International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Review Article

Securitization of Mortgage Backed Securities using Data Mining Techniques

Viraj Gada†*, Sumeet Deshpande‡ and Apoorva Dhakras†

†Computer Engineering, Mumbai University, Dadar, Mumbai, India
‡Computer Engineering, Mumbai University, Shivaji Park, Mumbai, India
†Computer Engineering, Mumbai University, Andheri, Mumbai, India

Accepted 10 July 2015, Available online 17 July 2015, **Vol.5, No.4 (Aug 2015)**

Abstract

This paper describes different data mining techniques used in securitization of Mortgage Backed Securities. Mortgage Default Risk Analysis is used in many financial institutes for accurate analysis of consumer data to find defaulter and valid customer. For this different data mining techniques can be used. The information thus obtained can be used for Decision making. In this paper we study about Mortgage Default Risk Analysis, factors considered before sanctioning a mortgage and different data mining techniques like C4.5 algorithm, Support Vector Machines (SVM), Apriori algorithm, Adaboost algorithm, Naïve Bayes Classifier and K-means algorithm.

Keywords: *Data Mining, Mortgage Default Risk Analysis, C4.5, Support Vector Machine, Apriori, Adaboost, Naïve Bayes, K-means algorithm*

1. Introduction

[1]The credit crisis of 2008 in USA was a wake-up call for bankers and real estate sectors to stop giving loans in the form of sub-prime mortgages (SPM). As more and more debtors defaulted on their payments, the banks were unable to retrieve their capital as property prices plummeted. This led to a full-blown crisis with banks going bankrupt due to high supply of properties but low demand. This situation had a cascading effect on the investment banks, investors, lenders and hedge fund managers who subsequently went bankrupt as well. In order to avoid such a situation again, it is essential for the banks to be sure that the person with whom it mortgages is financially sound and possesses the ability to repay credit without defaulting. Five main factors are taken into account by banks before granting a mortgage. They are down payment, job history, paper work, debt load and credit score. The bank sanctions a mortgage only after carefully scrutinizing these five factors. Traditionally, this assessment was done using statistical and mathematical methods by analysts. However, over time, data mining techniques have emerged as a potential solution due to their knowledge discovery process and their ability to transform them into useful information.

2. Mortgage Default Risk Analysis

Mortgage Backed Securities (MBS) are types of bonds representing an investment in a pool of real estate loans. Whenever a person lets say A wants to buy a real estate property (like house, farm,etc) but lacks the immediate capital to pay for the property, he goes to the bank and applies for a loan of a certain amount. The bank sanctions the loan against the mortgage of one of A's properties. This property acts as a security in case A defaults on his payments. Also, A has to pay the bank a principal amount along with some interest annually to cover the loan amount. In traditional times, bank would simply keep the records in its loan books and receive principal and interest for the duration of the loan till the entire amount was paid.

This recovery of loan lasted for around 5-10 years thus tying up bank's capital and resources. So, the banks decided to sell the stream of interest and principal payment of A's loan to investors to get it off bank's books and free-up capital. At the same time, bank would make money by simply servicing and offering Mortgage Home Loans and associated fees. To sell to investors, bank handles hundreds and thousands of such Mortgage Home Loans (MHL). This bundle of MHL is sold to the investment banks in the form of a single bond called Collateral Debt Obligation(CDO). The investment bank then partitions CDO according to quality (usually safe, ok and risky) and sells it to other investors. So, although A makes a payment to the bank, the loan is essentially in hands of the investors. Thus Mortgage Backed Securities are essentially a way for banks to free up their capital and provide a way for investors to buy off their mortgages. Seeing the potential profits this system could make, the investors

