

---

# A study of machine learning techniques to detect Mortgage Fraud

---

**Chuping Liu**  
**Rizwan Khan**  
**Rajiv Prithvi**  
**Sean Little**

University of Southern California, CA 90089, USA

CHUPINGL@USC.EDU  
RIZWANKH@USC.EDU  
RNC@USC.EDU  
SEANLITT@GMAIL.COM

## Abstract

Preventing Mortgage fraud and loan defaulting has been a major challenge of mortgage based financial institutions. The vulnerability and seriousness of the issue was amplified by the recent sub-prime market crisis. We wanted to know if there was a way using Machine Learning approaches to prevent or radically reduce fraud cases dogging the industry. We utilized the datasets acquired from a mortgage-based financial institution to experiment with machine learning algorithms using different tools to find out if it were possible to detect fraud instances. We used both supervised and unsupervised approaches. We conclude that with expert input and intelligent selection of attributes, it is possible to predict potential fraud with high accuracy.

## 1. Introduction

The purpose of this paper is to describe the methodology and findings of an attempt to de-

---

tect patterns of mortgage fraud. The data for this project was provided by IndyMac Bank (IMB) to train a model that could be generally applied to mortgage data. This model could be used to process large numbers of loan applications, and flag potential cases of fraud without the need for a specific audit investigation.

## 2. Method

### 2.1. Data Description and Preprocessing

IMB provided us with 43273 samples of loan data. The data for each loan contained attributes such as application date, loan status, loan amount, etc. We were also given a list of audit results. Of the 43273 loans there were about 1935 loans that were audited. Our first task in building our fraud model was to convert the IMB loan data into a format the modeling suite (WEKA) could utilize. This involved associating the audit data with the respective loan data, removing nonsensical and irrelevant loan information, and generating a *classification attribute*. This attribute is a binary flag used to indicate whether the associated loan was estimated to be fraudulent or not.

The classification attribute was generated using some of the audit data. There were different types of audits, and 11 different types of possible problems the audit could potentially dis-

cover. Some of these problems were more significant than others. For example an audit investigation for one loan might determine that the applicant did not fill out their documentation properly. And an applicant for another loan might have fabricated his or her credit score or employment information. An IMB representative gave us a list of specific findings that are most likely related to fraud. Those findings included problems with assets, collateral, credit, employment/income, and housing occupancy.

It is important to emphasize that we were never completely sure that the information from a particular loan was actually a case of mortgage fraud. We therefore modeled fraud according to an estimate, or potential for fraud. We built this estimate by using the logical OR of the audit findings we were told were significant. If none of those problems were found, we assumed that the loan was found to be free of fraud. Note that for further work, it would be very helpful if we were given access to audit data where no problems were found. For this project we think that IMB assumed that this data would not be interesting, and therefore did not provide us with these results.

Once we had a classification attribute, we had enough information to generate an ARFF file, the file format used by WEKA. Unfortunately there were still problems in the data that required further preprocessing. For our model to generalize acceptably, we needed to be completely certain that the patterns we were training against were real. But many of the audited loans were rejected because of the audit findings. This resulted in loan data that contained *artificial* and *distracting* patterns that if not removed would completely dominate the model and destroy the capacity of our results to generalize. To illustrate, of the 1935 loans that were audited, 512 of those were not funded. Presumably these loans were denied because of the adverse audit findings. Nearly all of those denied loans had several loan attributes that were completely out of family when compared with

a normal, funded loan. The loan term was set to zero, the loan product was set to NULL, and the repayment flags were all NULL.

We propose 3 ideas for removing these distracting patterns:

- Completely remove attributes that are too tainted by artificial patterns.
- Attempt to add new attributes that will hopefully provide more insight into fraud.
- Training should be applied only to loans that have a known result, i.e. the audited data. Then the models that perform well against the audited data can be applied to loans that were not audited.

