Using Data Mining Technology to Identify and Prioritize Emerging Opportunities for Mortgage Lending to Homeownership-Deficient Communities

Clark R. Abrahams, Community Development Banking, Bank of America
Frank R. Burnett, Community Development Banking, Bank of America
Jin-Whan Jung, Ph.D., Analytical Consulting, SAS Institute Inc

ABSTRACT

Advances in technology are fueling substantial consolidation in both mortgage origination and servicing. Internet lending is growing, and on-line companies and distribution channels will proliferate. Demographic and economic trends are driving changes in customer mix. Amid all of this change, there are some constants. One such constant is the recognition that in order for communities to realize their full potential, it is advantageous for as many residents as possible to become equity stakeholders. Homeownership affords the owner(s) the promise of capital appreciation, along with a sense of pride and fulfillment. There is an ever-growing need to bring more low/moderate income & minority households into the home-buying process.

To compete successfully, lenders will need to utilize data mining techniques to develop powerful models that can identify markets that are both under-tapped and also present profitable opportunities. This paper illustrates how to gather/ pre-process the requisite data, and develop / assess alternative models. Findings, together with a discussion of practical considerations, for an illustrative data-mining example are presented.

THE BANKING LANDSCAPE IN THE NEXT DECADE

OVERVIEW [1]

The banking sector has undergone rapid change during the past decade. Factors such as rapid technological advances, a booming economy, a broadening of products and services offered, and consolidation have had the greatest impact. As the industry has evolved, the core competencies required to successfully compete have undergone dramatic change.

Traditionally, banks have depended upon their strengths in: 1) personal sales, 2) branch location, construction, and operation, 3) risk management and financing skills, and 4) expense control. Today, banks depend upon strengths in: 1) technology management, 2) information management, 3) innovation, budgeted R&D, 4) service quality, 5) bundling / pricing of originated assets for sale to investors, 6) alliance management, and 7) brand management.

Traditional assets and services included: 1) on-balance sheet portfolios of loans & marketable securities, 2) DDA, card services, funds storage, payment authorization, clearing and settlement, payment accounting / reporting, and 3) specialized services, e.g. trust, safekeeping, credit enhancement. Today, competition is emerging around information technology-created assets, as the barriers of time, distance and form have dissolved. Bank's assets and services have been expanded to emphasize: 1) financial transaction speed / flexibility, 2) real-time reaction / response, 3) customer access / reach, 4) insight / knowledge, and 5) intelligence / creativity.

The players, and their respective roles, in this new financial sector landscape include: 1) owners of information and intelligence, i.e. content providers, 2) hardware manufacturers, who provide technology 3) telecommunications companies and transporters, who supply channel access and customer reach, 4) banks and e-commerce providers, who represent various utilities, focusing on: transaction, execution and service, and 5) 3rd party developers of business solutions, who are the context providers and packagers that offer efficiencies and new business capabilities.

HOUSING AND THE MORTGAGE LENDING ARENA

The United States is experiencing the longest housing boom on record. Last year marked the 50th anniversary of the 1949 Housing Act. Below is an excerpt from its' preamble:

"The Congress hereby declares that the general welfare and security of the Nation and the health and living standards of its people require housing production and related community development sufficient to remedy the serious housing shortage, the elimination of substandard and other inadequate housing … and the realization as soon as feasible of the goal of a decent home and suitable living environment for every American family … " [2]

Nicolas Retsinas, former head of HUD and current director of the Harvard Joint Center for Housing Studies (JCHS) has stated that "The Act, however, reminds us that rhetoric, or even vision, is not a substitute for programs and policies, and certainly not for resources." The primary burden of achieving the goal highlighted above rests squarely on the private sector. The table below illustrates some of the progress in housing that has been achieved since passage of the Act. [3]

	1949	1999
National Homeownership Rate	55.0%	66.6%
Total Number of Housing Units	42.8MM	119.0MM
Lacking Complete Plumbing Facilities	35.5%	1.1%
Crowded (> one person per room)	15.7%	4.9%
No telephone available	21.5%	6.6%
Total Single-Family Housing Starts	1.229MM	1.319MM
Total Multi-Family Housing Starts	201M	341M

Table A. Housing Trends

Specific to mortgage lending, we can point to several key trends[4]

- **Record-breaking homeownership rates are expected this decade, possibly crossing the 68% threshold!** "Over the next decade, the pace of household growth should match or slightly exceed the 1.1-1.2 million annual rate averaged in the 1990s. Including manufactured homes, the number of housing units added should thus be on a par with the 16 million or so built in [the past] decade." [5] At the end of the first quarter of this year, the government reported that 70.1 million families in the United States own their own home, which equates to a homeownership rate of 67.1 percent (72 percent of White households are homeowners, with minorities trailing by 26 percent).

- **Signs are that the gap between homeownership rates for Whites and rates for minorities will keep getting bigger**, according to Mr. Retsinas. Apparently, it has not narrowed during the past six years, despite the prosperous economy. Gaps are supposed to narrow in good times. The Harvard JCHS reports that African-Americans trail Whites by 26.6 percentage points, while U.S.-born Hispanics trail Whites by 23 percentage points. JCHS goes on to say that the homeownership rate for African-American and Hispanic-American college graduates with a BA degree is less than that for Whites having only a high school diploma. Home prices have been on the rise, and have hit record highs in over half of the nation's largest cities. Real mortgage costs are estimated to have risen 5 percent from 1998 to 1999. Down payments have increased $442 on average for the same period, up 3.5 percent. Naturally, the prevailing interest rate environment will have a significant impact on the attainment of the 68 percent homeownership milestone. "According to a study conducted by the Department of Housing and Urban Development (HUD), every percentage point increase in interest rates pushes 400,000 families out of the housing market because housing becomes unaffordable for them. … [Consequently, another trend is that] consumers are turning to adjustable rate mortgages (ARMs) to lessen the strain on budgets caused by the increase in mortgage rates. In February 2000, 32 percent of new mortgages were adjustable rate, versus 11 percent one year ago." [6]

  Using data mining techniques and census data, it is possible to locate communities having low-income renters, below poverty level households, etc. and assess opportunities for designing programs promoting homeownership (based on specially designed sets of underwriting factors). In addition, census data provides information combining ethnic and housing variables, so that one can compare the percentage of homeownership compared with percentage in the population for Native-Americans, Asian-Americans, African-Americans, Hispanic-Americans, Mixed-Race Americans, and Whites. This information can also be helpful from a fair lending perspective.

