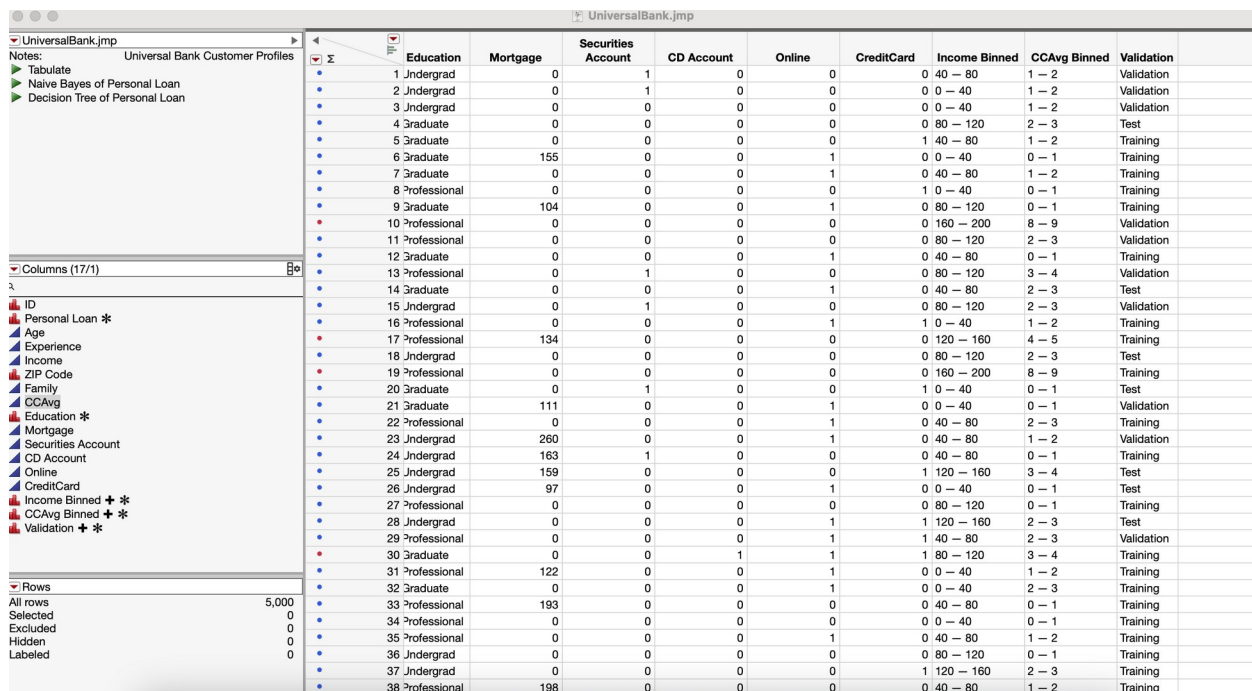


4KG3 Data Mining Assignment 2 - Classification

Student Name: Jieyu Zhang

Student Number: 400368960

1. Data preparation Bin continuous variable for Naïve Bayes. Make a validation column.



UniversalBank.jmp

Notes: Universal Bank Customer Profiles

- Tabulate
- Naive Bayes of Personal Loan
- Decision Tree of Personal Loan

Columns (17/1)

- ID
- Personal Loan *
- Age
- Experience
- Income
- ZIP Code
- Family
- CCAvg
- Education *
- Mortgage
- Securities Account
- CD Account
- Online
- CreditCard
- Income Binned + *
- CCAvg Binned + *
- Validation + *