*Corresponding author: **Viraj Gada**

want more CDO slices and so they call up the lender for the want of more mortgages. The lender in turn calls the broker for more home owners. But the broker cannot find any appropriate home owners who qualify for mortgage as he already has those. So the investors think that if home owners default on their payments, then they can sell off their houses to earn more profits since property prices are always increasing. Thus lenders start violating the five necessary conditions to be fulfilled (down payment, job history, paper work, debt load and credit score) and start taking risks by providing houses to owners without down payment, proof of income or necessary documents. So, instead of selling houses to responsible home owners called prime mortgages, the lenders started selling them of to less responsible home owners called Sub-Prime Mortgages (SPM). So, the whole process of home owner -> broker -> lender -> investment banks -> investor continues as each make profits from commissions obtained by selling MBS. No one was worried as they were selling of their risks to next person in the chain and if home owners defaulted they would just have to sell the house and recover their losses. However, this process had a catastrophic effect when many home owners started defaulting (primarily due to the greed of investors who encourages lenders to give properties to SPM. As there was a sudden spurt of properties on sale, the property prices went spiraling down as there was very less demand. Seeing the prices of property going down, the financially sound home owners also refuse to pay to lenders as the prices of the property which they bought in the past years for a high cost has decreased drastically and that they see no point in paying the high installments year on year. Thus, this leads to a capital crunch with banks going bankrupt. And as the cascading effect continues, the lenders, brokers, investment banks and hedge fund managers also go bankrupt. Thus, to avoid this situation it become very essential to separate the 'good credit' class from the 'bad credit' class. By using Data Mining Techniques on the vast information contained in the databases of banks, it is possible to segregate the 'good credit class' people from the 'bad credit class' people. Firstly, we take into consideration the 5 main factors: down payment, job history, paper work, debt load and credit score and allot them weightages. Now using data mining techniques, we can fill the information for these five criteria and accordingly get the information of the necessary minimum cut-off's of each category seen in the past as a marker for allotting mortgage to home owners. If a home owner does not satisfy a majority of these criteria, then his mortgage is not sanctioned as he falls in the 'bad credit class' and is more likely to default on his payments.

The information on the five factors for credit scoring and segregation is given below:

- *Down Payment:* The down payment is the capital paid upfront to buy something on credit. The down payment rates vary from country to country and is usually around 20% of the total cost.

- *Job History:* The banks want to make sure that you are can be steadily able to pay your installments in the future. Hence, they scrutinize your job history carefully to track your employment. This means that a period of unemployment or constant changing of jobs can disqualify you from seeking a mortgage.

- *Paper Work:* After the 2008 credit crisis, proper paperwork has become a norm in order for the lenders to know your financial background thoroughly before sanctioning a mortgage.

- *Debt Load:* Before sanctioning a mortgage, it is also essential for the lenders to know that you are handling your other debts(bills, car loans, etc) so as to be assured you won't default on anything and be able to pay off everything.

- *Credit Score:* Borrowers generally need a credit score of 640 to qualify for a conventional mortgage as per many broker institutions. Hence, if the credit score is low, it can severely hinder your chances of getting a mortgage although a low credit score due to emergency or medical reasons can be taken into consideration.

3. Data Mining Techniques

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics and database systems.The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining involves six common classes of tasks:

- *Anomaly detection (Outlier/change/deviation detection)* – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- *Association rule learning (Dependency modelling)* – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- *Clustering* – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- *Classification* – is the task of generalizing known structure to apply to new data. For example, an e-

mail program might attempt to classify an e-mail as "legitimate" or as "spam".

- *Regression* – attempts to find a function which models the data with the least error.
- *Summarization* – providing a more compact representation of the data set, including visualization and report generation.

4. C4.5 Algorithm

Classification algorithms particularly suit the needs of the mortgage default risk analysis. By classifying the data input into the 5 factors by using C4.5 algorithm, we can obtain sufficient information to segregate the 'good creditors' from the bad 'creditors'. Amongst the classification algorithms, C4.5 ranks as one of the best data mining algorithms for several reasons. This tree-induction algorithm not only trace back the machine learning results obtained from ID3 but also provide greater classification accuracy. This algorithm is also the fastest amongst the compared main memory algorithms for machine learning and data mining. This system takes as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs

C4.5 contains mechanisms for proposing three types of tests:

- The "standard" test on a discrete attribute, with one outcome and branch for each possible value of that attribute.
- If attribute Y has continuous numeric values, a binary test with outcomes Y£Z and Y>Z could be defined, based on comparing the value of attribute against a threshold value Z.
- A more complex test based also on a discrete attribute, in which the possible values are allocated to a variable number of groups with one outcome and branch for each group.

C4.5 Algorithm

Let the classes be denoted $\{C_1, C_2, \ldots, C_k\}$. There are three possibilities for the content of the set of training samples T in the given node of decision tree:

- T contains one or more samples, all belonging to a single class C_j. The decision tree for T is a leaf identifying class C_j.

