Loan Prepayment Modeling

Afshin Goodarzi
Risk Monitors Inc.
50 Main Street
White Plains, NY, 10606
afshin@riskmonitors.com

Ron Kohavi
Silicon Graphics, Inc.
2011 N. Shoreline Blvd
Mountain View, CA 94043
ronnyk@engr.sgi.com

Richard Harmon
Risk Monitors Inc.
50 Main Street
White Plains, NY, 10606
r_harmon@riskmonitors.com

Aydin Senkut
Silicon Graphics, Inc.
2011 N. Shoreline Blvd
Mountain View, CA 94043
asenkut@engr.sgi.com

Abstract

Loan level modeling of prepayment is an important aspect of hedging, risk assessment, and retention efforts of the hundreds of companies in the US that trade and initiate Mortgage Backed Securities (MBS). In this paper we review and investigate different aspects of modeling customers who have taken jumbo loans in the US using MineSet™. We show how refinancing costs differ across states and counties, and which attributes make good predictor variables for prepayment forecasts. Our data comes from the McDASH Analytics database containing real data, which tracks loans over the past nine years at monthly intervals.

Introduction

Loan level modeling of prepayment is an important aspect of hedging, risk assessment, and retention efforts of the hundreds of companies in the US that trade and initiate Mortgage Backed Securities (MBS) (Richard & Roll 1989; Brown 1992; Harmon 1996). With at least 52 million mortgages (according to the Mortgage Bankers Association estimates of end of the year 1997) outstanding in the US and the securities being traded every day the stakes are very high and the potential gains/losses are substantial. Our studies indicate that different prepayment estimation/forecast methodologies can easily introduce a 20% to 30% difference in the cash flow of a portfolio. For a typical portfolio, such differences could easily be measured in the hundreds of millions of dollars per year in cash flow alone.

Despite the importance of having loan-level prepay models, models are unavailable except possibly to the large institutional investors that can put the research resources together to come up with these models. Such companies would maintain the secrecy of these models as a competitive advantage.

In a collaboration between Risk Monitors, Inc. and Silicon Graphics Inc., we have embarked on building such models using MineSet™ (Brunk, Kelly, & Kohavi 1997). This project involved the identification and verification of drivers for prepayment forecasts. Even though most of the drivers of prepayment are rooted in economic theories and analysis, we still need to verify these theoretical assumptions against the wealth of historical data that is available to us today. This paper lays out some of the results of this ongoing effort.

Prepayment in Mortgages

A typical mortgagee makes a commitment to pay the mortgagor in equal payments on a monthly basis for the term of a loan. Included in each contract is the right of the mortgagee to exercise his/her right to payoff (or prepay) the loan at any point in time. Furthermore, this option is typically exercisable with no financial penalties to be paid to the mortgagor. A mortgage loan is prepaid due to the sale of the underlying property or due to refinancing into another loan. The mortgagor may also terminate the mortgage loan when the mortgagee defaults on the required payments.

There are numerous reasons for the mortgagee to prepay the loan but the most significant factors are typically driven by changes in interest rates, employment status, family status, income, relocation, retirement, health related impacts, etc. Among the financial incentives that contribute to the prepayment of mortgage loans, the most significant is the incentive to refinance an existing loan into a loan with a lower interest rate and payment requirements.

Mortgage investors, mortgage servicers, and other owners of mortgage related financial instruments, are exposed to significant interest rate risk when loans are prepaid and to credit risk when loans are terminated due to default. Prepayments will halt the stream of cash flows that owners of mortgage related financial instruments expect to receive. In may cases this will result in a lower than expected return on their investment. For example, if interest rates decline, there will typically be a subsequent increase in prepayment activity which forces investors to reinvest the unexpected additional cash flows at the new lower interest rate level. This will result in a lower expected return on their mortgage-related investment. On the other hand, if interest rates increase, there will typically be a subsequent decrease in prepayment activity which will force investors to wait for a longer period before they can reinvest the cash flows at the new higher interest rate level. The result will be a lower return than available at prevailing market rates. Therefore, the ability to accurately predict a mortgagees likelihood to prepay the loan is vital to the estimation of the expected return of investors and mortgage servicers.

