In *Computational Intelligence in Financial Engineering*, pages 232–236, Piscataway, NJ, March 1997. IEEE Press.

Mortgage Data Mining

George H. John
Global Business Intelligence Solutions
IBM Almaden Research Center
gjohn@almaden.ibm.com
http://www-cs-students.stanford.edu/~gjohn/

Ying Zhao
Global Business Intelligence Solutions
IBM Almaden Research Center
650 Harry Rd, San Jose, CA 95120
zhao@almaden.ibm.com

Abstract

This paper reports a preliminary investigation of the use of modern data mining tools for mortgage scoring. Using IBM's Intelligent Miner (a data mining toolbox), we built a model of serious delinquency on a sample of data from Mortgage Information Corporation's Loan Performance System, which contains over 20 million loans with a volume of over $1.6 trillion. Currently, two technologies prevail in mortgage scoring: logistic regression, a very old and very simple method, and neural networks, newer and more complex types of models that can be extremely difficult to interpret. The radial basis function (RBF) algorithm in Intelligent Miner combines the mathematical complexity and generality of neural networks with a comprehensible visualization that explains the RBF model. Due to the performance and understandability of the RBF model, as well as other unique technologies not described here, the Intelligent Miner should be a useful tool for mortgage bankers, facilitating development of customized systems for mortgage scoring and other mortgage banking applications.

1 Introduction

The mortgage industry is currently adopting many technologies—to lower servicing costs, improve origination risk exposures, and to price mortgage-backed securities more efficiently. Data mining is a technology that many originators and servicers are interested in using, and are already using in limited forms. Working with the Mortgage Information Corporation's Loan Performance System database and IBM's Intelligent Miner tool, we analyze the risk in the market portfolio represented by MIC's LPS database, and we explain the model of delinquency we discovered. The model was able to successfully identify seriously delinquent mortgages from a database of information that would typically be available at the time the underwriting/origination decisions are made, thus the model should be a useful tool for mortgage scoring during this stage of the process. In addition to predicting future delinquency, the RBF model also segments mortgages into groups with similar risk. During the process of trading a pool of mortgages, a seller might be able to obtain a higher price by publishing the risk group composition of the pool, allowing buyers to better estimate the effect that adding the pool would have on their overall risk exposure.

Many technologies for managing stock portfolios are used by financial engineers in the mortgage industry as well, for example, in option-theoretic pricing of mortgage-backed securities (Fabozzi & Modigliana 1992, Cheyette 1996). Recently, several authors have explored the use of data mining technologies in stock selection (John, Miller & Kerber 1996, John & Miller 1996, Apte & Hong 1995, Levin 1996), and now the mortgage industry is turning to this technology as well: 27% of large mortgage servicers identified data mining as the most important technology for the future (Lebowitz & Bechhoefer 1996). Similar technology exists in the mortgage industry in a fragmented state right now, for example, as mortgage scoring systems (DeZube 1996), but these systems are largely black boxes, and are based on decades-old technologies such as linear and logistic regression (von Furstenberg 1971, Pinder 1996). A user of the system knows neither how the system was built, nor how the system is making its decisions, nor whether the system is really appropriate for her own problems (perhaps the scoring system was built using a training database with a different bias than the data she intends to score). Additionally, some mortgage scoring models are built using a very small sample of data, thus the model might not be as sophisticated as one built using large amounts of data.

LTV	Product	Amount	Purpose	Occupancy	···	Seriously DLQ
80%	Fixed 30yr	$150,000	Purchase	Owner		No
85%	Fixed 15yr	$350,000	Refinance	Non-owner		Yes
60%	30yr ARM (Tbill Index)	$200,000	Purchase	Owner		No
⋮						

Table 1: Hypothetical fragment from the historical database.

In contrast, data mining methods strive to build models of data that give insight to a human business analyst. A user of a data mining system knows exactly how the model was built, since the user herself specifies which training database to use to build a model. Finally, data mining tools generally are built using advanced computational technologies (efficient data structures, fast algorithms, and parallel processing) to handle extremely large databases, and modern statistics to squeeze the data dry without discovering spurious patterns.

In the following, we describe an "open" process of building a scoring system. An analyst at a mortgage bank could independently build such a model using the tools and data described, or by using their bank's own data and internal algorithms. The experiment we describe is one example of how that process would work.