- **There will be a redistribution of business flows among intermediaries in the value chain, which will lead to "deconstruction," the dismantling and reformation of traditional mortgage lending industry business structures.** Andy Woodward**,** president-elect of the Mortgage Bankers Association of America (MBAA) pointed this out recently, and the fact that mortgage bankers must change their mind-set as well as their value system at an MBAA-sponsored Business Strategies Conference earlier this year. He identified two key drivers of change within the mortgage value chain, which consists of investors, guarantors, aggregators, and originators, all serving the customer. The first key driver is consolidation, both horizontal and vertical. Mr. Woodward observed that "Consolidation is a very real thing that is going on, and it's happening on other planes as well. You are starting to see appraisers, collateral insurers, credit repositories, [and others] forming a type of horizontal chain." [7]

- **Internet lending is growing, and on-line companies and distribution channels will proliferate.** Mr. Woodward said that the second key driver of change is technology, and he went on to say that "Business strategies have become dominated by technology. … What's changing is the 'virtualization' of the mortgage banking industry, and it's driving the value chain. … [The Internet] is a process enabler that allows us to reach our customers faster, cheaper, simpler and friendlier." [8]

- **Wholesale business will rise relative to traditional retail business, but the competition aspect between the two will change.** "Today, it's retail competition against your company's own Web site. It's wholesale competition against business-to-business exchanges. And it's the correspondent competition against eliminating the middlemen and aggregators." [9]

- **Customer mix will change due to demographic trends and a greater emphasis on bringing nontraditional borrowers (low income & minorities) into the home-buying process.** Concerning homebuyer demographics, the proportion of 1st time homebuyers will rise over the next decade. The echo boomer generation is 80 million strong, and lenders will push to reach out to minorities. In 1998, the greatest share of 1st-time homebuyers by race was Hispanics (82%), followed by African-American (71%), and Asian (68%). Growth of immigration to the United States has been dramatic, with 50% decade-to-decade growth for the 1980's to the 1990's. The number of immigrants 1981-1990 was 7.3 million, and for 1991-2000 the figure was 11 million. Back in 1994, Fannie Mae made a historic commitment to finance $1 trillion in loans to minorities, people living in inner cities, recent immigrants and people with incomes below the area median. Fannie Mae estimates that by the end of 2000, twenty-five percent of all households in America will be headed by minorities. By the end of the decade, that number will grow to twenty-nine percent. Of the 12 million additional households by the year 2010, 8 million will be minority households.

WHAT EMERGING TECHNOLOGY HAS TO OFFER

Ability to Manage Data and Tap Information Sources

We are all familiar with the term "information overload." A great deal of information is already being routinely collected and reported to the federal government with respect to mortgage lending. A primary source of loan applicant data for home purchase, improvement, and refinance is the publicly available HMDA database, which is compiled annually. The 1998 edition of the HMDA database contained over 24 million 60-byte records for a total of 1.4 gigabytes of information. There are other publicly available information sources, including government and commercial websites. Powerful search engines allow users to quickly locate alternative providers. In addition, there is a wealth of information available to lenders via the credit bureaus and their own internal data warehouses.

To help organize and manage all of this information, powerful computing platforms, equipped with query and analysis tools, have emerged. Due to the way Bank of America has evolved (largely through mergers and acquisitions), we currently access our information via a variety of sources, including Teradata, DB2, flat files, SAS datasets, Sybase, Oracle, etc. A dedicated NDM Gateway Server provides us with connectivity and the data extracts on it are purely transient. We utilize an *access hub*[10] concept for producing analysis-ready data. Our philosophy is "Get It - Use It - Archive It – Discard It." In so doing, we avoid the possibility of discrepancies stemming from two independently maintained data sources, and our storage requirements are not excessive. Construction of analysis-ready data is often performed on an ad hoc basis. Our primary programming tool is SAS, with some SQL.

Ability To Identify Under-Served Communities And Customer Segments

Data mining is a technology that has the ability to quickly locate significant relationships and patterns in very large databases. These same patterns might take several analysts months to uncover using the traditional approach of first postulating a model, then constructing model input data flows, and coding a program to link the data to a standard statistical routine for analysis. Our sole data-mining tool is SAS Enterprise-Miner™.

Data visualization is useful in exploring key drivers and their interactions. Multi-way effects associated with performance metrics such as origination, decline, and fallout rates can be viewed and subjected to further multi-dimensional drill-downs. Technological capabilities for viewing key mortgage lending performance drivers, and their combined effect on a variety of target variables is evolving. At Bank of America we have developed a *hypercube*[11], which enables one to simultaneously view nine dimensions. Three reside on a main grid, with drill-down capability to two additional three dimensional expansion grids. The boxes floating within the various grids can represent a number of different metrics via length, width, color, and intensity. In this example the grid dimensions consist of categorical variables which serve to partition the mortgage applicant population in a desirable fashion and the boxes are characterized by continuous variables which represent the number of mortgage applications and the dollar volume of originations. Our primary data visualization tool is a generic model developed several years ago using Visual Insight's Discovery 2.2.1 product.

The ability to identify under-served communities and qualified low-to-moderate income mortgage loan applicants is of increasing importance. Data mining-based exploratory analysis, aided by data visualization, deepens our understanding of our communities and customers and can enable us to locate, and even anticipate, emerging mortgage lending opportunities that continue to surface nationwide. Data visualization can prove invaluable as a means of conveying less than intuitive answer sets returned by data mining algorithms.

