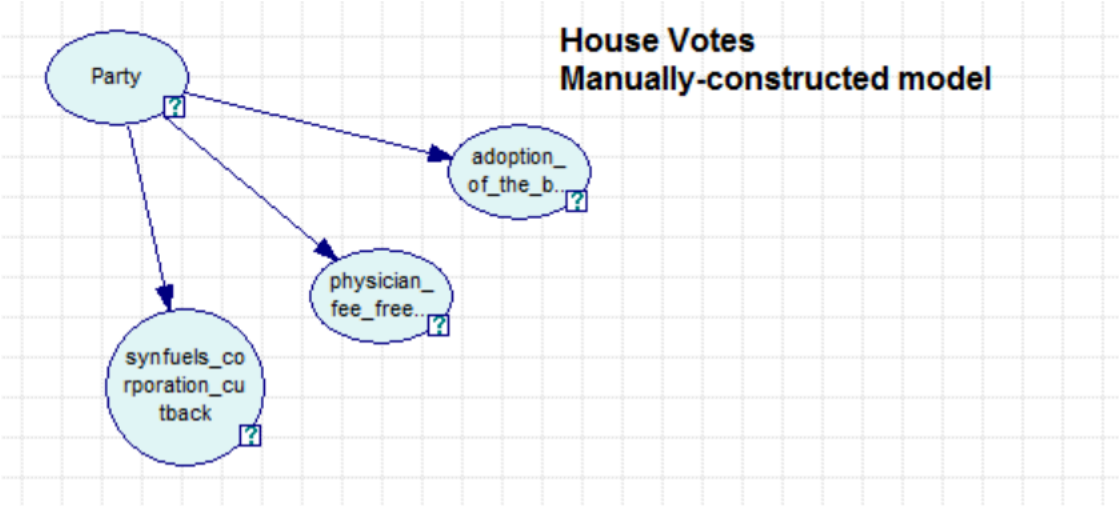

Figure 1: Tree View for Manually-constructed model

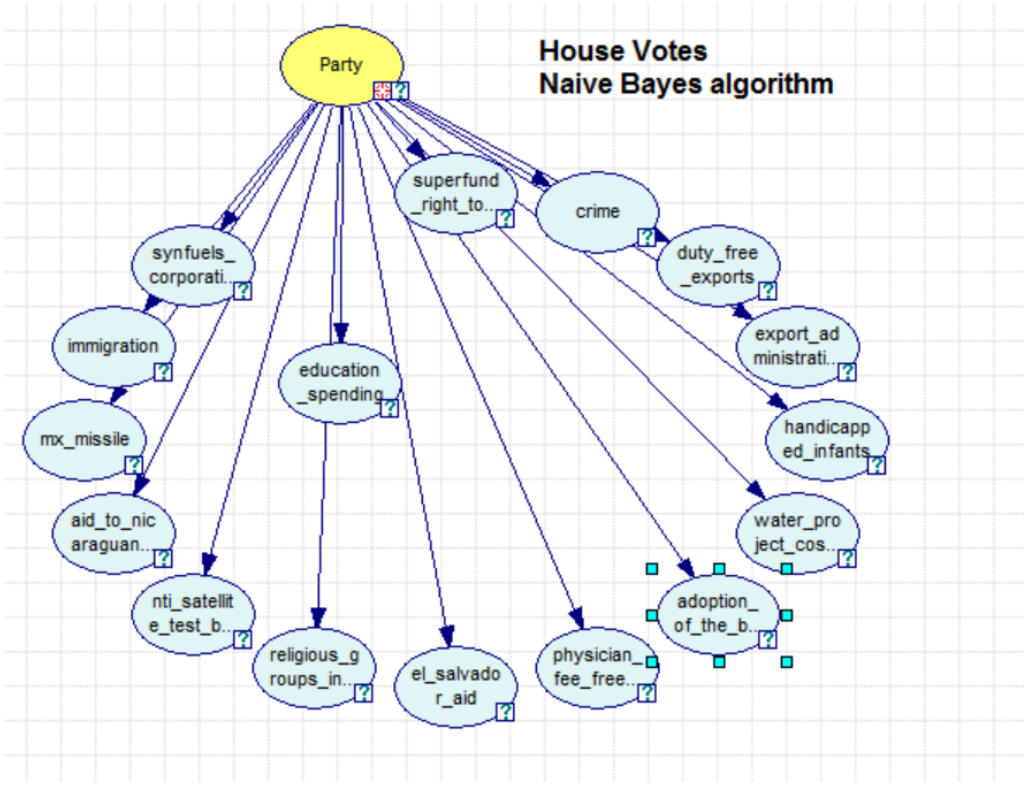

Figure 2: Tree View for Naïve Bayes model

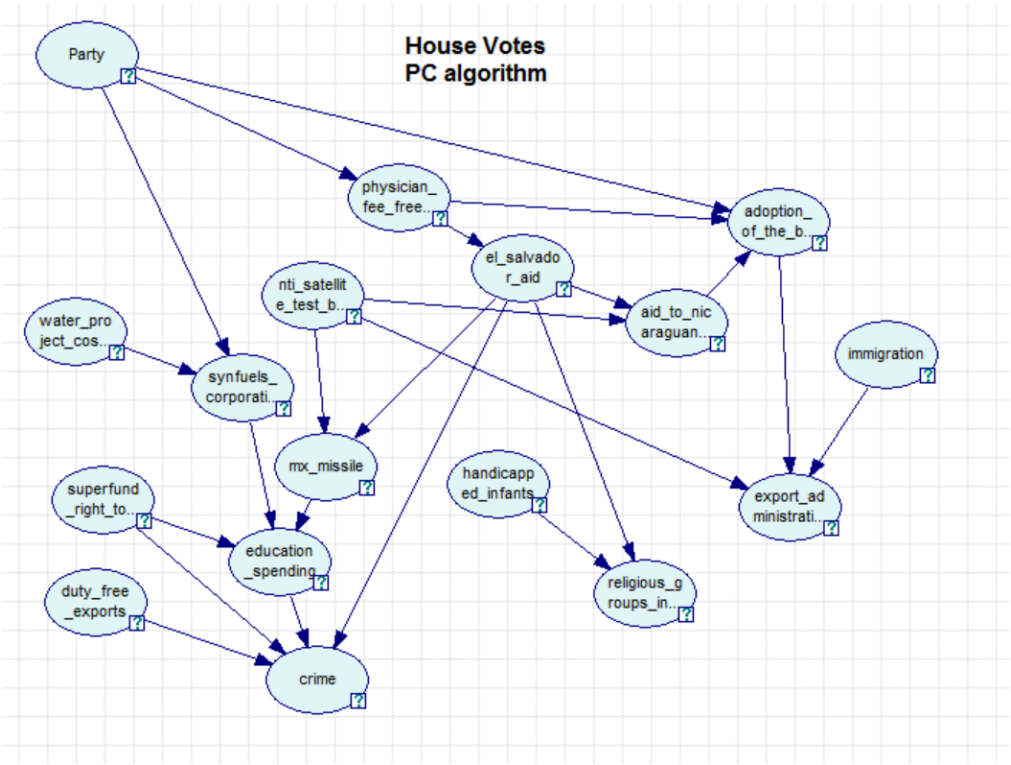

Figure 3: Tree View for PC model

1. Classification Accuracy

Tables below illustrate the accuracy for three models after running “leave-one-out” cross-validation through Genie. We can observe that Manually-constructed has the highest accuracy. "Democrat" is regarded as "positive", thus "republican" is accordingly recognized as "negative".

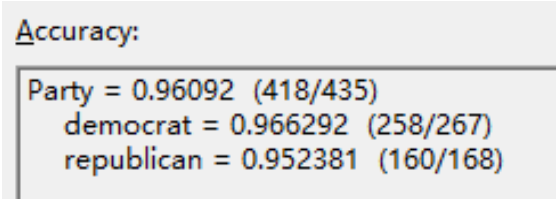

Table 1. Manually-constructed model

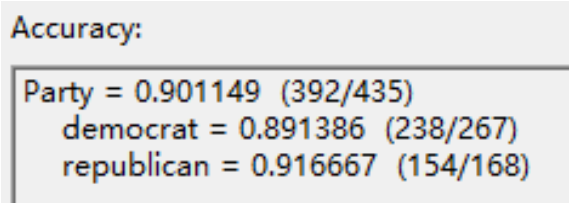

Table 2. Naïve Bayes

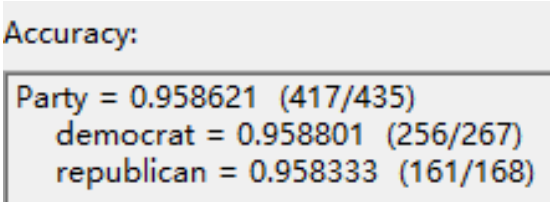

Table 3. PC

2. Sensitivity and specificity

Tables below are the confusion matrix for three models. All of them are counted when the cutoff of the prediction probability is 0.5.

	democrat	republican
democrat	258	9
republican	8	160

Table 4. Manually-constructed model

	democrat	republican
democrat	238	29
republican	14	154

Table 5. Naïve Bayes

	democrat	republican
democrat	256	11
republican	7	161

Table 6. PC

Table 7 illustrates the sensitivity and specificity for three models calculated according to the following equations referring to data on Table 4, 5, 6. As can be seen from table below, the sensitivity of manually-constructed model is highest while the specificity of PC model is highest. Both the sensitivity and specificity of the Naïve Bayes model are lowest.