	Education	Mortgage	Securities Account	CD Account	Online	CreditCard	Income Binned	CCAvg Binned	Validation
1	Jndergrad	0	1	0	0	0	40 - 80	1 - 2	Validation
2	Jndergrad	0	1	0	0	0	0 - 40	1 - 2	Validation
3	Jndergrad	0	0	0	0	0	0 - 40	1 - 2	Validation
4	3graduate	0	0	0	0	0	80 - 120	2 - 3	Test
5	3graduate	0	0	0	0	1	40 - 80	1 - 2	Training
6	3graduate	155	0	0	1	0	0 - 40	0 - 1	Training
7	3graduate	0	0	0	1	0	40 - 80	1 - 2	Training
8	Professional	0	0	0	0	1	0 - 40	0 - 1	Training
9	3graduate	104	0	0	1	0	80 - 120	0 - 1	Training
10	Professional	0	0	0	0	0	160 - 200	8 - 9	Validation
11	Professional	0	0	0	0	0	80 - 120	2 - 3	Validation
12	3graduate	0	0	0	1	0	40 - 80	0 - 1	Training
13	Professional	0	1	0	0	0	80 - 120	3 - 4	Validation
14	3graduate	0	0	0	1	0	40 - 80	2 - 3	Test
15	Jndergrad	0	1	0	0	0	80 - 120	2 - 3	Validation
16	Professional	0	0	0	1	1	0 - 40	1 - 2	Training
17	Professional	134	0	0	0	0	120 - 160	4 - 5	Training
18	Jndergrad	0	0	0	0	0	80 - 120	2 - 3	Test
19	Professional	0	0	0	0	0	160 - 200	8 - 9	Training
20	3graduate	0	1	0	0	1	0 - 40	0 - 1	Test
21	3graduate	111	0	0	1	0	0 - 40	0 - 1	Validation
22	Professional	0	0	0	1	0	40 - 80	2 - 3	Training
23	Jndergrad	260	0	0	1	0	40 - 80	1 - 2	Validation
24	Jndergrad	163	1	0	0	0	40 - 80	0 - 1	Training
25	Jndergrad	159	0	0	0	1	120 - 160	3 - 4	Test
26	Jndergrad	97	0	0	1	0	0 - 40	0 - 1	Test
27	Professional	0	0	0	0	0	80 - 120	0 - 1	Training
28	Jndergrad	0	0	0	1	1	120 - 160	2 - 3	Test
29	Professional	0	0	0	1	1	40 - 80	2 - 3	Validation
30	3graduate	0	0	1	1	1	80 - 120	3 - 4	Training
31	Professional	122	0	0	1	0	0 - 40	1 - 2	Training
32	3graduate	0	0	0	1	0	0 - 40	2 - 3	Training
33	Professional	193	0	0	0	0	40 - 80	0 - 1	Training
34	Professional	0	0	0	0	0	0 - 40	0 - 1	Training
35	Professional	0	0	0	1	0	40 - 80	1 - 2	Training
36	Jndergrad	0	0	0	0	0	80 - 120	0 - 1	Training
37	Jndergrad	0	0	0	0	1	120 - 160	2 - 3	Training
38	Professional	198	0	0	0	0	40 - 80	1 - 2	Training