EXAMPLE: IDENTIFICATION AND PRIORITIZATION OF EMERGING OPPORTUNITIES FOR MORTGAGE LENDING

MOTIVATION

In May 2000, Bank of America announced that the first-year results of its 10-year, $350-billion community development lending and investment commitment totaled $39.6 billion in 1999. The total for affordable housing in 1999 was $25.1 billion. The commitment to affordable housing includes loans and investments as well as mortgages. "Through this lending and investment strategy we are working to help build a stronger America, one community at a time. We will build on the momentum and success of this first year to spur more lending and investment in low- and moderate-income communities, neighborhoods and rural areas."[12] In setting its 10-year goal, Bank of America also committed to acquiring, building and rehabilitating 50,000 affordable housing units. In 1999, Bank of America financed, developed or rehabilitated more than 18,000 affordable multi-family and single-family units.

The motivation for undertaking this particular study was many-fold. Needless to say, Bank of America will continue to rely on traditional methods, and also on strategic partnerships, to enable it to receive input and guidance from organizations and individuals within our communities, so that the bank can be confident that its resources are going where they are most needed. This study was intended to see how these existing processes might be effectively augmented. We had five primary objectives in mind, namely: 1) positive impact on community, 2) profitable business growth, 3) expand banking franchise, 4) mitigate fair lending risk, and 5) promote brand name. In concept, these are meaningful goals that are certainly worthy of research. A discussion of how one might go about quantifying these factors would be a worthy topic for an entire article. For the current effort, we chose percent mortgage loan origination rate as our target variable.

The purpose of this example is to illustrate how one would use data mining and visualization techniques to locate and rank homeownership-deficient communities for mortgage lending marketing initiatives. In 1999, Bank of America provided more than $23 billion in mortgages to low-and moderate-income census tracts and to low- and moderate-income individuals. More than half—$12 billion—went to minority borrowers. We, the authors, did not attempt to define the term "homeownership-deficient community" directly. Instead we focused on low-to-moderate income census tract groupings within most MSAs (over 400 MSAs were examined). The information utilized for this study was comprised of HMDA, Census Bureau, and lifestyle/socio-economic profiling data. This information was first summarized at the census tract level, and then further aggregated, within MSA, into three mutually exclusive sets based upon average income ranges. The three ranges defining our basic operational unit are: low, moderate, and combined middle-high groups, where the groups are as defined on page 4. Pre-processing and sub-setting of the data resulted in 713 observations for this analysis.

DATA PRE-PROCESSING

The most arduous and time-consuming task revolved around the sourcing and pre-processing of the data. Sourcing involved: 1) the specification of the sources and scope of information required, 2) identification of data elements available and the mapping rules used to extract the desired fields into our research database. Pre-processing entailed: 1) the validation, 2) the aggregation, 3) the transformation of existing variables, 4) generation of new variables, sampling, and selection of final candidate variables for the modeling stage.

Sourcing: Following is a review of the data sources for this example.

- 1998 HMDA Data

1998 HMDA is composed of over 24 million real estate secured HMDA reportable applications. HMDA data is publicly available and the majority of all real estate secured applications is HMDA reportable. Therefore the HMDA database provides a comprehensive view of real estate lending activity for the whole nation. This database contains the following information for every loan application.

1. Loan Type: Conventional, FHA, VA, and FSA/RHS
2. Purpose: Home purchase, home improvement, refinance, and multifamily dwelling
3. Owner Occupancy: Owner occupied, not owner occupied, and not applicable
4. Action: Loan originated, application approved but not accepted, application denied, application withdrawn, application incomplete, loan purchased
5. Applicant and Co-applicant race
6. Applicant and Co-applicant gender
7. Financial Institution

- Macro Economic Data

Macro economic data was provided by an internal bank source. All of the data was available in a quarterly time series starting from March of 1997 and ending in June of 1998. The data can be obtained from the following three sources: National Association of Realtors, Office of Federal Housing Enterprise Oversight, and the Regional Financial Associates, Inc. The time series are described below.

- Regional Financial Associates, Inc. Data

1. Existing home sales at the state level (thousands)
2. Disposable personal income at the state level (billions $)
3. Total personal income at the state level (billions $)
4. Single family permits at the state level
5. Nonagricultural Employment at the MSA level (thousands)
6. Unemployment Rate at the MSA level

- National Association of Realtors

1. Housing Affordability Index

- Office of Federal Housing Enterprise Oversight

1. House Price Index

- Marketing Demographic Data

This marketing demographic data was obtained from an internal bank source. This data was composed of 51 variables supplied by Claritas. The Claritas data is at the census tract level

- Low Income Census Tract Field

This is a categorical variable that flags low and moderate-income level census tracts, as defined below:

  - low - < 50% of area median income
  - mod - 50% - <80% of area median income
  - mid - 80% - <120% of area median income
  - up - 120% and > of area median income

- Census Bureau Data

This analysis incorporated 54 variables from the 1990 U.S. Census. All of the variables were selected from the Standard Statistical File 3A. The census data is at the census tract level

Preprocessing: Following is a description of the steps preceding the model building.

- 1998 HMDA Data Processing

1. Merge low-income census tract field with the 1998 HMDA data using state, county, and census tract as merge keys.
2. Only keep home purchase and refinance HMDA applications.
3. Calculate the total number of HMDA applications and the total number of loans originated at the state, MSA, and census tract income level. This aggregation produces 1,225 observations.
4. Calculate the proportion of HMDA loans originated. This is the ratio of the total number of loans originated to the total number of HMDA applications

- Census Bureau Data Processing

1. Merge low-income census tract field with the 1990 Census data using state, county, and census tract as merge keys.
2. Sum all variables to the state, MSA, and census tract income level.

- Marketing Demographic Data

1. Merge low-income census tract field with the marketing demographic data using state, county, and census tract as merge keys.
2. Sum all variables to the state, MSA, and census tract income level.

- Combine Data

1. Merge aggregated HMDA, Census, and marketing data using state, MSA, and census tract income level as merge keys.
2. Next the MSA level macro economic data was merged with the combined HMDA, Census, and marketing data using state and MSA as merge keys.
3. Finally, the state level macro economic data was combined with the HMDA, Census, marketing, and MSA economic data using state as a merge key. Approximately 9% (112 observations) of the aggregated HMDA file did not merge with the Census, marketing or economic data. These observations were not included in the analysis.