The Data

The data that was used for this study was supplied by Risk Monitors through an exclusive arrangement with McDASH Analytics. The McDASH database consists of loan level mortgage information on over 11 million loans on a monthly basis dating back to 1989, from many of the largest nationwide mortgage servicers in the US. The raw servicing data files are cleansed by each servicer before being passed onto McDASH Analytics. After receipt of the data, McDASH undertakes additional data integrity checks before the data is added to the McDASH Analytics Database. The Risk Monitors data feed requires some additional data processing to ensure that individual loans cannot be identified as belonging to a specific mortgage servicer. This cleansed data is then passed onto Risk Monitors in its entirety. We then segment this data into several different categories: Investor Type (GNMA, Conventional, Jumbo, Home Equity and B/C) and Product Type (30Yr FRM, 15Yr FRM, 7Yr FRM, 5Yr FRM and ARMs).

The data for each loan dates back to the origination of the loan or to 1989 which ever comes first. In each record the initial properties of the loan as well as the current status of the loan are described. The initial properties include:

Unique identifier Each loan in the McDASH data base is identified to us via a unique number.In this way the confidentiality of the mortgage servicer as well as mortgagee are maintained. The data set contains nothing to identify the servicer and the mortgagee.

Note-rate The interest rate at which the loan was secured.

Closing date The date of transfer of ownership from the seller to the buyer(mortgagee) of a property.

Loan amount The total loan amount in US dollars. Typically this amount is smaller than the property value.

Type of loan Other than the general categories of fixed rate and adjustable rate mortgages, there are numerous more specific types of mortgage products available. A few examples include Balloons, Hybrid ARMs, Mortgages with prepayment penalties, negative amortization mortgages, etc.

State The state in which the property resides.

Zip code The Zip code for the property. The State and the Zip code are the only two indicators of the location of the property. No other information about the the location is available to us.

The current status refers to attributes that change monthly, including:

Current principal balance The amount of principal outstanding. If this amount is paid off the loan is then considered paid off. If the pay off occurs earlier than the term of the loan then the loan is considered prepaid.

Amount of escrow for the month Property taxes to the municipality and the state are maintained in an escrow account by the servicer. The balance of this escrow is reported on a monthly basis.

Current date date of observation.

Status of the loan Active, paid in full, foreclosed, going through foreclosure.

Due date Date of monthly payments coming due. This can be used to determine loans that are 30, 60 or 90+ days delinquent.

Risk Monitors filtered the data further for “bad” data. Where “bad” data is described as records that have large amount of missing data or grossly incorrect attributes. Additional factors derived from other data sources (Historical mortgage rates and Treasury rates were downloaded from EJV Bridge) were appended to each record of the McDASH data set, including:

Burnout A measure of the refinance incentive the loan has been exposed to since the loan was created. If a loan has been exposed to interest rate incentives to refinance and still has not done so, other circumstances may exist that prohibit the borrower from exercising the prepayment option. For example, the credit worthiness of the homeowner may have dropped or a lack of equity in the property.

One year treasury Interest rate. 1-year Treasury CMT Rate.

Ten year treasury Interest rate. 10-year Treasury CMT Rate.

Yield Curve slope (YCS) The difference between the ten year treasury (T10Y) and one year treasury (T1Y) (Litterman, Scheinkman, & Weiss 1991):

$$\mathrm{YCS} = (\mathrm{T10Y} - \mathrm{T1Y})/\mathrm{Average}_{\mathrm{monthly}}(\mathrm{T10Y} - \mathrm{T1Y})$$

Typically, the ten year treasury note has a higher interest rate, but the difference between the curves at a given point in time changes. Regardless of the mortgage rates, different yield curve shapes provide incentives or disincentives for additional refinancing activity.

Present Value Ratio (PVR) The refinance incentive is measured by the ratio of the present value of the existing mortgage’s payments to the annuity value of a new mortgage. The equation we use is from Richard & Roll (1989).

$$\mathrm{PVR} = \frac{I}{R} \cdot \frac{1-(1+R)^{-M}}{1-(1+I)^{-M}}$$

where I is the note rate on a monthly basis (WAC/1200), R is the current mortgage refinance rates on a monthly basis (Mortgage Rate/1200), and M is the remaining life of the loan in months..