- 2. T contains no samples. The decision tree is again a leaf, but the class to be associated with the leaf must be determined from information other than T, such as the overall majority class in T. C4.5 algorithm uses as a criterion the most frequent class at the parent of the given node.

- T contains samples that belong to a mixture of classes.In this situation, the idea is to refine T into subsets of samples that are heading towards single-class collections of samples. An appropriate test is chosen, based on single attribute, that has one or more mutually exclusive outcomes $\{O_1, O_2, \ldots, O_n\}$:

- 4.T is partitioned into subsets $T_1, T_2, \ldots, T_n$ where T_i contains all the samples in T that have outcome O_i of the chosen test. The decision tree for T consists of a decision node identifying the test and one branch for each possible outcome.

Test – entropy

If *S* is any set of samples, let *freq* (C_i, *S*) stand for the number of samples in S that belong to class C_i (out of *k* possible classes), and ½*S*½ denotes the number of samples in the set *S*. Then the entropy of the set *S*:

For I = 1 to k

$$\text{Info}(S) = -\sum \left((\text{freq}(C_i, S)/ \tfrac{1}{2}S\tfrac{1}{2}) \times \log_2 (\text{freq}(C_i, S)/ \tfrac{1}{2}S\tfrac{1}{2})\right)$$

After set *T* has been partitioned in accordance with n outcomes of one attribute test *X*:

$$\text{Info}_x(T) = \sum ((\tfrac{1}{2}Ti\tfrac{1}{2}/ \tfrac{1}{2}T\tfrac{1}{2}) \times \text{Info}(Ti))$$

When i=1 to n

$$\text{Gain}(X) = \text{Info}(T) - \text{Info}_x(T)$$

Criterion: select an attribute with the highest Gain value.

5. Support Vector Machines (SVM)

Sometimes, in the absence of a considerable amount of data, it is necessary to make decision of granting mortgage based on a handful of cases. This scenario takes place in new banks yet to register data for mortgages and also in developing economies where such data maybe hard to obtain. In such a scenario, the Support Vector Machine algorithm can be of great use. SVM is one of the most full-proof and accurate among well-known algorithms. Also, it is insensitive to the number of dimensions. Apart from mortgage default risk analysis, this algorithm is currently being used in many fields of application like bioinformatics, text, image recognition, etc.

- *SVM Algorithm*

- *Choose a kernel function*

For Linear Function

$k(X_i, X_j) = X_i * X_j$

For Polynomial Function

$k(Xi, Xj) = (\Upsilon Xi * Xj + C)^d$

For RBF function:

$k(X_i, X_j) = exp(-\Upsilon|X_i - X_j|^2)$

For Signoid Function:

$k(X_i, X_j) = tanh(\Upsilon X_i * X_j + C)$

where $k(X_i, X_j) = \phi(X_i) * \phi(X_j)$

that is, the kernel function, represents a dot product of input data points mapped into the higher dimensional feature space by transformation .

Gamma is an adjustable parameter of certain kernel functions.

The RBF is by far the most popular choice of kernel types used in Support Vector Machines. This is mainly because of their localized and finite responses across the entire range of the real x-axis.

Choose a value for C

Classification SVM Type 1

For this type of SVM, training involves the minimization of the error function:

$0.5w^Tw + C * \zeta_i$ *where i =1 to n*

subject to the constraints:

$y_i(w^t \Phi(x_i) + b) \geq 1 - \zeta_i$ *and* $\zeta_i \geq 0$ *, i= 1,...,N*

where C is the capacity constant, w is the vector of coefficients, b is a constant, and ζ_i represents parameters for handling nonseparable data (inputs). The index i labels the N training cases. Note that y belongs to +1 or -1 represents the class labels and x_i represents the independent variables. The kernel Φ is used to transform data from the input (independent) to the feature space. It should be noted that the larger the C, the more the error is penalized. Thus, C should be chosen with care to avoid over fitting.