Info. Gain	Index	Name
0.0427733	30	entry_type <sup>1</sup>
0.0420738	21	loan_product_category
0.0399902	28	ever90at12
0.0398064	27	ever60at6
0.0394483	22	creditlevel_final
0.0391543	25	ever30at3
0.0390755	22	ever30at2
0.0390562	23	ever30at1
0.0389114	26	ever30at4
0.0385689	4	loan_status <sup>1</sup>
0.0378609	20	funded
0.0361001	8	loan_term
0.0322413	1	application_date <sup>2</sup>
0.0034037	11	cltv
0.0028694	29	current_lp_dqcode

Table 1. Information Gain Values

<sup>1</sup>These attributes were removed because they contained what seemed to be artificial patterns. For example the entry type field contained a *qc\_fail* state. Assuming this meant quality control fail, we didn't want to assume that all loans would be subject to quality control analysis. These assumptions should be further discussed and verified with IMB.

<sup>2</sup>We kept the application date attribute for some analysis rather than removing it completely because according to IMB, the types of loans granted changed dramatically when the mortgage indus-

We used an information probability ranking to verify that the top performing attributes were providing valid fraud insight. See the rankings on the *Table 1*. Entries are by attribute and sorted according to the most salient first. The attributes indexed [30,4,20,1] were completely removed from the data before the model was built, for reasons that will be described shortly. The attributes indexed [21,28,27,25,24,23,26,8] indicate that were significantly different when the loan was not funded.

## 2.2. Software Tools

We used Weka (Waikato Environment for Knowledge Analysis) software tool to perform all our operations on our dataset. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms were applied directly to a dataset. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. We used both supervised and unsupervised techniques incorporated within Weka as part of project. For SVMs we used libSVM which like Weka is publicly available tool for Machine Learning.

## 2.3. Algorithms

- *Logistic regression* is a discriminative learning algorithm which directly attempts to estimate the probability of fraud given the attribute data for a company-quarter. Our examination included several hypothesis classes of increasing complexity within the setting of logistic regression.
- *Naive Bayes* is a generative method that determines a model of  $p(x|y)$  and then uses Bayes rule to generate  $p(y|x)$  rather than fitting parameters to a model of  $p(y|x)$  di-

---

try began having problems during min 2007. We wanted to preserve the legitimate patterns that might be present if fraud was more likely to occur during certain time frame

rectly. Consequently, this approach can be expected to work well if the conditional distributions of the attributes are significantly different for different values of the label. Intuitively it is reasonable to expect that a borrower who has decided to act fraudulently would behave in a manner similar to other borrowers who have made the same decision, and in a manner inconsistent with those borrowers not acting fraudulently.

- *J48 decision tree* represents a supervised approach to classification. A decision tree is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes. J.R. Quinlan has popularized the decision tree approach with his research. The implementation goes by the name C4.5. We used the weka classifier package has its own version of C4.5 known as J48.
- *Bayesian Net* is a directed graph, together with an associated set of probability tables. The graph consists of nodes and arcs. The nodes represent variables, which can be discrete or continuous. The edges of the graph represent causal/influential relationships between variables. The key feature of BBNs is that they enable us to model and reason about uncertainty.
- *K-nearest neighbor algorithm (k-NN)* is a method for classifying objects based on closest training examples in the feature space. k-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The IBK classifier in weka works as nearest neighbor algorithm in weka.
- *SVM (Support Vector Machines)* is a very useful technique for data classification. It separates the data by a hyper-plane such that the margin width between the hyper-plane and the examples is maximized. Sta-

tistical learning theory shows that maximizing the margin width reduces the complexity of the model, consequently reducing the expected general risk for error.

- *Unsupervised Learning* is used to correlate the results with supervised learning and aims for discover odd or interesting patterns. Three tools provided in WEKA (Witten and Frank, 2005) were explored to study the fraud detection. They are clustering, rule association, and feature selection. Rule association represented the data in an *if Then* format. It was the most efficient way to extract information, if it works. Feature selection helps to discover the information contribution of the attributes such that relatively smaller set of attributes may be zoomed in to detect fraud. Clustering was the most used method in this project since it can be an important approach to factor out interesting customer group in fraud. Three clustering methods were explored and compared: fartheseFirst, simpleKmeans, and makeDensityBasedClusters.