Housing delta price index A measure of the change in the value of housing in each state. These are calculated from FNMA/FHLMC repeat sales estimates which are derived from conventional mortgage loans at the aggregate state level.

FNMA commitment rates Four Week Moving Average FHLMC mortgage commitment rates.

Lag Burnout, Yield Curve Slope and Present Value Ratio are all computed with 0 through 12 weeks of lag. We suspect that since the process of closing on a refinance application is several weeks long that the refinancer is reacting to market conditions that are indeed several weeks old. While the exact lag differs, the rule of thumb is to use a lag in the range of four to ten weeks with a lag of eight being the rule of thumb.

For the purposes of this study we will look at one cut of the McDASH data, which included Jumbo loans. Jumbo loans are characterized as loans that are larger than a certain threshold principal balance. In 1997 the jumbo loan threshold is loans of $227,150. The dataset under study has over 1.08 million records representing over 55000 loans. The reporting for this data set start in 1994 and ends in April of 1998. The data set has the following statistical properties:

Number of Loans 55000. (Note: many loans appear and disapear during the time horizon specified for this study, 1994-1998)

Loan range Loans ranged up to $1,065,295.

Loan age The maximum loans were 464 months (38 years).

Note rates Interest rates varied from 3.21% to 16.75% with a mean value of 7.97%

The process of mining this data involves not only finding a good set of predictor variables that best predict the prepayment of a loan, but also attributing such findings to sound recognizable economic principals. Furthermore, we should disentangle interrelationships that might confuse a model or modeler.

The Knowledge Discovery Process

We now describe some of the processes we went through in the knowledge discovery process (Fayyad, Piatetsky-Shapiro, & Smyth 1996; Brachman & Anand 1996).

Data Cleansing

During the initial phase of the analysis we found that data mining can serve as a great tool for showing bad data, thus facilitating cleansing. We found unlikely patterns in the results that Mineset was generating, which prompted further analysis that yielded corrections to the data cleansing and preparation efforts.

Overview of the Data

Figure 1 shows boxplots for several key attributes. These boxplots show basic statistics about key attributes. Such statistics allowed filtering records that were incorrect (*e.g.*, negative loans).

Our data is distributed across the US but a large proportion of it comes from California because we are looking at jumbo loans and California has relatively expensive housing. Figure 2 (left) shows a map of prepayments across the US per state. Figure 2 (right) drills down to zip codes. By far, the zip code with the highest prepayments is unknown with a prepayment percent of 0.5% (five times the national average). This requires further investigation.

Refinancing Costs

One of the factors in refinancing of a loan is the fixed cost of refinancing.(i.e. closing cost on a per county basis). Should the refinance cost end up being large enough (say .25 points or higher), the cost may deter some from exercising the refinance option. It is well known in the industry that refinancing costs differ from state to state. Figure 3 shows that costs not only vary between states but also between counties in the same state. Figure 4 shows a closeup of Oregon and New York. This fact implies that even in a portfolio which is entirely within a state the refinance incentive is quite diverse. In Maryland, the difference between the highest cost and lowest cost is over 0.69 points. You could then roll this penalty into the actual note rate to achieve a higher Annual Percentage Rate (APR). For example, a 30 year mortgage for $150,000 with a note rate of 8.5% will have an APR of 8.73% if the total penalties and points are 1.5%. [1]

Determining the Lag Period

In our original data, we appended to each record several measurements, such as different treasury rates, yield curve slope, and burnout. Each of these measurements were added as multiple attributes with lags varying from one to twelve weeks back.

When we built the Simple-Bayes model (described below), we consistently found that a lag of four weeks was either the best or second best in the attribute ranking order for a given measurement.

We then removed all the other lags to make our dataset narrower and avoid highly correlated attributes, but this fact, by itself, was useful insight as to the time it takes people to react to changes in economic conditions.