- Subset data

This analysis only examined low and moderate-income level census tracts. This final data set contained 713 observations. For this analysis we were fortunate that all input variables were previously scrubbed, so we did not have to deal with missing values or out-of-range values. Furthermore, eighty-seven additional input variables were constructed from the raw census and marketing data by further aggregation and creating ratios.

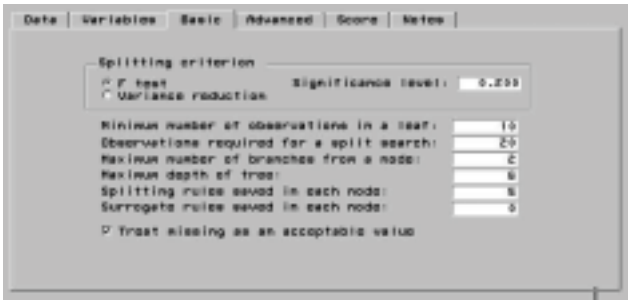
STATISTICAL MODELING

Variable Selection

The input dataset (final.sd2) used for modeling contained 713 observations with 216 potential explanatory variables. All of the explanatory variables are interval scaled. We employed ten bootstrap iterations in the Tree[13] Node to find a subset of variables with the most potential predicative power. The bootstrap algorithm is described below.

Iteration Description

1. Sample 713 observations with replacement from the input data set.
2. Inside the Tree Node, set the parameters in the Basic and Advanced Tab to the settings displayed below.



3. After fitting a tree model, output the variables used in the model to a SAS dataset.

Combining Bootstrap Iterations

1. Merge the 10 datasets with the variables used in each of the 10 tree models. Determine how many times each of the 216 input variables were included in a tree model.
2. Use the input variables that occurred at least four times (18 variables) in a tree model as the final subset variables to use in modeling. Based on business knowledge, two variables were excluded from the 18 and eight more were added from the input variables that appeared in tree models three times or less. Therefore, 24 variables were selected as the final subset to use in modeling.

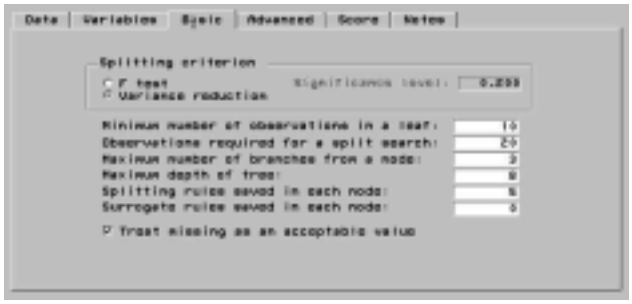
Sampling and Target Variable Transformation

For all of the modeling techniques describe below, the target variable analyzed was the logit transformation of the proportion of loans originated. The logit transformation was used to insure that predicted probabilities would be between zero and one. After transforming the target variable, the modeling dataset with 24 input variables was partitioned into training and validation datasets (80% and 20%

partition respectively). The validation dataset was used to find the models with the smallest validation dataset errors. These models are expected to perform better than their competitor models on new datasets. As a result, the final model selected was a model with one of the smallest validation errors.

Tree Models

Using the Tree Node, eight tree models were fit to the input data. The eight models varied the maximum number of branches from a node (2 through 5) and the splitting criterion (F test and variance reduction). The basic and advanced tab for one of the tree models is displayed below.



The best tree (maximum four branches from a node using variance reduction) had a validation average squared error of 0.1396. This corresponds to a sum of squares error of 19.96.

Logistic Regression Models

Using the Regression Node, eight stepwise logistic regression[14] models were fit to the input data. The regression models varied the transformations applied to the input variables and whether or not two-way interactions were allowed in the model. The transformations used were no transformation, maximize input correlation with the target variable, maximize normality of input, and bucket input into three categories using quantiles. When a transformation was used, the transformation was applied to all of the input variables. Transformations were not mixed within a regression model. The next figure shows the Selection Method Tab in the Regression Node for one of the eight stepwise regression models.

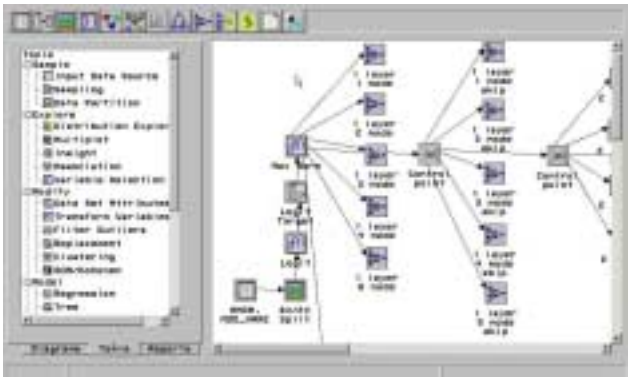
For each stepwise regression model development, the stepwise selection method generated several candidate models. The final model selected was the one with the smallest validation error (error sum of squares). Among the eight stepwise regression models, the model with the smallest validation error (error sum of squares = 17.09) was the main effects regression model with nine input variables using the maximize normality of input transformation.

A correlation matrix was produced for the nine inputs in the best regression model. Two of the inputs were highly correlated (0.99) indicating a collinearity problem between these two variables. Based on business judgement, one of the two variables involved in the collinearity problem was dropped because collinearity problems cause affected variables to have unstable parameter estimates. When one of the inputs was removed, the error function changed to 17.50. The final eight variables selected appear in the table below:

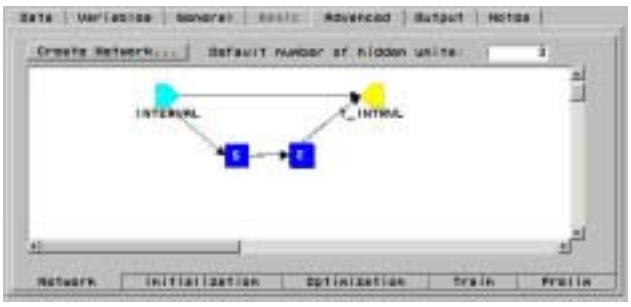
Name	**Description**
AI98Q1	Affordability Index – Q1 1998
HI97Q4	House Price Index – Q4 1997
WHTOW	Proportion of Non-Minority Home Owners
TOTHISP	Number of Hispanics
ASNAM	Proportion of Asians
LOWFAM	Proportion of Low Income Families
POVHH	Proportion of Poverty Households
RURDNSC	Number of Rural Downscale Households

Neural Network

Because neural networks[15] can require substantial training time, only the eight input variables in the best regression model were used as inputs for a neural network. Using the neural network node, 42 neural networks were trained. The schematic below shows a portion of the diagram containing the 42 neural networks.