### 3. Results and Discussion

#### 3.1. Unsupervised Learning

The fraud detection problem was studied from unsupervised learning perspective. The results shows in *Figure 1* may be summarized as follows:

- Farthest First was the best clustering method that achieve highest correctness in clusters to classes evaluation. The reason may be that farthestFirst method started the clustering by finding odd instances.
- A particular customer group with 89.3% tendency to be fraud was detected, although this customer group only represents about 3.9% of all the fraud cases.

- Unsupervised learning did not yield as high as fraud detection percentage as that of supervised learning (about 15% deficit). Its performance sensitivity to the fraud percentage in the training dataset was opposite to that of the supervised learning. This may be due to the reason that unsupervised learning aims to capture patterns rather than training a discrimination function on the data.
- Feature reduction and rule association seems not efficient or appropriate in the fraud detection in this project.
- Unsupervised learning has helped to clean up *artificial noise* data in the data pre-processing stage.

#### 3.2. Supervised Learning: Logistic, J48, BN and KNN

##### 3.2.1. 30% NO-FRAUD AND 70% FRAUD

- *With date attribute:* The date attribute was analyzed to have the highest information gain of the all the attributes. We ran the experiments on the dataset with the date attribute intact which gave us some key insights as to when the loans given out were risky and time intervals between loan payments were all insightful. The best performing algorithm for this dataset was IBK=25. The Logistic regression algorithms had accuracies comparable to IBK=25. The accuracies drop when we use cross validation compared to the splits. For this reason we believe that the performance of the algorithms is amore accurately represented when we use cross validation as compared to splits. Accuracies that we get by using 10-fold cross-validation is probably the best measure of the algorithm.
- *Without date attribute:* As mentioned earlier the data application had significant

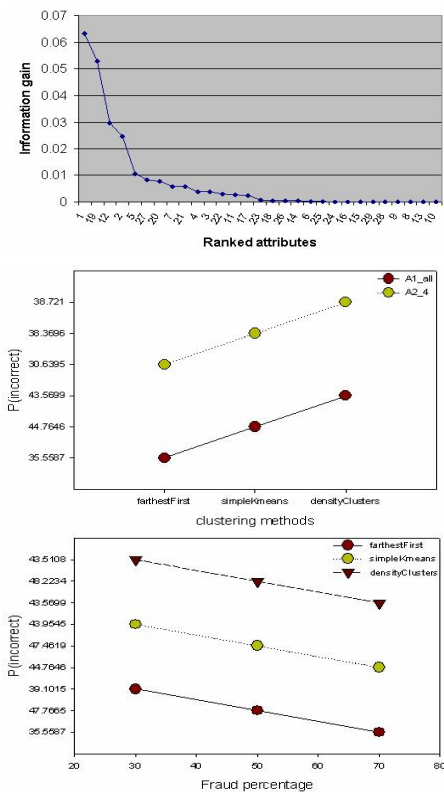


Figure 1. Feature Selection, Clustering on Attribute and Fraud Sets.

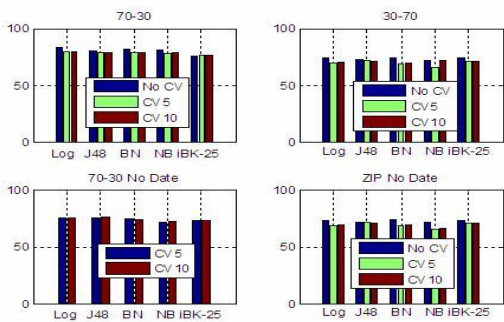


Figure 2. Results without bagging and boosting with 5 and 10-fold cross-validation.