[1] the solution for this computation is algorithmic based on (Fabozzi 1995). The equation that is being solved is $Mn/B0 = (R * (1+R)^N)/((1+R)^N - 1)$ where R = Noterate/1200, Mn = Monthly payment, B0 = Balance, N = Original term of the loan.

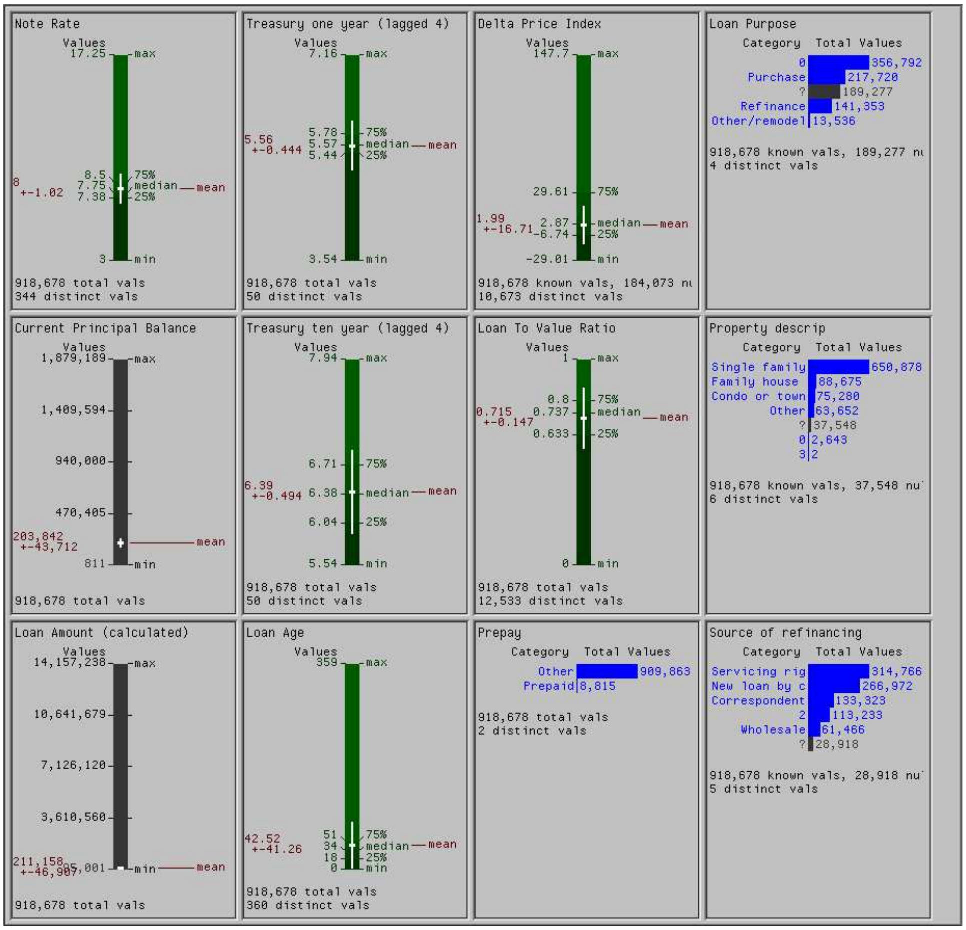

Figure 1: Basic statistics about some attributes.

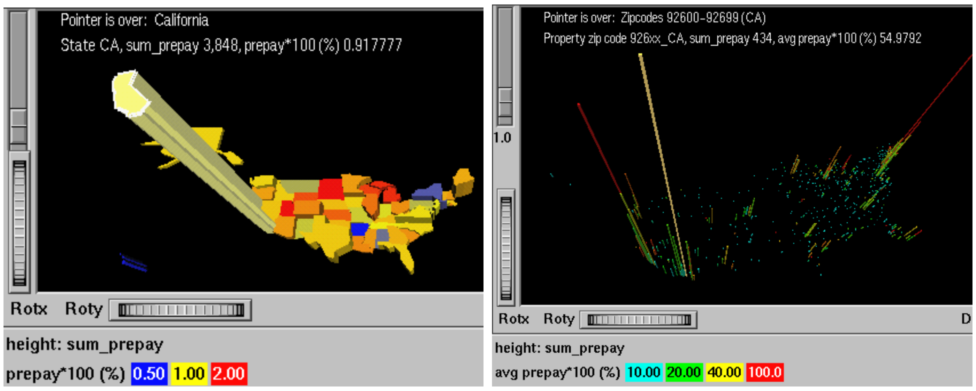

Figure 2: The prepaid loans by states (left) and zip codes (right). The height is the number of prepaid loans; the color is the averge percent times 100.