Sensitivity: TPR = TP / (TP+FN)

Specificity: TNR = TN/ (TN+FP)

	Manually-constructed model	Naïve Bayes model	PC model
Sensitivity	97%	89%	96%
Specificity	95%	92%	96%

Table 7. Sensitivity and Specificity for three models

3. Positive and negative predictive value

Table 8 shows the positive and negative predictive value for three models calculated according to the following equations referring to data on the Table 4, 5, 6. From this table, we can see that the Manually-constructed model is the best one and the Naïve Bayes model is the worst one.

Positive predicative value: TP / (TP+FP)

Negative predicative value: TN/ (TN+FN)

	Manually-constructed model	Naïve Bayes model	PC model
Positive predicative value	97%	94%	97%
Negative predictive value	95%	84%	94%

Table 8. Positive and negative predictive value for three models

4. Calibration curve for a selected bin count

Figures below (Figure 4, 5, 6) are the calibration curve with bin count equals to 10 for three models. It is noticeable that the calibration curve of Manually-Constructed model and PC model are alike. We can observe from both two curves that the the real prevalence of democrat surges when classifier probability increases to more than 0.6. Moreover, the actual probability remains relatively high when the predicted ability is also high, suggesting the great predictive ability of Manually-Constructed model and PC models. As for the Naïve Bayes model, the real prevalence of democrat remains at around 0.65 even though the classifier probability increases from around 0.15 to 1, indicating the poor performance of this model.