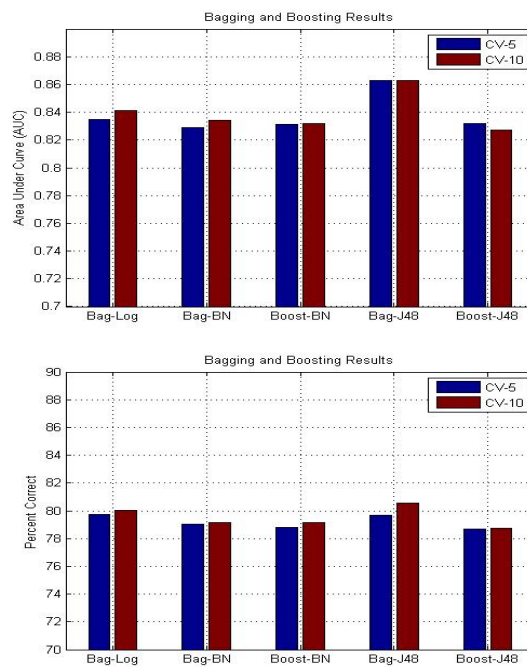


Figure 3. Results with bagging and boosting with 5 and 10-fold cross-validation.

strength as an attribute, influencing performance of algorithms which was good for training. From practical point of view, we understand that banks and other mortgage based institutions would be hard pressed to make business based on dates, so we decided that we should extend the algorithm comparisons to datasets without the date attribute. There was a general increase in performance of algorithms especially J48 decision tree, which performed better on removal of the date attribute. J48 and IBK=25 algorithms were the best performing algorithms for this dataset.

- *Zip ratios:* We wanted to look beyond the datasets given to see if any of the web information about mortgage gave us any evidence or interesting information that could be used with existing data attributes to pick out any interesting pat-

tern on the datasets that would increase the accuracies on the algorithms. We scraped census information on the website [www.brainyzip.com](http://www.brainyzip.com) and used the ratio of loan given versus median mortgage property value to see if there were any anomalies on loan sanctioned to borrowers. This was based on the assumption that the loan sanctioned would be directly proportional to the property value as claimed by the borrowers. This ratio would indicate any possible red flags like false documentation of the property worth or deliberate overestimation of the property value pledged for loans. The census information was gathered using the borrower's zip code which was part of the dataset given to us. The best performing algorithm was IBK=25 and J48 decision tree had comparable accuracies with respect to IBK=25. Again we felt the accuracies on cross validation was better indicator of how the algorithms would fare against new unknown datasets. One interesting observation that we saw was the improvement in accuracy of J48 algorithm on this dataset as compared to accuracies on the previous dataset with the date attribute intact.

### 3.2.2. 70% NO-FRAUD AND 30% FRAUD

- *With date attribute:* On comparison with previous dataset with 70 percent fraud, the 30 percent fraud dataset had significant rise in accuracies in all our algorithms. We were not able to reason out why exactly there would be significant changes by changing fraud content (percentage) of the datasets. The accuracy changes on different fraud percent datasets requires further research and experimentation. It would be fantastic to know if there was an optimal way to arrange the datasets for best performances of algorithms. Since all algorithms had positive changes on change of dataset we concluded that it's not algorithm dependent but data dependent. The

logistic classifier was the best performer for this dataset. The Bayesian Nets and Naïve Bayes had comparable accuracies relative to logistic classifier, with Bayesian nets algorithm performing better.

- *Without date attribute:* By this stage of experimentation it had dawned to us that the 90% split of the dataset as training set and remaining as test set somehow had less reliable performance wise compared to cross validation performance of the dataset. We chose to spend more time with cross validation of the dataset. Logistic regression worked best for this dataset. J48 was second best performing algorithm in this case as compared to the previous dataset where Bayes Nets performed better than J48 in terms of accuracies but Bayesian nets still worked better AUC wise.

### 3.3. Supervised Learning: SVM

There are three criteria that we use for judging the performance of the different SVMs. They are as follows:

- *Number of Support Vectors:* As we can see from *Table 2* that the number of support vectors does effect the accuracy of the SVM. The more the support vectors are the better the accuracy is but it could also suggest over-fitting and thats why we have used cross-validation to reduce the over-fitting.
- *ROC (Receiver Operating Characteristic) Curve and AUC (Area Under the Curve):* As obvious from *Table 3* and *Figure 4* and *5* we can see that SVMs perform the best for the dataset with 70% No-fraud and 30% Fraud. The AUC gives an indication of the accuracy of the SVM.
- *Feature Selection:* We do a comaparison of the features that an SVM comes up with for classification. If two classifiers come up

with the same important features that improves the reliability of the classifiers. *Figure 5* gives an example of selected features for a particular SVM.