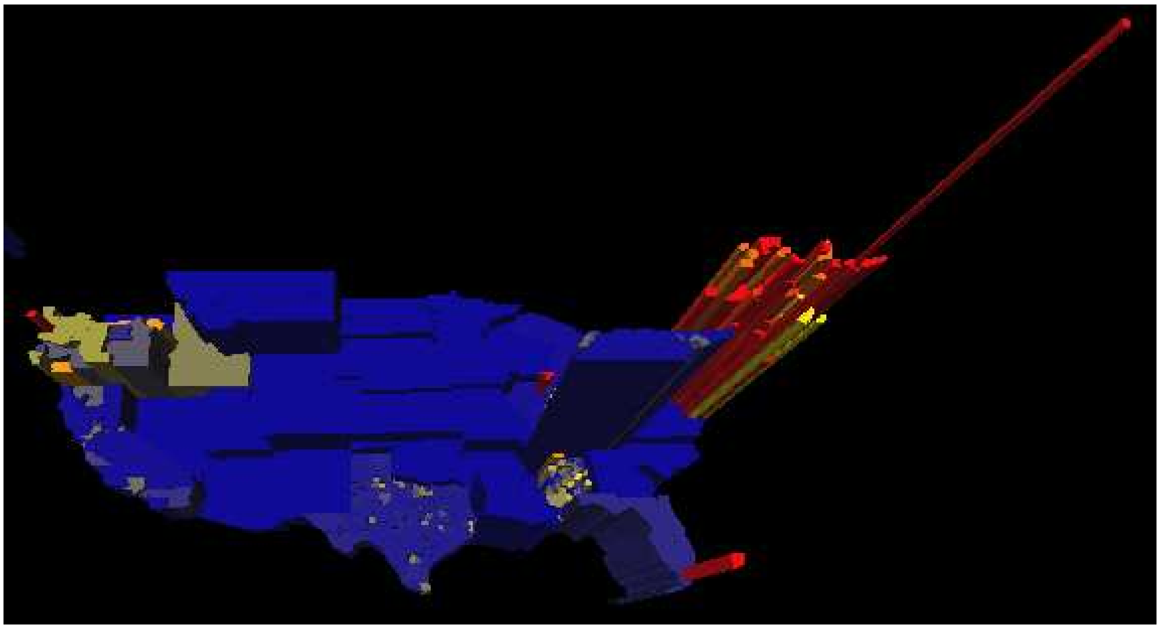

Figure 3: Refinancing costs (mapped to height) for every county based on FIPS codes. Deviations from each state's average are colored from blue (zero deviation) to yellow (0.005) to red (0.01).