Rows

All rows 5,000

Selected 0

Excluded 0

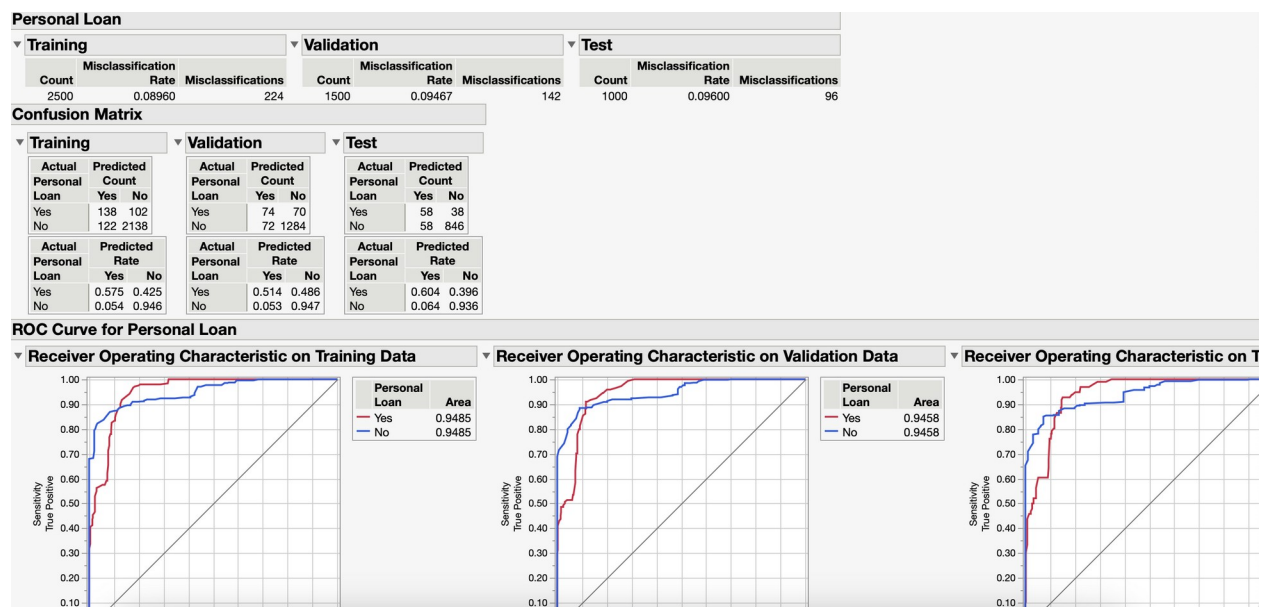
Hidden 0

Labeled 0

2. Running Naïve Bayes for classification

Check the misclassification rate and the confusion matrix. You may also look at the lift curve. Are you happy with these results? If not, why? Can you further improve it? How?

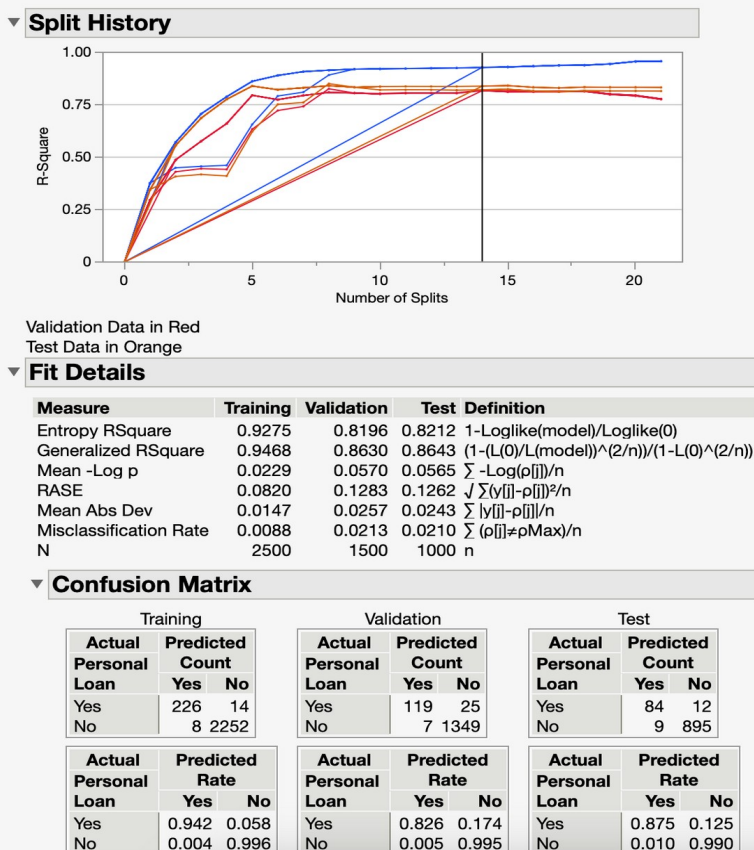
The overall performance of the model is good, with a low misclassification rate (training: 8.96%, validation: 9.47%, testing: 9.6%) and a high AUC score (~0.95), indicating strong discriminative ability. However, although the specificity (correctly identifying "no" cases) is high (~94-95%), the sensitivity (correctly identifying "yes" cases) is slightly low (~51-60%), which means that many actual loan approvals are misclassified. To enhance, the model can use random forest or gradient augmentation, these two models could better capture complex relationships between variables. In addition, cost sensitive learning can be used to prioritize appropriate classification for loan approval. If the goal is to reduce false positives, then this model is sufficient; However, if increasing the recall rate of 'yes' cases is crucial, additional optimization is needed.