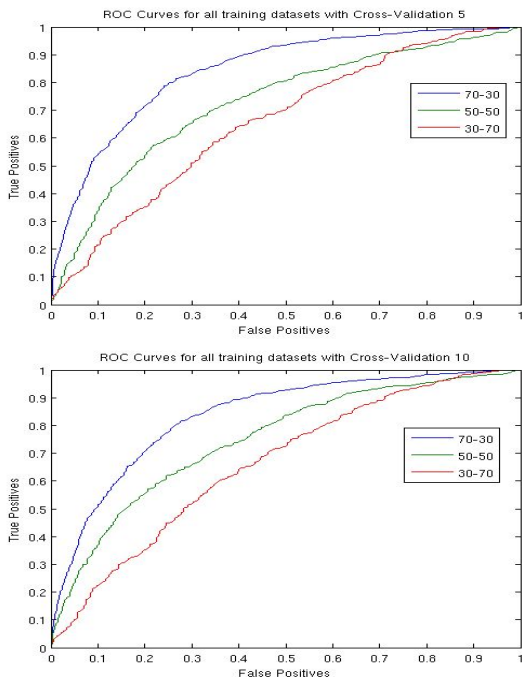


Figure 4. ROC Curves for SVMs.

CV	30%-70%	50%-50%	70%30%
5-fold	965	1270	1699
10-fold	965	1381	1637

Table 2. Number of Support Vectors

## 4. Conclusion

The result set that we obtained was encouraging. The biggest hurdle that we had to face was to cut down irrelevant data attributes and include ones which were more pertinent to train the algorithms to perform well on future datasets. The IBK (k=25) algorithm worked well on data sets which had more percentage of fraud. The J48 algorithm performed better

CV	30%-70%	50%-50%	70%30%
5-fold	965	1270	1699
10-fold	965	1381	1637

Table 3. AUC (Area Under the Curve)

```

start: 2007-12-02 19:40:00.734325

try feature sizes: [29, 14, 7, 3]

%#Feat est. acc.
29: 71.74980
14: 72.59310
7: 72.24170
3: 70.13350
max validation accuracy: 72.593100

select features: [2, 5, 1, 21, 27, 7, 19, 4, 22, 11, 12, 20, 10, 16]
14 features

end:
2007-12-02 19:44:16.414484

```

Figure 5. Feature Selection using SVMs.

when more informative data attributes were included such as the zip ratio and when date attribute was removed. Logistic Regression gave better accuracy on datasets with less fraud percentage. It was the most stable algorithm on our algorithm list. The SVM's performed pretty well on our datasets. Bayes Net performed well but we believe that we need more research on how well Bayes Net as an algorithm fits in to predict fraud. We believe that bagged J48 and logistic were our winning models on audited data which needs to be verified against un-audited data. It would most logical to perform cost sensitive classification of future datasets with insights obtained till now. We conclude after our experiments and many data preprocessing that with expert advice, wise selection and distribution of attributes and more thorough investigation, it is possible to predict potential fraud with high accuracy.

## References

Baensons B, van Gestel T, V. S. S. M. S. J., & J, V. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring.

*Journal of the Operational Research Society*,  
54, 1082–1088.

Bramer, M. Principles of data mining. *Springer*  
2007.

Chang, C.-C., & Lin, C.-J. (2001).  
*LIBSVM: a library for support vec-  
tor machines*. Software available at  
<http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Fawcett, T. (2006). An introduction to roc anal-  
ysis. *Elsevier's science direct- Pattern Recogn-  
ition Letters*, 27, 861–874.

Kaufmann, M. (2005). *Data mining: Practical  
machine learning tools and techniques, 2nd  
edition*. San Francisco.

N, C., & J, S.-T. (2000). *Support vector ma-  
chines and other kernel-based learning meth-  
ods*. Cambridge University Press.