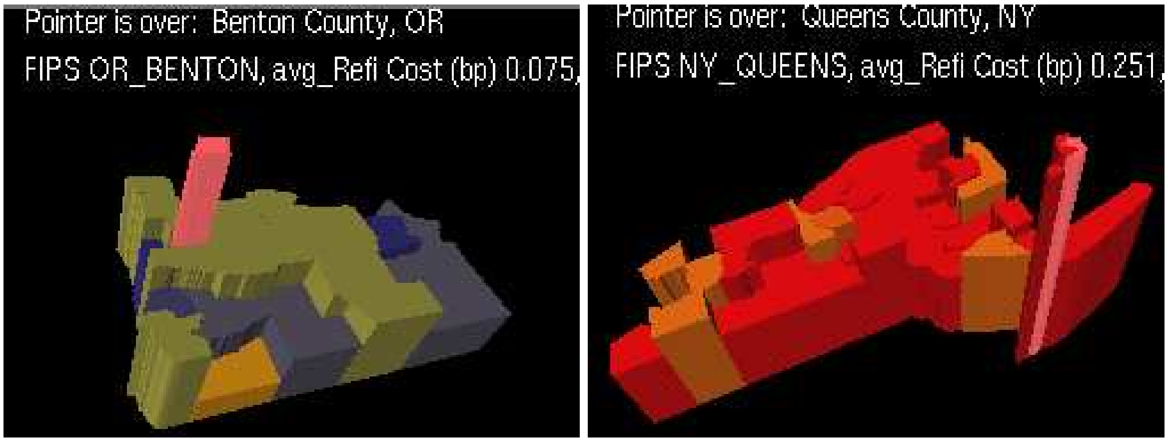

Figure 4: Refinancing costs for Oregon (left) and NY (right). Differences between neighboring counties are on the order of hundreds of dollars: enough to deter some people from refinancing.

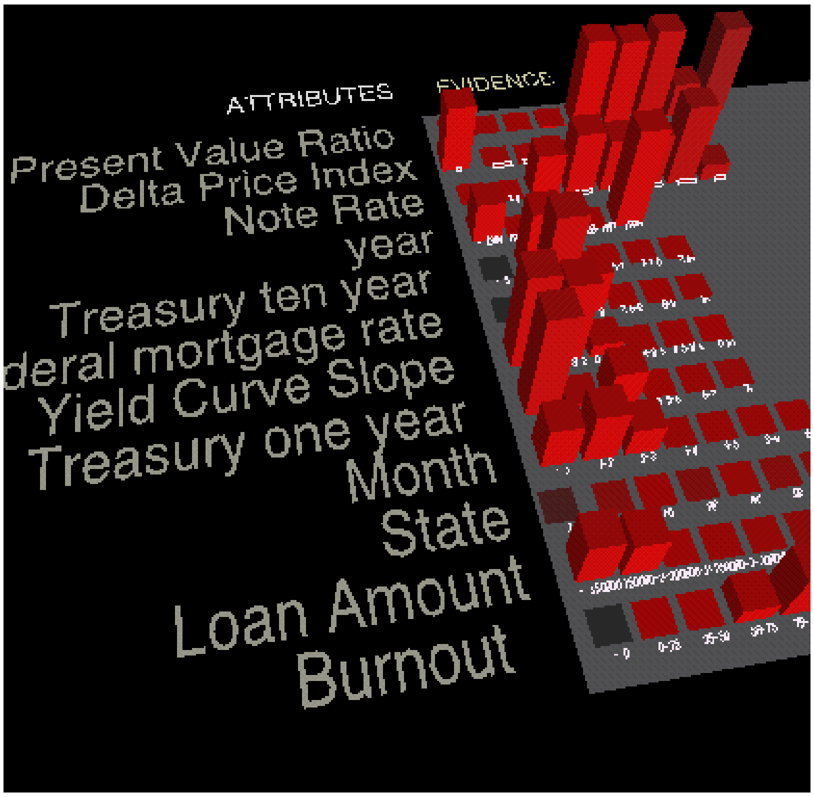

Figure 5: The simple/naive-Bayes model for prepayments.

Building a Simple-Bayes Model

Figure 5 shows a visualization of the Simple/Naive-Bayes model (Domingos & Pazzani 1997; Good 1965; Duda & Hart 1973) for prepayments.

The visualization consists of rows of bars, one for each attribute. The attributes are sorted in order of importance computed as the conditional entropy (Cover & Thomas 1991) of each attribute and the label. The bar height gives evidence that is *additive* as opposed to multiplicative. Formally, the Naive-Bayes model computes the probability of the class as:

$$\mathrm{P}(C_i \mid X) \propto \prod_{j=1}^{n} \mathrm{P}(A_j \mid C_i) \cdot \mathrm{P}(C_i) \ ,$$

where C_i is class i, X is the instance, and A_j is attribute j. Taking logs and looking at a single attribute represented by a bar for class i, the bar height is drawn proportional to

$$-\log\left(1 - \mathrm{P}(A_j \mid C_i)\right)$$

This kind of representation is excellent for characterizing a class of interest. The colors become less saturated (*i.e.*, grayer) if the confidence interval for the estimated evidence is large, signifying that the estimate is based on a small number of instances. The tool is interactive and users can highlight a bar by moving the cursor over it. More details about this visualization are given in Becker, Kohavi, & Sommerfield (1997).

