Assessing Loan Risks: A Data Mining Case Study

R. Gerritsen, IT Prof. Nov./Dec. pp. 16-21, 1999
1. Problem Domain

- Six hundred thousands loans administered by USDA
- Need to better understand the loans
  - Improve management of lending program
  - Reduce incidence of problem loans
- Requirement – to sort out borrowers who does not pay promptly
- Likely to be heading for trouble
2 Data Preparation

- A 2% sample with 12000 records extracted from data warehouse
- Information available
  - Loan (amount, payment size, lending date, purpose, etc)
  - Asset (dwelling type, property type, etc)
  - Borrower (age, race, marital stat, income)
  - Region loan made (state, minorities)
3 Accuracy of Data Mining Algorithm

- Measurement of the success rate of the prediction
- No “most accurate” algorithm exists
- Different algorithm achieves better or worse accuracy in different cases
- The best algorithm for a data set cannot be determined in advance -> need to use at least two algorithm and apply evaluation techniques
4 Interpretability of Algorithm

- k-NN does not produce a model -> the worst in interpretability
- Neural network model produces little information to use directly
- Naïve Bayes model tells you which variables (attribute values) are most important for a particular outcome
- Decision trees can highlight interactions on attributes
5 Speed of Algorithm

- Two components
  - Speed to build the model (training time)
  - Speed to use the model (making predictions on new cases)
- k-NN: zero training time but very slow prediction
- Other 3 algorithms have similar prediction speed but different in training speed:
  - Naïve Bayes the fastest (one pass)
  - Decision trees require 20-40 passes
  - Neural networks may take 100 to >1000 passes
6. Building the Models

- Build a Naïve Bayes model for initial exploration
- Follow up with a decision tree model
- Output class value: problemless, substandard, loss, unclassified and not available
- Decide Holdout procedure: two third (8000) of the input dataset for training, the remaining for testing.
6.1 First Model – Initial Result

- Naïve Bayes is used because it is fast to build and provides good interpretability
- A poor success prediction rate of around 50% on the testing dataset
6.1 First Model – Refinement

- Problem – default binning of the payment amount into equal interval not appropriate for skew distribution
- Refinement: rebin so that each bin contains equal population
- Result improved to 67% (76% achieved for problemless and loss categories)
6.1 First Model – A Problem Highlighted

- Problem: an important attribute is “total loan amount due”: large value -> problem loans
- However, large value often means borrower stop payment already which is a after the fact information -> not useful to sort out potential bad loans
- Remove the attribute result in the success rate dropped to 46% (37% for the loss category!)
6.1 First Model – Further Refinement

• No need to predict the unclassified and not available classes which <1% -> discard instances
• No need to distinguish between substandard and loss classes -> combine into a superclass “Not OK” and rename the problem less to OK.
• Revised model achieves 82% accuracy!
• However, only 20% success to predict not OK loans -> needs further improvement
6.2 Followup with A Decision Tree

- A success rate of 85%
- Accuracy on “Not OK” class is 23%
- Some improvement over the Naïve Bayes model
- More analysis -> counting the cost
  - Savings in early prevention of problem loans
  - Cost of intervening with good accounts