Classification SVM Type 2:

In contrast to Classification SVM Type 1, the Classification SVM Type 2 model minimizes the error function:

$0.5w^Tw - v\rho + 1/N(\zeta_i)$ *where i=1 to n*

subject to the constraints

$y_i(w^t \Phi(x_i) + b) \geq \rho - \zeta_i$ *and* $\zeta_i \geq 0$ *i= 1,...,N and* $\rho \geq 0$

The quadratic equations can be derived from the corresponding factors taken into consideration and the classification is presented accordingly.

Construct the discriminant function from the support vectors **:**

The classifier is said to assign a feature vector *x* to class w_i if

gi(x) > gj(x) for all j is not equal to i

For 2 category cases:

g(x)= g1(x)-g2(x)

Decide w1 if g(x) > 0 ; else decide w2

6. Apriori Algorithm

To find frequent itemsets from a transaction dataset and derive association rules would help us a lot in compiling the data by maintaining similar data together. Also, finding frequent itemsets (itemsets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. In this regard when we have a data in which majority of the data is similar, the mortgage determing algorithm which can be used is Apriori algorithm.

The Apriori algorithm is an influential algorithm for mining frequent item sets of boolean association rules. It uses a 'bottom up' approach where frequent subsets are extended one item at a time. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

- *Algorithm*

1. Scan the transaction basket to get support of the search one item set., compare S with min_sup and get a support of 1-item seta, L1.

2. Use Lk-1 join Lk-1 to generate a set of candidate k-item sets. And use apriori property to prune the unfrequented k-itensets from this set.

3. Scan the transaction database to get the support S of each candidate k-itemset in the find set , compare S with min_sup and get a set of frequent k-itemsets Lk.

4. Compare if set = Null
5. If NO, then go to step 2.

6. If YES, then go to step 7.

7. For each frequent item set 1, generate all nonempty subsets of 1.

8. For every non-empty subsets of 1, output the rule "s=>(1-s)" if confidence C of the rule "s=>(1-s)" (support s of 1/support S of s) min_count.

7. Adaboost Algorithm (Adaptive Boosting):

Boosting refers to a general and provably effective method of producing a very accurate classifier by combining rough and moderately inaccurate rules of thumb.It is based on the observation that finding many rough rules of thumb can be a lot easier than finding a single, highly accurate classifier.To begin, we define an algorithm for finding the rules of thumb, which we call a weak learner. The boosting algorithm repeatedly calls this weak learner, each time feeding it a different distribution over the training data (in AdaBoost). Each call generates a weak classifier and we must combine all of these into a single classifier that, hopefully, is much more accurate than any one of the rules. Thus by applying these rules into the mortgage default risk analysis, we can classify bad creditors from good creators and can thus sanction loans accordingly.

- *Algorithm*

0) Set W˜ (0) i = 1/n for i = 1,...,n

1) At the mth iteration we find (any) classifier h(x; ˆθm) for which the weighted classification error "m "m = 0.5 − 1 2 !" n i=1 W˜ (m−1) i yih(xi; ˆθm) # is better than chance.

2) The new component is assigned votes based on its error: αˆm = 0.5 log((1 − "m)/"m)

3) The weights are updated according to (Zm is chosen so that the new weights W˜ (m) i sum to one): W˜ (m) i = 1 Zm · W˜ (m−1) i · exp{ −yiαˆmh(xi; ˆθm) }

8. Naive Bayes Classifier

The Naive Bayes(NB) Classifier is based on Bayes' Theorem and which is based on independent assumptions between predictors.The model can be particularly used for large data sets and hence it is used for data analysis in finance related applications.It outperforms the more sophisticated classification methods.

The NB classifier uses class conditional independence which assumes that the effect of a predictor(X) on a given class(c) is independent of the values of other predictors.

P(c|x) = ((x|c) * P(c))/ P(x)

P(c|x) = Posterior probability
P(x|c) = Likelihood
P(c) = Class prior probability
P(x)= Prediction prior probability

P(c|x) = P(x1|c) * P(x2|c) * P(x3|c) *--* P(xn|c)

The posterior probability can be calculated first, constructing a frequency table for each attribute against the target.Then the frequency tables are converted to likelihood tables and finally NB equation is used to calculate posterior probability.The class with the highest posterior probability is the final outcome of prediction.