Interpreting Figure 5, we can see that the following factors increase the evidence for prepayments:

1. High present value ratio (over 1.1) with a four-week lag. Loans with present value ratios greater than one are referred to as 'premium' mortgages. This behavior is consistent with previous empirical analysis (Brown 1992; Harmon 1996).
2. High delta price indices (large change in housing prices from origination to current period). The left-most bar represent unknowns, which correlate with recent 1998 data (where refinancing is on the rise). High delta price indices imply an increase in a borrower's home equity value, which can put the owner in a better rate bracket (Peristiani *et al.* 1996).
3. Note rates with rates over 8%. These people are likely to refinance at lower rates. This behavior is consistent with previous empirical analysis (Brown 1992; Harmon 1996).
4. "Current year" (of record) equal to 1994 or 1998. This is not an important factor in this analysis. It merely reflects increased prepayment activity that had occurred in some years due to very low mortgage rates.
5. Low Ten-year Treasury (5-6%), low Federal Mortgage

rates, or low One-Year Treasury (4-5%) (all lagged 4 weeks). Low Ten-year Treasury rate yield high prepayment activity. When One-year Treasury rates are low, prepayment activity is higher. This reflects an increased incentive for thirty-year Fixed Rate Mortgage borrowers to refinance into shorter term fixed rate mortgage instruments (15-year FRMs, 7-year and 5-year Balloons) and adjustable rate mortgages (ARMs). This behavior is consistent with previous empirical analysis (Cunningham & Jr 1990; Harmon 1997).

6. Month in January to March. This may be explained by the 1998 boom in refinancing because we only have the first three months. There was no obvious seasonable pattern, something we expected but did not see.
7. Low values for yield Curve Slope (lagged 4 weeks). A yield curve slope that is smaller (flatter) than the sample average slope results in greater prepayment activity. This probably reflects declines in long term interest rates which stimulate prepayment activity (Litterman, Scheinkman, & Weiss 1991). (Note the results of ten-year Treasury Rate and FHLMC 30-year Mortgage Rate factors examined above.)
8. Small loan amounts (up to $200,000). While we had expect loans with larger original loan balances to prepay at a faster pace, the current evidence indicates an opposite effect. This observation may be due to these loans being more seasoned (older loans) having higher note rates, resulting strong incentive to prepay. Additional data analysis confirmed this conjecture. These were once Jumbo loans but may now be refinanced at a non-Jumbo loan with better rates.
9. High burnout. Theoretically, a proper burnout measure should show reduced prepayment activity as the burnout counter increases reflecting reduced interest rate sensitivities consistent with increased refinance opportunities over time. We find an opposite impact that most probably is attributable to the current burnout counter which should not accumulate if the borrower is prevented from refinancing due to equity and credit constraints (Peristiani *et al.* 1996). Further reasarch and additional data will examine this hypotheses in more detail.
10. Low Current Principal Balance (not shown). Lower principal balances reflect higher prepayment activity.
11. Specific Loan Purposes (not shown). Loans originated as the result of a purchase of a home experienced higher prepayment activity as compared to loans that were originated through refinance activity.
12. Specific Origination Sources (not shown). Correspondent/Co-issued loans and wholesale originated loans showed greater prepayment activity than retail loans or loans purchased through servicing transfers. Brokers are economically incented to originate loans through multiple refinancings. This fact is consistant with the results observed for this variable.

The above analysis confirmed several known factors affecting prepayment but also raised several problems that have prompted us to rethink about definitions (e.g., burnout rates).