2 Mortgage Information Corporation's Database

The patterns discovered by data mining are only as good as the tools that produced it and the training database being mined. MIC has developed the nation's largest, most detailed database on the credit and prepayment performance of whole loans, mortgage-backed securities, and mortgage markets throughout the U.S. For our delinquency scoring model, we used MIC's Loan Performance System, which contains more than 20 million residential first mortgages. With a loan volume of $1.6 trillion, this data base captures approximately one half of the total dollar volume of mortgage loans outstanding in the United States. The data represents loans held or serviced by more than 20 of the largest financial concerns in the country, as well as loans purchased by Fannie Mae and Freddie Mac. The LPS database is updated frequently as new performance data becomes available. By using such a vast database, a data miner is sure to find a significant number of mortgages matching virtually any combination of features. For example, even if one is only interested in 1995 origination year Ohio loans with over 90% LTV, there are still thousands of loans in the database matching this characteristic.

The database we used to build our scoring model can be thought of as a single table with one row per loan and one column per attribute such as loan-to-value ratio and loan product (fixed 30 year, 15 year ARM, *etc.*). Table 1 shows a hypothetical fragment of the database. Our goal was to predict the delinquency status of a mortgage, which is also available in the LPS database. We chose to investigate whether a loan was *seriously* delinquent, which was defined as 90 days past due or foreclosure.

3 IBM Radial Basis Function Algorithm

One kind of technology for data mining is radial basis function fitting (Ripley 1996, Ch. 4). Radial basis function fitting addresses the supervised learning problem: given a database, build a model which allows the prediction of one target attribute given the rest. In the particular mortgage delinquency risk analysis we have performed, the field of interest is whether or not the mortgage subsequently becomes seriously delinquent, and the database would include such other fields as Loan-To-Value ratio (LTV), origination amount, product type (30 year fixed, 15 year variable, etc), and so forth. Radial basis function fitting produces a complex model that can be hard to understand. Fortunately, the radial basis function fitting method implemented in IBM's Intelligent Miner product includes visualization routines that truly allow a business analyst to understand the resulting model.

Refer ahead to Figure 1, which shows the visualization of the RBF model built on the database described above. Each horizontal band describes one discovered segment of mortgages. In standard segmentation, or clustering, segments are discovered so that all items in a segment are similar, and items in different segments are dissimilar. In the RBF model, this is also true, but the algorithm for identifying segments is primarily concerned with predict-