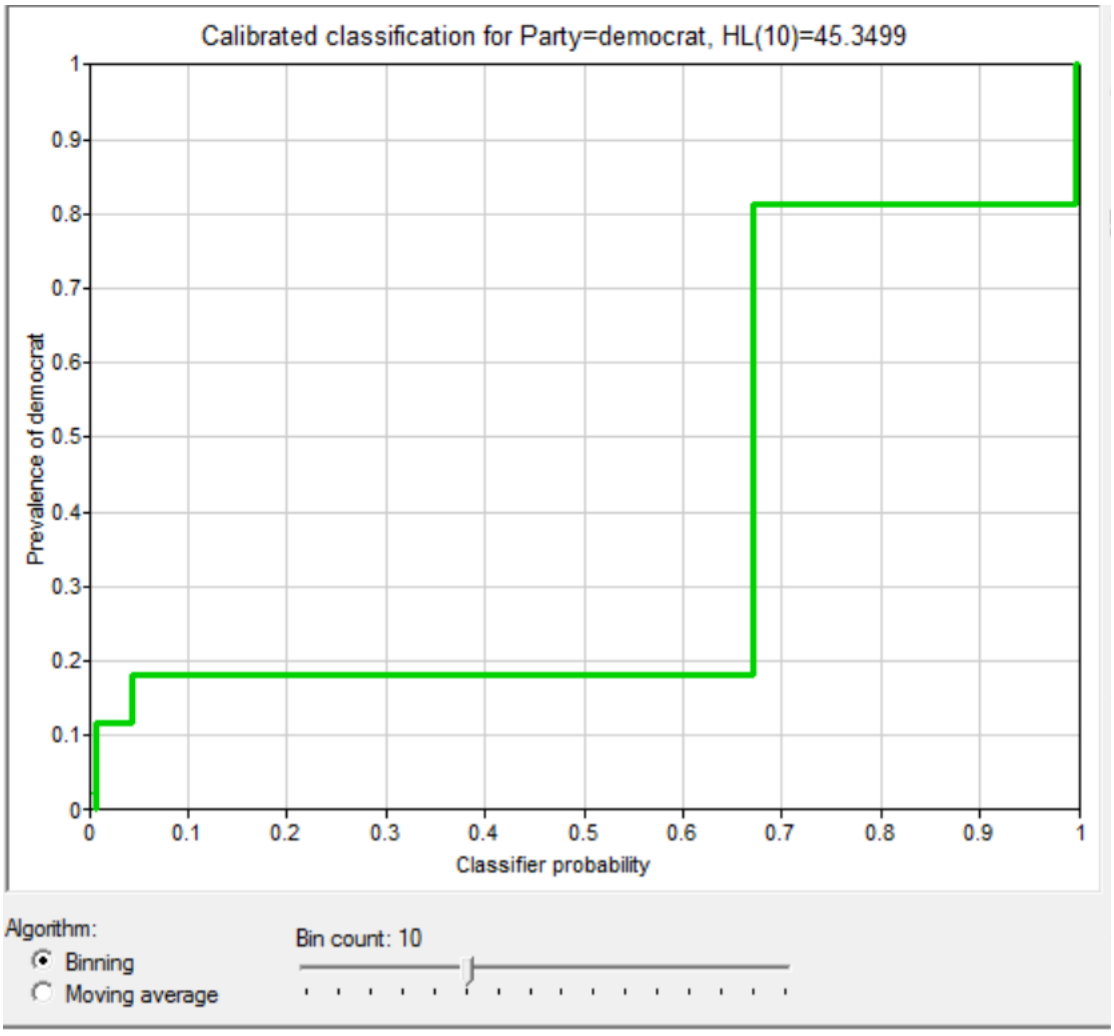

Figure 4. Calibration Curve for Manually-constructed model

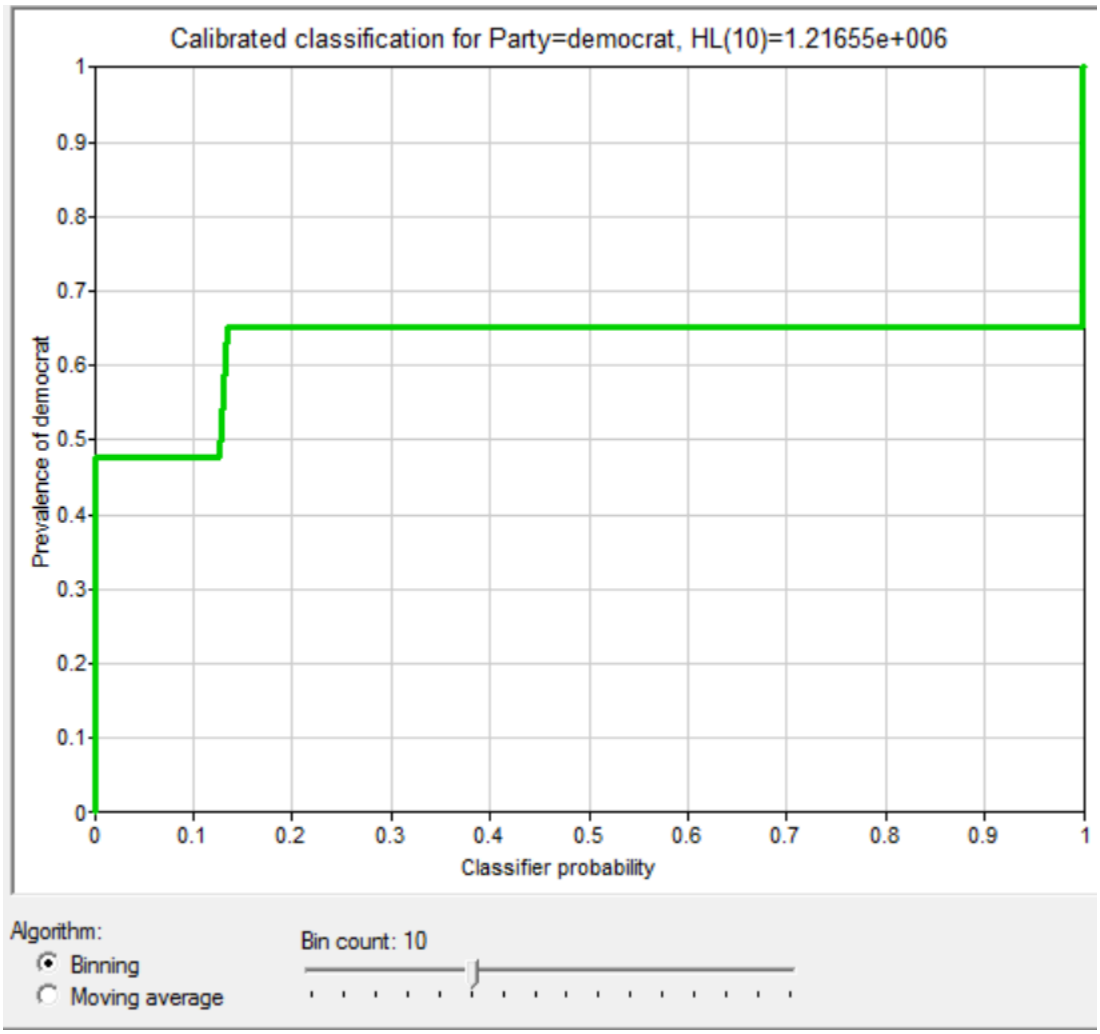

Figure 5. Calibration Curve for Naïve Bayes model

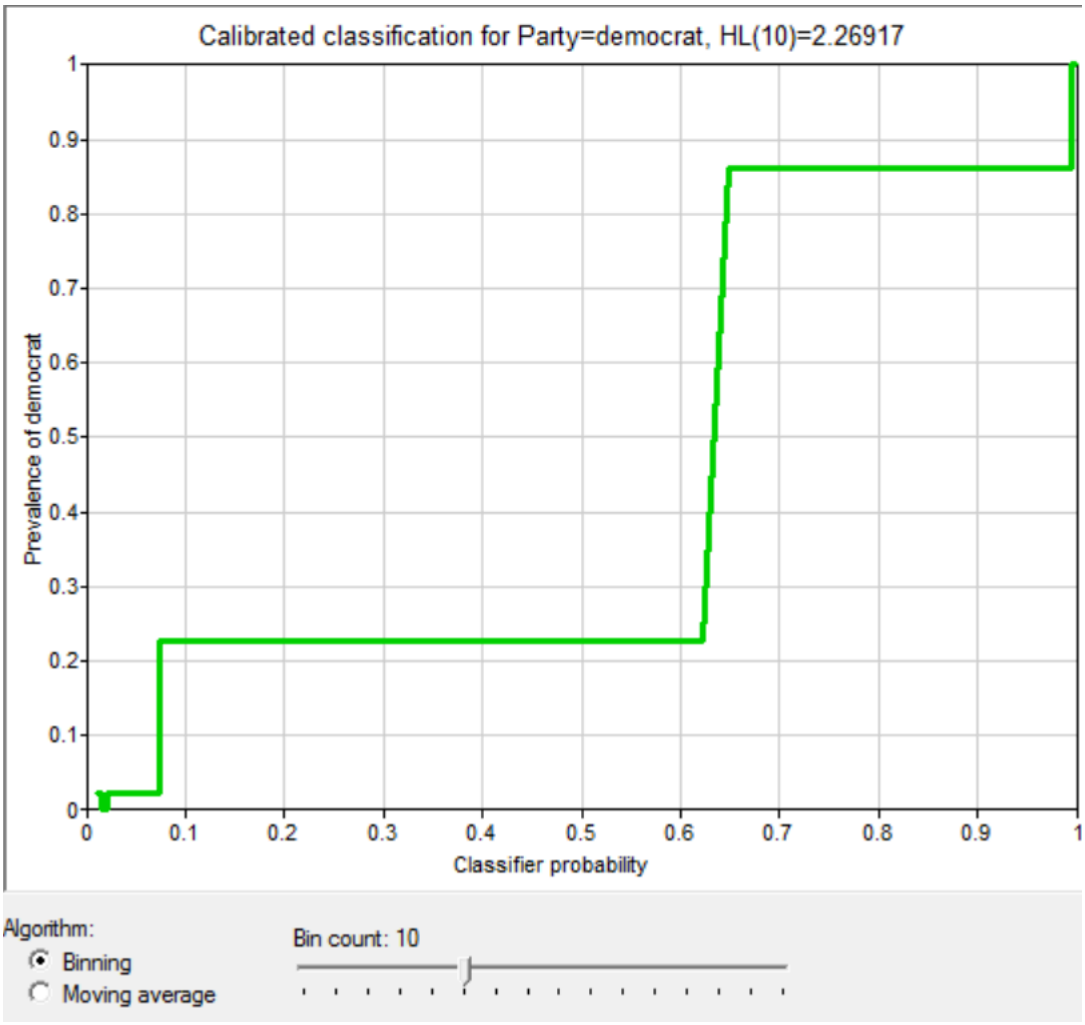

Figure 6. Calibration Curve for PC model

5. Conclusion

With respect to the accuracy, sensitivity and specificity, manually constructed model has overall better performance than other two models. This can be attributed to fact that the three predictors(adoption_of_the_budge_resolution,physician_fee_free,synfuels_corporation_cutback) this model uses have strong correlation with party affiliation of the representative, enhancing the predicative ability of this model. PC model is only second to manually constructed model. The Naïve Bayes has the worst performance concerning with all evaluation metrics, which could possibility resulted from the high correlation among predictors as Naïve Bayes assumes the independence of predictors.