All 42 networks were multi-layer perceptrons. The 42 networks varied the number of nodes in the first hidden layer (one through five), whether or not the network included a skip layer, whether or not the network had a second hidden layer (two nodes), and whether or not a weight decay parameter was used during training. Using the advanced tab feature in the neural network node, one can vary the network architecture. The diagram below shows how to construct a network with two hidden layers and a skip layer.



The errors of the best three neural networks are displayed in the table below.

First Hidden Layer	Second Hidden Layer	Skip Layer	Weight Decay	Error

Nodes	Nodes			
1	0	No	No	16.95
4	2	No	Yes	17.13
5	2	No	Yes	17.15

Table B. Best Three Neural Networks

Ensemble Models

Sometimes an average of several results from different models is more accurate than a result from a single model. For example, a neural network may produce different results, especially when early stopping is used, since the results may be sensitive to the random initial weights. Tree-based models are also very sensitive to minor changes in the input data. To stabilize the results from these models, the ensemble model approach, based on multiple models (or multiple samples), has been introduced.[16] Enterprise Miner™ provides a variety of ways to combine models using the Ensemble Node and the Group Processing Node. In this paper, we used only two types of ensemble models for the purpose of illustration.

The first ensemble model is a combined model using the best tree, logistic regression, and neural network models selected from above. Best means smallest validation error. The ensemble model predicted values are produced by averaging predicted values from the best tree, logistic regression, and neural network models. The second ensemble model estimates predicted values via the bagging (bootstrap aggregation) algorithm[17]. Instead of varying the algorithm, the data is modified using the bootstrap algorithm (sets of random samples with replacement). We applied the bagging algorithm to a logistic regression, tree, and three different neural networks. Then the predicted values are averaged in each modeling approach separately. For each type of model the predicted values are calculated by averaging across the bootstrap samples. The validation errors for the combined model, logistic regression with bagging, tree with bagging, and three neural networks with bagging are shown in the following table.

Bagging	Validation Error
Combined Model (tree, logistic regression, neural network)	17.22
Regression	17.63
Tree	19.36
Neural Net, 1 Layer, 1 Node	18.35
Neural Net, 2 Layers, 4-2 Nodes	17.26
Neural Net, 2 Layers, 5-2 Nodes	17.11

DATA VISUALIZATION: 9-D HYPERCUBE

We deployed a generic, nine-dimensional data visualization model for the purpose of analyzing loan application data over a three-year period. It allows the user to impose a structure on the data via categorical variables, and define a set of measures that summarize the observations contained within each cell of the resulting *hypercube*. The user can navigate the hypercube by selecting three factors of interest, as shown in Figure 1.

Figure 1 Hypercube XYZ-Axis 1 for HMDA Data Exploration

The dimensions in this primary grid are:

1. **Region:** West, Southwest, Midwest, Central, Northeast, Southeast
2. **Census Tract:** Low Income, Mod Income, Upper Income
3. **Institution:** Competitor, Bank of America

By clicking on one of the floating bars in the primary grid, a drill-down procedure is invoked, which renders the result in Figure 2.

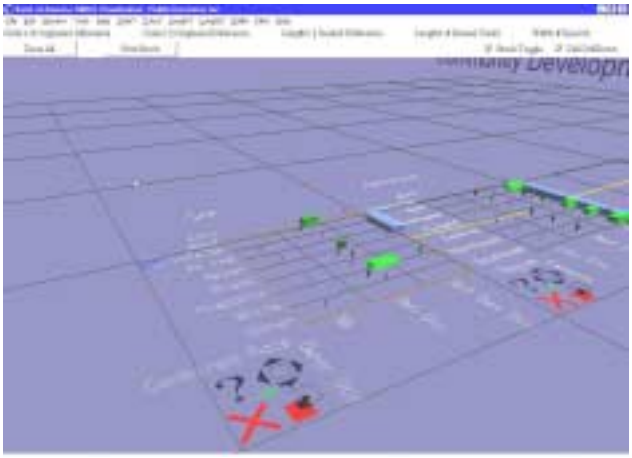


Figure 2. Hypercube Second XYZ-Axis: Drill-Down for Competitor, Mod-Income and Central Region

The dimensions in this secondary grid are:

4. **Combined Race:** No Information/NA, Am Indian, Asian, Black, Hispanic, White, and Other
5. **Owner Occ:** NA, Owner Occupied, Not Owner Occupied
6. **Type:** Government, Conventional

By clicking on one of the floating bars in the secondary grid, the drill-down procedure is invoked again, which renders the result in Figure 3.

Figure 3. Hypercube Third XYZ-Axis: for Conventional Mortgage, African-American, and Owner-Occupied Housing

The dimensions in this final grid are:

7. **Purpose:** Home Improvement/Multi-Family, Purchase, Refinance
8. **Loan Amt:** <$100K, 100K to $200K, $200K +, Unknown
9. **Income:** <$25K, $25K to <$75K, $75K +, Unknown

There are two bars associated with each cell, and the length and color of each pair of upper and lower bars can be associated with a different measure. The width may also be associated with a measure, but is the same for both bars. By touching a bar with the cursor, a brush appears, as in Figure 3, which lists all of the values for the following measures:

Measures: Number Records, Number Bank of America, Number Competitor, Number Originated, $K Originated, Number Originated Difference, $K Originated Difference, Number Denied Credit, $K Denied Credit, Number Denied Difference, $K Denied Difference, Number Fallout, $K Fallout, Number Fallout Difference, $K Fallout Difference, Number Non Purchased, Number Purchased, $K Purchased, % originated, % denied, and % fallout.