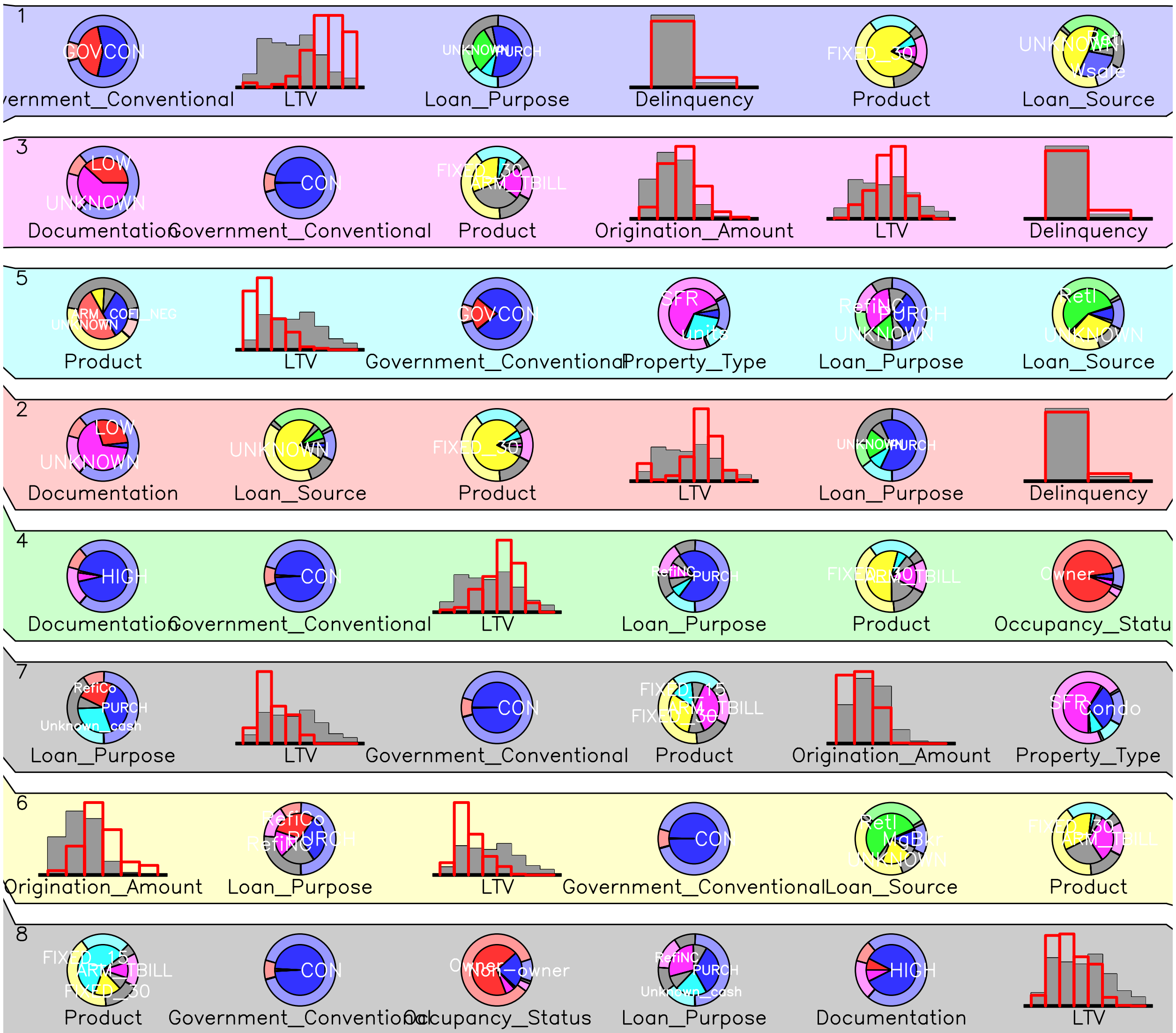

Figure 1: Visualization of radial basis functions.

ing the target attribute. The standard visualization also shows the percentage of mortgages that were assigned to each segment, and the average value of the target attribute within each segment, but we have removed that part of the picture so that the remainder is clearer.

The visualization of each segment shows the variables that most distinguish the segment from the rest of the data. For example, in the top segment, the leftmost variable is "Government or Conventional." Whereas in the entire database there are only about 10% government loans, in this segment there are nearly 45%. One can read this from the "pie within pie" chart above the "Government or Conventional" label—the outer pie shows the distribution of the entire database, showing a small wedge (about 10%) under the label "GOV" and a large wedge (about 90%) under the word "CON". The inner pie describes the distribution within the segment, where the wedge under the GOV label is almost half (say 45%), and the wedge under the CON label is about 55% of the circle. Looking at the next segment, we can read that the most distinguishing variable is the documentation, and that the second segment from the top has about 40% low documentation loans and 60% unknown documentation. (The labels on the outer circle of the pie are hard to read. Not all information can be neatly displayed on a single page—when running the Intelligent Miner, the user can zoom in on parts and see the actual numbers behind each graph.)

Categorical or nominal attributes are characterized by pie-within-pie charts as described above, but we have not discussed numeric variables yet. Those are shown using double histogram graphs. For example, in the first segment the second most distinguishing variable was the loan to value ratio. The solid grey part of the graph above "LTV" shows that the overall database contains loans roughly evenly spread over the range of loan to value ratios (except for the very high and low LTV loans, where the numbers drop off). In contrast, the loans in this segment are almost all very high LTV loans, as indicated by the hollow black-bar histogram. As another example, looking at the third segment from the top we can see that it contains mostly low loan to value loans.

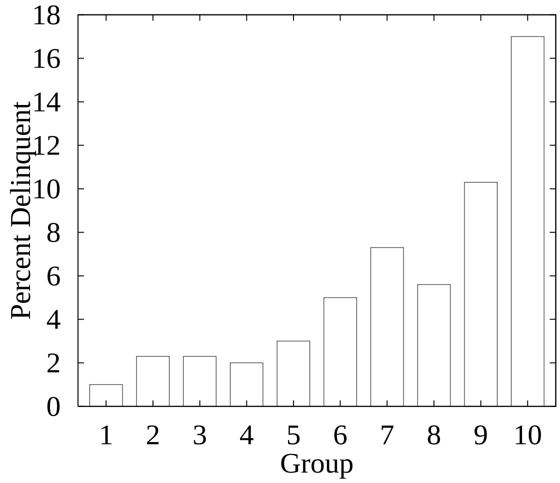

Figure 2: Results on an independent test set.