Summary

During the KDD process, we discovered several interesting things, including:

1. An oftened unrecognized use of data mining is for data cleansing. We went through several iterations, each time finding problems with the data. This use of data mining requires that the models be comprehensible. Had we used opaque models such as Neural Networks, it is unlikely that we would have found several problems.
2. When the refinancing costs (Figures 3 and 4) were shown to a customer, they observed that the refinance cost in certain states were much higher than others. Here the visualizations were very effective and the customers, whose business is to solicit refinance by targeting a state at a time, shifted priorities to the lower cost states from the higher cost states, thus saving large telemarketing costs of blanketing a large high cost low refinance state.
3. The best lag between events (*e.g.*, treasury rate changes) and prepayment was about four weeks. This is shorter than conventional wisdom of about eight weeks to ten weeks.
4. The data still has many unexplained unknown values. Specifically, the unknown zip codes show very large prepayments, a factor of five higher than any other zip code!
5. In several cases we found spurious correlations that were the result of processes or events in the world. This again highlights the importance of understandability of the models built. For example, it was observed that a large group of loans were paid off during a specific shift in the yield curve. After some analysis and investigation, we found that a servicer had sold off a portion of their portfolio at a specific time of the year. Concidentally the yield curve had a substantial shift at the same time. These two items may be unrelated. MineSet's ability to drill-down and recursively build a model to explain a pattern was very useful.

This study was relatively low-tech in terms of the models used from the spectrum of possible models available to modelers in the MineSet tool or otherwise. The Simple/Naive-Bayes model assumes conditional independence, an assumption which is known to be false. However, our purpose (as in many other pilot studies the authors were involved in) was not to create the most accurate model; in fact, we never even looked at the estimated model accuracy. Our purpose was to get insight about the data, factors, and behaviors that will help define what other attributes might be needed for the next stage.

We intend to continue the study and use more complex models, but had we started with those without noticing the data quality and issues that were seen during the simpler, we would have missed key issues.

One factor that was crucial in deriving insight quickly is having a data mining environment that supports modeling, drill-downs, drill-through, and integration of different tools and visualization. Users need more than a single algorithm to effectively mine data.

References

Becker, B.; Kohavi, R.; and Sommerfield, D. 1997. Visualizing the simple bayesian classifier. In *KDD Workshop on Issues in the Integration of Data Mining and Data Visualization*.

Brachman, R. J., and Anand, T. 1996. The process of knowledge discovery in databases. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press and the MIT Press. chapter 2, 37–57.

Brown, Hayre, L. 1992. Analysis of mortgage servicing portfolios. *Journal of Fixed Income* 60–75.

Brunk, C.; Kelly, J.; and Kohavi, R. 1997. MineSet: an integrated system for data mining. In Heckerman, D.; Mannila, H.; Pregibon, D.; and Uthurusamy, R., eds., *Proceedings of the third international conference on Knowledge Discovery and Data Mining*, 135–138. AAAI Press.
`http://www.sgi.com/Products/software/MineSet`.

Cover, T. M., and Thomas, J. A. 1991. *Elements of Information Theory*. John Wiley & Sons.

Cunningham, D. F., and Jr, C. C. 1990. The relative termination experience of adjustable to fixed rate mortgages. *Journal of Finance* XLV(5).

Domingos, P., and Pazzani, M. 1997. Beyond independence: Conditions for the optimality of the simple Bayesian classifier. *Machine Learning* 29(2/3):103–130.

Duda, R., and Hart, P. 1973. *Pattern Classification and Scene Analysis*. Wiley.

Fabozzi, F. J. 1995. *The Handbook of Mortgage Backed Securities*. Probus Publishing Co., 4th edition.

Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*. AAAI Press and the MIT Press. chapter 1, 1–34.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. M.I.T. Press.

Harmon, R. 1996. Risk monitors prepayment models. Technical report, Risk Monitors.

Harmon, R. 1997. The gnma arm prepayment model. Technical report, Risk Monitors.

Litterman, R.; Scheinkman, J.; and Weiss, L. 1991. Volatility and the yield curve. *Journal of Fixed Income* 49–53.

Peristiani, S.; Bennett, P.; Monsen, G.; Peach, R.; and Raiff, J. 1996. Effects of household creditworthiness on mortgage refinancings. Technical Report 9622, Federal Reserve Bank of New York.

Richard, S. F., and Roll, R. 1989. Prepayments on fixed-rate mortgage-backed securities. *The Journal of Portfolio Management* 73–82.