3. Building classification tree and evaluating performance

You may try split and prune tree and try find a best or desirable tree. What is your criteria and approach to identify the best tree? What method you have used to evaluate the classification performance? Show your evaluation results.

The best tree is selected based on minimizing misclassification rate (training: 0.88%, validation: 2.13%, testing: 2.1%), maximizing predictive ability, and balancing model complexity. The high entropy R-squared (test data of 0.8212) and generalized R-squared (0.8643) indicate strong explanatory power, while the segmentation history graph confirms that additional segmentation does not significantly improve performance, thus preventing overfitting. The confusion matrix shows excellent classification performance, with a sensitivity of 87.5% for predicting loan approvals and a specificity of 99% for predicting non loan approvals. The model was evaluated using confusion matrix, misclassification rate, R-squared, RASE, and mean absolute deviation, all of which confirmed that the decision tree is the best model, achieving the best balance between accuracy and generalization.



4. Make evaluation and comparisons of the two methods

Compare Naïve Bayes and Decision Tree for classification. What are their advantages and disadvantages? How is their performance? Are you happy with the results? Which one you prefer? Why?

Based on the analysis of Decision Trees and Naive Bayes, Decision Trees are significantly better than Naive Bayes in terms of classification accuracy and misclassification rate. The misclassification rate of the decision tree model on the test set is only 2.1%, with high sensitivity (87.5%) and specificity (99%), while maintaining strong explanatory power (entropy R-squared: 0.8212, generalized R-squared: 0.8343). In contrast, the Naive Bayes model has a higher misclassification rate and encounters difficulties in feature dependency because it assumes that variables are independent, which may not hold true in this dataset. Although Naive Bayes is computationally efficient and suitable for high-dimensional data, Decision Trees provide better interpretability, capture complex feature relationships, and generalize well without overfitting. Given these results, I am more satisfied with the Decision Tree model as it provides higher accuracy, better classification performance, and lower error rates. Therefore, I prefer Decision Trees to complete this classification task because they provide a more reliable and interpretable solution.

5. Possible real word applications

What is the potential application of classification for mortgage approval? What data a bank may need to collect and what analytic tool the bank should consider in order to make a good mortgage approval decision? Search web to find out the criteria for mortgage application and approval.

Classification models are critical for automated mortgage loan approval decisions, allowing banks to accurately assess applicants' creditworthiness while lowering risks and manual work. Banks can use models like decision trees, random forests, and logistic regression to estimate loan default risks and make data-driven lending choices. To develop an effective model, banks must collect essential financial and demographic data, such as credit scores, income, job history, debt-to-income ratio, loan-to-value ratio, payback history, and existing liabilities. JMP Pro, SAS, IBM SPSS, AI-powered credit risk models, and other analytical tools can enhance forecast and approval processes. These classification-based systems improve loan accuracy, reduce defaults, and increase profitability by ensuring that applicants with good reputations are granted loans while high-risk candidates are flagged for further evaluation.

6. Report how much time you have spent on this assignment, what problem you have faced and what you have learned.

I worked on this homework for almost five hours, mostly concentrating on deciphering the implications of the different indicators and analytical results. Explaining various assessment measures and their effects on model performance, such as entropy R-squared, generalized R-squared, misclassification rate, and confusion matrix, is my biggest issue. Additionally, more research is required to differentiate between the benefits and drawbacks of decision trees and naive Bayes in classification tasks. I gained knowledge on how to analyze classification performance data, assess the precision and dependability of models, and comprehend how these insights may be used to make judgments in the real world, like approving a mortgage. My ability to efficiently evaluate and interpret machine learning outcomes has improved as a result of this work.