4 Results

Figure 1 shows the visualization of a model we built from a small sample of the MIC LPS database. From the model shown in Figure 1, we understand that the most risky group of mortgages (the top band) are what we would expect: high LTV, government loans. The second most risky group might be a surprise: LTV around 75 and low or unknown documentation—this is the "Low doc 75" phenomenon: in return for 75% or lower LTV, lending institutions would approve mortgages with low documentation. The RBF algorithm automatically identified such loans as a segment, based only on the database from which it was built. The model suggests these loans are more risky, and also that the documentation field might be "unknown" for strategic reasons.

The bottom band contains the safe mortgages, low LTV conventional loans with 15 year terms, indicating a financially stable debtor. The next-lowest band contains extremely low LTV loans with high origination amounts and many ARM refinances.

To evaluate the radial basis function model as a tool for aiding underwriting decisions, we used it to score an independent test database of mortgages not used during the initial training. We then sorted the scored mortgages into ten groups, where each group contained 10% of all of the test mortgages, and the first group contained the 10% with the lowest scores, the second group contained the 10% with the next lowest scores, and so on. We observed a high correlation between the score and the actual delinquency of a mortgage. In the data we obtained from MIC, about 5% of the mortgages were seriously delinquent after the database was processed. Figure 2 shows that in the lowest decile of scores, only 1% of mortgages were actually delinquent, while in the highest decile of scores, over 17% were seriously delinquent. This compares well with Pierzchalski (1996), which shows that when using FICO scores on Mortgage Guaranty Insurance Co.'s database, the highest scoring quintile experienced eight times the foreclosure rate of the lowest scoring quintile, and using MGIC's mortgage score the ratio was 34 foreclosures in the high risk quintile to 1 in the low risk quintile. (It should be noted that the test database in this case was the same database on which the MGIC score was developed.) Our own ratio was 8.27 to 1.

5 Conclusion

Any modeling tool used for mortgage portfolio management should produce a comprehensible model. With millions of real dollars at stake, the ability to understand the model is a necessity. Neural networks are very popular, but their opacity makes them a less than ideal tool (Barr & Bhagat 1994). We believe that the ability of IBM's radial basis function technology to discover patterns and present them as easily understood graphs makes it an excellent tool for use in many areas of mortgage banking. Data mining can be used in nearly every phase of mortgage banking: developing targeted marketing campaigns for solicitation, customer retention models, mortgage and credit scoring for assisted underwriting and mortgage insurance pricing, credit collection response models for use during servicing, and prepayment and pricing models for use in trading of mortgages and mortgage-backed securities.

Of course, to be a useful tool, the learned model must be a good predictor of future performance. Our experiments show that the radial basis model was exceptionally good at distinguishing good loans from loans that would later become seriously delinquent.

References

Apte, C. & Hong, S. (1995), Predicting equity returns from securities data with minimal rule generation, in U. Fayyad, G. Piatetsky-Shapiro & P. Smyth, eds, *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, chapter 22, pp. 537–556.

Barr, D. S. & Bhagat, S. A. (1994), "Computers as investment strategists", *Pension Management*, November, pp. 12–15.

Cheyette, O. (1996), "Implied prepayments", *Journal of Portfolio Management*, Fall, pp. 107–114.

DeZube, D. (1996), "Mortgage scoring rules of thumb", *Mortgage Banking* **56**(11), August, pp. 50–57.

Fabozzi, F. J. & Modigliana, F. (1992), *Mortgage and Mortgage-Backed Securities Markets*, Harvard Business School Press, Boston, Massachusetts.

John, G. H. & Miller, P. (1996), Building long/short portfolios using rule induction, in *Computational Intelligence in Financial Engineering*, IEEE/IAFE, New York, New York.

John, G. H., Miller, P. & Kerber, R. (1996), "Stock selection using rule induction", *IEEE Expert* **11**(5), October, pp. 52–58.

Lebowitz, J. A. & Bechhoefer, I. S. (1996), *MORTECH 96: The Biennial Analysis of Technology Use in the Mortgage Industry*, SSP/RES Research.

Levin, A. U. (1996), Stock selection via nonlinear multifactor models, in *Advances in Neural Information Processing Systems*, Morgan Kaufmann Publishers, San Francisco, California.

Pierzchalski, L. (1996), "Guarding against risk", *Mortgage Banking*, June, pp. 38–45.

Pinder, J. P. (1996), "Decision analysis using multinomial logit models: Mortgage portfolio valuation", *Journal of Economics and Business* **48**(1), pp. 67–77.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, New York, NY.

von Furstenberg, G. M. (1971), *Technical Studies of Mortgage Default Risk*, Center for Urban Development Research, Cornell University, Ithaca, NY.