At any point, the user has the option of specifying the XYZ dimensions, and also check-off which groups within each dimension to include in the exploration, via the panel displayed below.

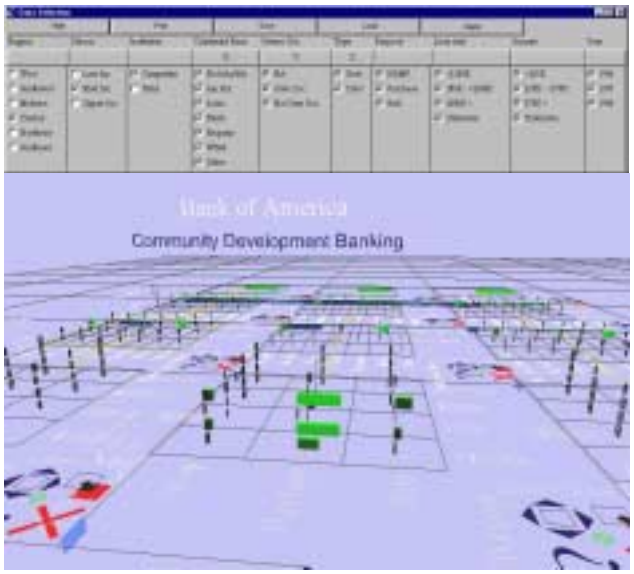


Figure 4. Navigating the Data – Temporal View for 1996-1998

The tenth dimension is time, in this case we have chosen annual periods, namely 1996, 1997, and 1998. Hypercubes for the three years are aligned side-by-side for comparative purposes in Figure 4. We have also defined measures as the year-to-year difference for a set of variables, so that we can easily spot positive (green-colored) and negative (red-colored) trends over time. This tool enables the user to view non-intuitive combined effects and navigate the data mining answer set to gain further insight. Sessions can be animated to produce a visual story for communicating the underpinnings of empirical results to an audience in real-time mode.

DISCUSSION OF FINDINGS AND SUGGESTIONS FOR FURTHER STUDY

I. The chief objective of our modeling exercise was to find the characteristics that jointly possess the greatest information value for identifying communities where we can achieve the greatest success in originating mortgage loans.

We selected a bagged neural network with two hidden layers (5 and 2 hidden nodes in each layer, respectively) as our champion model. We selected this model because it had one of the smallest validation data set errors. Even though the neural network with one hidden layer and one hidden node had the smallest validation error, this architecture did not perform very well when *bagged*. We felt that the bagged neural network with two layers would provide better performance when scoring new datasets, i.e. the predicted values would be closer to the actual values. Figure 5 displays predicted and actual probabilities for the champion model.

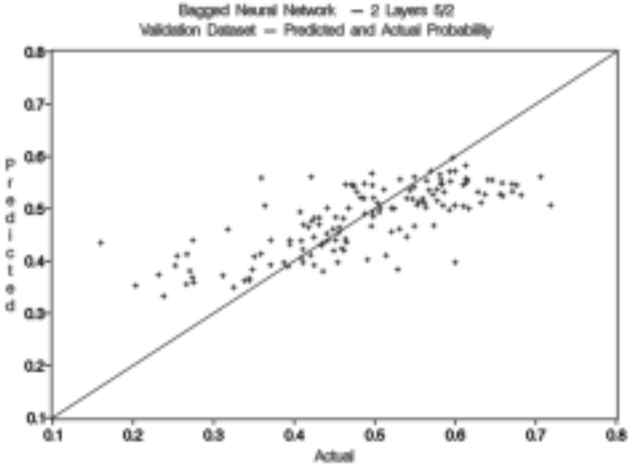


Figure 5. Predicted vs. Actual Values for Target Variable

The model is high on the low end of the actual target variable range, and it also tends to underestimate the probability of origination on the higher end. This is a deficiency in the current model. Similar patterns were witnessed in all alternative formulations, i.e. the regression model, the tree model, etc. In a future study we plan to examine this data at a lower level, e.g. census tract (approximately 65,000 observations). We will broaden the scope of candidate variables by incorporating non-public Bank of America loan performance data and customer relationship information. We are also looking forward to periodically refreshing our data sources as new information becomes available, e.g. the results from the 2000 census.

Figures 6-8 were used to gain insight as to the structure of the correlation between the explanatory variables and the probability of origination. Figure 6 shows the relative distribution of each of the model's eight explanatory variables with respect to the predicted probability of origination.

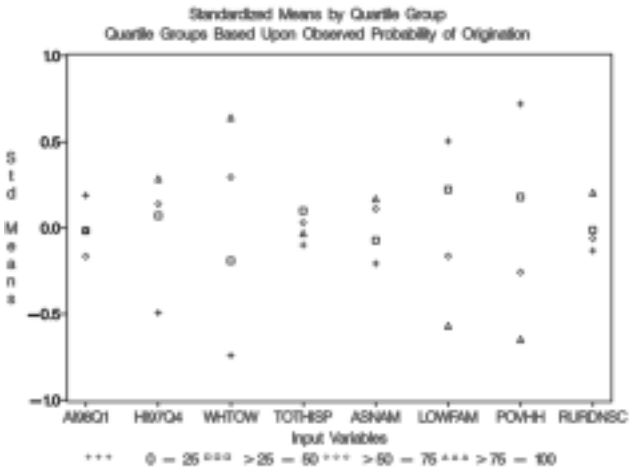


Figure 6. Comparison of 8 Variables in Final Model

The data was divided into four groups, based upon quantiles of the observed probability of origination. Each of the 713 observations was assigned to one of four groups, namely:

1. Observations with origination rates < 25th percentile
2. Observations with origination rates > 25th percentile, but less than the median
3. Observations with origination rates > median, but less than the 75th percentile
4. Observations with origination rates > 75th percentile

The mean of each input variable is calculated for the four groups. For each variable, the group means were standardized using the following formula:

$$\frac{\text{(group mean – overall mean)}}{\text{overall standard deviation}}$$

The y-axis displays the individual standardized means for each variable and group combination. For example, consider the variable percent of households at the poverty level (POVHH). The lowest probability of origination occurs when POVHH is the greatest.