Numerical Predictors
Numerical variables need to be transformed to their categorical counterparts (binning) before constructing their frequency tables. The other option we have is using the distribution of the numerical variable to have a good guess of the frequency. For example, one common practice is to assume normal distributions for numerical variables.
The probability density function for the normal distribution is defined by two parameters (mean and standard deviation).
Mean: $\mu = \frac{1}{n} \sum_{i=1}^{n} xi$
Standard Deviation: $\sigma = [\frac{1}{n-1} \sum_{i=1}^{n} (xi - \mu)^2]^{0.5}$
Normal Distribution: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

K-Means Clustering

K-means is a clustering method that aims to find the positions μi,i=1...k of the clusters that minimize the *distance* from the data points to the cluster. K-means clustering solves

$$\text{argmin}c\sum_{i=1}^{k}\sum_{x\in ci} d(x,\mu i) = \text{argmin}c\sum_{i=1}^{k}\sum_{x\in ci} \|x-\mu i\|_2^2$$

where ci is the set of points that belong to cluster i. The K-means clustering uses the square of the Euclidean distance $d(x,\mu i)=\|x-\mu i\|_2^2$.

The Lloyd's algorithm, mostly known as k-means algorithm, is used to solve the k-means clustering problem and works as follows. First, decide the number of clusters k. Then:

1. Initialize the center of the clusters	μi= some value ,i=1,...,k
2. Attribute the closest cluster to each data point	ci={j:d(xj,μi)≤d(xj,μl),l≠i,j=1,...,n}
3. Set the position of each cluster to the mean of all data points belonging to that cluster	μi=1|ci|∑j∈cixj,∀i
4. Repeat steps 2-3 until convergence	
Notation	|c|= number of elements in c

Table of Comparison

Algorithm	C4.5	Svm	Apriori	Adaboost	Bayes	K-means
Type	Regression	Regression	Association rule	Classification	Classification	Clustering
Accuracy	High	High	Low	High	May vary	May vary
Complexity	High	Fair	Easy	Easy	Easy	Easy
Drawback	Cpu time & Memory	No Transparency	Low performance, requires large data	Sub-optimal results errors prone	Dependencies, loss accuracy	Not used for small data sets
Advantage	Deals with noise, builds model that can be interrupted	Robust, Exact	Easy, Simple	Simple, Precise, Strong, Abstract Base	Easy, No iteration	Easy, Simple Scalable

The algorithm eventually converges to a point, although it is not necessarily the minimum of the sum of squares. That is because the problem is non-convex and the algorithm is just a heuristic, converging to a local minimum. The algorithm stops when the assignments do not change from one iteration to the next.

Conclusion

In this paper, we assess and predict the best data mining algorithm for reduction in mortgage default risk. C4.5 algorithms are preferred by banks for their precise results but the cost in time and space required to make the decision trees nullifies its advantages . The discrimination made by support vector machine with limited data is accurate and the people can understand its working easily. This enables the banks and other financial institutions to provide an account for accepting or rejecting an applicant. AdaBoosting has already increased the efficiency of classification but gives sub-optimal solution and is largely error-prone thus making it unreliable. Apriori algorithm while being easy and simple is low on performance and requires a large data set to increase efficiency which may not always be possible. Naive-Bayes algorithm while being easy and non-iterative is prone to inaccuracy and dependencies loss. K-means algorithm while requiring large data set is simple, easily implementable and gives precise results. Thus taking into account all factors into the decision-making process, it can be predicted that SVM and K-means algorithm provide the best possible results for segregating 'good credit' from 'bad credit'. Thus, they will help in reducing the mortgage default risk for banks and ultimately avert a potential credit crisis.

References

Boris Kovalerchuk, Evgenii Vityaev (2002), Data mining for financial applications

Wu,V.Kumar,J.Ross, Quinlan,J.Ghosh,Q.Yang,H.Motod (2007), Top 10 algorithms in data minin

Jiawei Han, Micheline Kamber,Data Mining Concepts and Technique, 2nd edition

Salvatore Ruggieri,Efficient C4.5

S.V.N. Vishwanathan, M. Narasimha Murty,SSVM:A simple Svm algorithm

http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

http://math.mit.edu/~rothvoss/18.304.3PM/Presentations/1-Eric-Boosting304FinalRpdf.pdf

http://www.onmyphd.com/?p=k-means.clustering

http://wikipedia.com

http://www.saedsayad.com/naive_bayesian.htm