Figures 7 and 8 attempt to show the relationship between the probability of origination and two explanatory variables (POVHH and AI98Q1) for the champion neural network. Within each graph, three lines are plotted. These correspond to all variables being held to constant value except for the variable being displayed on the x-axis. For example, the solid line in Figure 7 shows the probability of origination as a function of poverty households with the other seven explanatory variables held at their 25th percentile value. We created these plots for all eight explanatory variables, but have shown only two for this paper due to space considerations.

The plots have provided us with insight as to how explanatory variables impact the probability of origination, and they also provide a check to ensure that we do not have nonsensical associations between the target variable and the explanatory variables. We felt these plots were important for a neural network because there is no intuitive interpretation for parameter values.

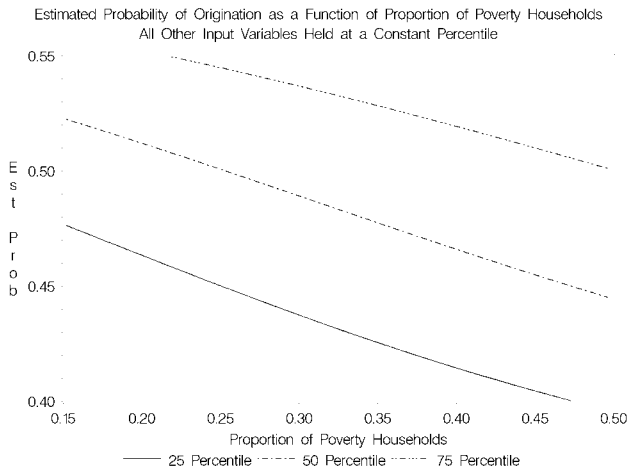


Figure 7. Relationship of POVHH and Predicted Probability of Origination in the Final Model

Examination of Figure 7 shows that when all variables are held constant, the probability of origination declines as the proportion of poverty households increases. This relationship agrees with the univariate relationship displayed in Figure 6.

Figure 8 shows a higher affordability index translates to a higher probability of origination, when all other variables are held constant at their 25th, 50th, and 75th percentiles, respectively. This contradicts the relationship shown in Figure 6 for AI98Q1, namely that lower affordability implies higher proportions of origination.

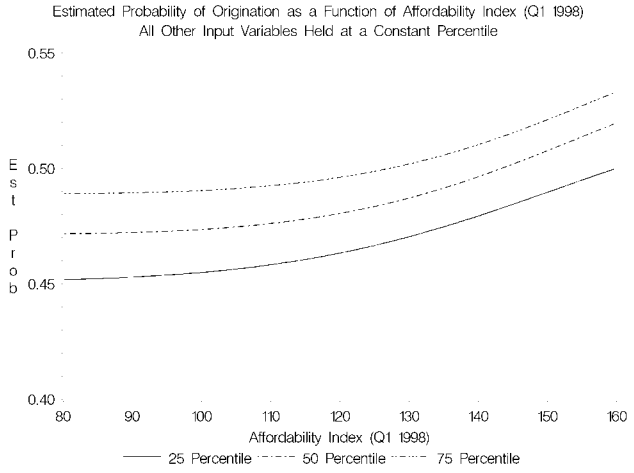


Figure 8. Relationship of Affordability and Predicted Probability of Origination

This seeming contradiction is due to the fact that the affordability index is negatively correlated with the housing price index (Pearson Correlation was -.0.55) However, when you adjust for house price index, the relationship between affordability index and proportion of origination changes, which is intuitive. In other words, if house prices are stagnant and income levels rise, the proportion of origination increases.

Rank	MSA Name	State	Low/Mod Income	Predicted Probability
1	Portland,	ME	M	0.5971
2	Hagerstown	MD	M	0.5904
3	Rochester	MN	M	0.5902

4	Minneapolis--St. Paul	MN	M	0.5845
5	Appleton--Oshkosh--Neenah	WI	M	0.5843
6	Green Bay	WI	M	0.5841
7	La Crosse	MN	M	0.5832
8	Sioux Falls	SD	M	0.5829
9	Lewiston--Auburn	ME	M	0.5828
10	Wilmington--Newark	MD	M	0.5799

Figure 9. Top Ten Observations, Ranked by Predicted Probability of Origination

Based upon the champion model, the top ten list of observations affording the greatest probability of success in originating mortgage loans appears in Figure 9. Next year, when more current data becomes available from our key sources, we plan to score low-mod census tracts and compare the results with our current findings.

II. Another objective of our modeling exercise was to identify opportunities to better serve minority communities. By using the model, we were able to identify minority communities exhibiting a higher probability of origination and a historically lower share of homeownership. We compared the results of the final model with a minority housing index which we constructed by dividing the percentage of minority households in the population by the percentage of minority homeowners for each of the 713 low and moderate income census tract group observations. The results appear in Figure 10.

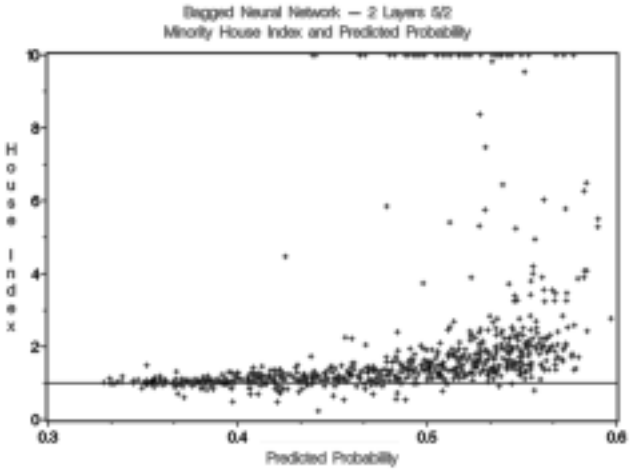


Figure 10. Relationship of Minority Homeownership and Predicted Probability of Origination

For instances where there were no minority homeowners, or very few, the housing index was very large. For presentation purposes, we opted to set the housing index value for those observations to a constant value of 10 (these points appear at the top border of the plot). This result demonstrates that there are communities where we have both a disproportionately low homeownership rate among minorities, and also a higher probability of origination (refer to points in the upper right-hand side of the plot). We conclude that significant opportunity exists to achieve profitable mortgage loan growth, while mitigating fair lending risk. Further study will be required, at a more granular level, to explore and quantify the market potential and determine the most suitable approaches and options.

Data mining, through the use of SAS Institute's Enterprise Miner, can help find potential new mortgage loan customers in low-to-moderate income neighborhoods and communities. Bank of America has a staff of specialists and mortgage loan programs especially geared towards assisting applicants who are oftentimes constrained by a variety of circumstances. Bank of America remains committed to finding solutions to make the dream of homeownership become a reality for residents in all of the communities that it serves.

REFERENCES

1. "The Role of Financial Services in the Year 2020," Joel P. Friedman, Managing Partner, Strategic Services – Americas, Anderson Consulting, Presentation at *Achieving Peak Performance*, Vision '97: The Experian Credit Conference, October 7, 1997. Mr. Friedman's enlightening observations formed the basis for much of the discussion.
2. "A Rallying Cry for Affordables," Kim Renay Anderson, *National Mortgage News*, December 20, 1999, page 14.
3. All housing characteristics are based upon Census Bureau reports. Source: National Association of Home Builders cited in *National Mortgage News*, December 20, 1999.
4. The authors wish to acknowledge that, in addition to the ultimate goal of homeownership, there is certainly a continuing need for affordable rental housing for low-income families. One-third of the nation's rental units is single family homes. Fifty-four percent of the rentals are outside of the central cities. Two-thirds of the landlords are private landlords owning five or fewer units. Since the early 90's, federally subsidized rental housing has fallen by ninety thousand units. In some instances, landlords have dropped out of federal programs so they can obtain market-value for their rentals. In other cases, housing projects have been demolished, without replacement, or replaced with lower capacity housing developments. Sources: "Despite Boom, Minorities Still Lag in Homeownership," Tracie Rozhon*, The Charlotte Observer*, July 15, 2000, Section E, Pages 1,5 and reference 5 below.
5. "The State of the Nation's Housing 1999," Joint Center for Housing Studies of Harvard University, p. 3
6. "Statistics and Industry Trends," Dr. Charlene Sullivan, *Consumer Trends,* published by Creditors International, Volume 1, Issue 5, May 2000.
7. "Woodward: Internet Alters Industry's Value System," Mike Sorohan, *Real Estate Finance Today*, Volume 17, Issue 5, February 7, 2000, p.1
8. "Woodward: Internet Alters Industry's Value System," Mike Sorohan, *Real Estate Finance Today*, Volume 17, Issue 5, February 7, 2000, p.28
9. "Internet Offers Challenges and Opportunities," Mike Sorohan quotes David Matthews, president of Ultraprise Corp. of Shepherdstown, W.Va., *Real Estate Finance Today*, Volume 17, Issue 5, February 7, 2000, p.1
10. "Operational Considerations in Data Mining," Dr. Jim Sattler, *Computing Science and Statistics*, Interface Foundation of North America, Proceedings of the 29th Symposium on the Interface, Houston, TX, May 14-17, 1997, Editor David W. Scott, p. 103
11. "Opening Remarks for Business Data Mining: Application Views," Clark Abrahams, *Computing Science and Statistics*, Interface Foundation of North America, Proceedings of the 29th Symposium on the Interface, Houston, TX, May 14-17, 1997, Editor David W. Scott, pp.357-359
12. "Bank of America Community Development Lending Exceeds $39 Billion -- More Than $14 Million Every Business Hour ", Press Release — May 11, 2000 Washington, DC, quote from Hugh l. McColl, jr., Chairman and Chief Executive Officer of Bank of America.
13. "Decision Trees for Predictive Modeling," Padraic G. Neville, SAS Institute Inc. publication, July 16, 1999
14. Applied Regression Analysis, A Research Tool, John O. Rawlings, Wadsworth & brooks / Cole Advanced Books & Software, 1988
15. Neural Networks for Pattern Recognition, Christopher M.

Bishop, Clarendon Press - Oxford, 1998
16. "Machine Learning Research: Four Current Directions", Thomas G. Dietterich, *Artificial Intelligence Magazine 97-136*, Winter, 1997,
17. "Bagging Predictors", Leo Breiman, Machining Learning, vol. 24, 123-140, 1996

ACKNOWLEDGMENTS

The authors wish to thank Diane M. Comstock, for sourcing and programming assistance associated with data gathering and pre-processing, and also Frank Tucker, for designing and programming associated with the data visualization hypercubes. We are also indebted to Jerry Oglesby for his support, and the overall encouragement provided by the management of our respective organizations.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author(s) at:

Clark R. Abrahams
Bank of America
101 S. Tryon Street, NC1-002-18-02
Charlotte, NC 28255-0001
Work Phone: (704) 386-1079
Fax: (704) 386-6662
Email: clark.abrahams@bankofamerica.com
Web: http://www.bankofamerica.com

Frank R. Burnett
Bank of America
201 N. Tryon Street, NC1-022-03-02
Charlotte, NC 28255
Work Phone: (704) 388-8884
Fax: (704) 386-4746
Email: frank.burnett@bankofamerica.com
Web: http://www.bankofamerica.com

Dr. Jin-Whan Jung
SAS Institute Inc.
SAS Campus Drive, Bldg. T
Cary, NC 27513
Work Phone: (919) 677-8000 Ext. 1-4365
Fax: (919) 677-4444
Email: jinwhan.jung@sas.com
Web: http://www.sas.com