Project Report

CREDIT RISK MODEL-TAIWAN CREDIT CARD CUSTOMERS

Abstract

This paper uses four data mining techniques to analyze the probability of default by the credit card customers in Taiwan by comparing the accuracy among them. A binary classification model is used to label the customers as defaulters or non-defaulters through the best parameters estimated from various techniques. The aim was to come up with features that will give the best predictability. Among the four techniques used, Adaptive Boosting is the one which works best for this data in estimating the defaulters.

Introduction

The credit card industry of the banking domain has always been a major concern for the banks in terms of identifying legitimate customers. There is a strong need for risk prediction especially in the financial industry to help manage uncertainty. Banking operations are something that we all come across in our daily lives. In the recent years the use of credit card has become very popular as it is one of the most convenient payment options for everyone. However, this convenience does come with its own risk for the banks. As the number of customers using credit cards increase, more efforts need to be taken to consider managing the risk involved in terms of delinquency. The overall objective of risk management is to utilize the past behavioral information of the customers- financial, demographic, personal information, and understand the patterns to make sound decisions for optimizing their profit.

The traditional approach to building a credit risk model, wherein the probability of default is to be estimated, utilizes a Logistic Regression methodology which not only gives good accuracy rate but also has easily interpretable results. But with the recent advancements in machine learning techniques, it is a good time to explore other ways to build risk prediction models. Thus, for the purpose of this paper four data mining techniques were explored: **Naïve Bayes, Logistic**

Regression, Classification Trees, AdaBoost. Credit risk here means the probability of a delay in the repayment of the credit granted (Paolo, 2001).

We expect to address one main question: Are there other methods apart from Logistic Regression that can perform well on this credit risk data to predict the defaulters. In the next section, we try and understand the four data mining techniques and their implications from related work that has been done on this subject. In Section 3 we discuss the problem settings, the models, and parameter estimation methods. Section 4 will walk through the experimental results and include model performance comparison. In Section 5, we try to analyze why certain models works better than others and further analysis. Section 6 concludes with the relevant findings.

Related Work

In order to predict the probability of default Yeh and Lien, 2009 used 6 different data mining techniques namely, - K-nearest neighbor (KNN), Logistic Regression, Naïve Bayes, Artificial Neural Networks, Classification trees and Discriminant Analysis. Their study focused on predicting the probabilities rather than just classifying the customers as defaulters and non-defaulters. It applied the Sorting Smoothing Method (SSM) for estimating the real probability of default from the model. For evaluating and comparing the performance of different models the Area ratios in the lift charts were used.

From the lift curves and the accuracy rate of the 6 techniques, they observed that on the training data, K-nearest neighbor classifiers and classification trees had the lowest error rate with KNN having a higher area ratio than other models. However, on the validation data, Artificial Neural Networks achieved the best performance with lowest area ratio and relatively lower error rate.

The above-mentioned research paper used scatter plot diagrams, regression line and R square to estimate the real default probability. Of the above methods, only Artificial Neural Networks had the highest explanatory ability in terms of R square as well as regression line.

In this we try to study 4 data mining techniques: Logistic Regression, Classification Trees, AdaBoost and Naive Bayes and check their performance through cross validations, ROC and other metrics. Following section explains the models and methods followed by the comparative analysis of the experimental results.

Models and Methods

Data Description:

The dataset contains information on default payments, demographic factors, credit data, history of payment, and monthly bill statements of 30,000 credit card clients in Taiwan from April 2005 to September 2005. We used the following 23 variables as explanatory variables:

X1: Amount of the given credit.

X2: Gender (1 = male; 2 = female).

X3: Education (1 = graduate school; 2 = university;3= high school; 4 = others).

X4: Marital status (1 = married; 2 = single; 3 = others).

X5: Age (year).

X6–X11: History of past payment. X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1= pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12–X17: Amount of bill statement. X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005.

X18–X23: Amount of previous payment. X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .; X23 = amount paid in April, 2005.

Exploratory Data Analysis:

There are 25 columns, all numeric values. Our target attribute is 'default payment next month', there are almost 4times cases of non-default versus default cases.

Average age of the applicants being 35.5 years, with a standard deviation of 9.2. The average value of the amount of credit card limit is 167,484. The standard deviation is unusually large, max value being 1M indicating more variance in the credit limit amount.

Females constitute higher proportion of credit card applicants (60%) and the education level of the applicants are mostly graduate school and university. The marital status is either married or single. Also, the correlation of the repayment status is decreasing between months with lowest correlations between Sept-April.

Predictive Modelling:

The dataset is randomly sampled by splitting into 70-30 proportion for training and test respectively. Since there are only 24 features in the dataset, we have used all of them for creating our baseline models. Moreover, we are using derived variables, which are the percentage transformation of the payment amounts. These variables have a better predictive power and are very stable. Hence, using these variables will give better prediction.

Decision Tree

The core idea is to recursively split a sample of the data with the best possible choice until some conditions are met.

Baseline model:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5698 1212
         1 1290  838
```

Final model: Only the important variables, derived payment variables considered and maxdepth=6.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6661 1345
         1  327  705
```

Naïve Bayes Classifier

The naïve Bayesian classifier is based on Bayes theory and assumes that the effect of an attribute value on a given class is independent of the values of the other attributes.

Baseline model:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5437  821
         1 1551 1229
```

Final model: The important variables, derived payment variables considered and no laplace smoothening.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6498 1235
         1  490  815
```

Logistic Regression

Logistic regression is used to predict the probability of occurrence of an event by fitting data to a logistic curve. A logistic regression model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables.

Baseline model:

Final model:

The important variables and derived payment variables considered.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6806 1553
         1  182  497
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6806 1576
         1  182  474
```

Adaptive Boosting

AdaBoost is a method in which the output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

Baseline model:

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6633 1321
         1  355  729
```

Final model: The important variables, derived payment variables considered and mfinal=100.

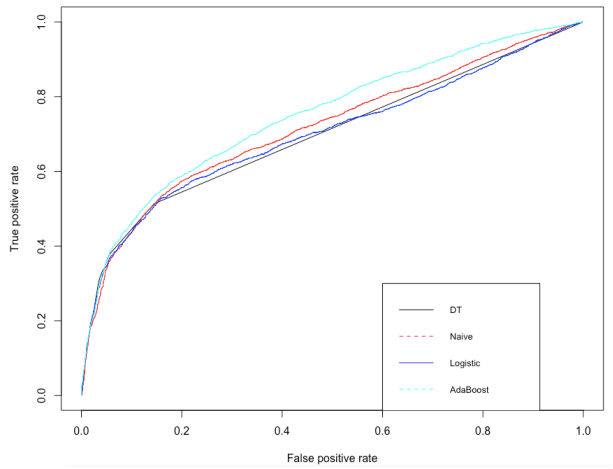
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6639 1309
         1  349  741
```

Experimental Results

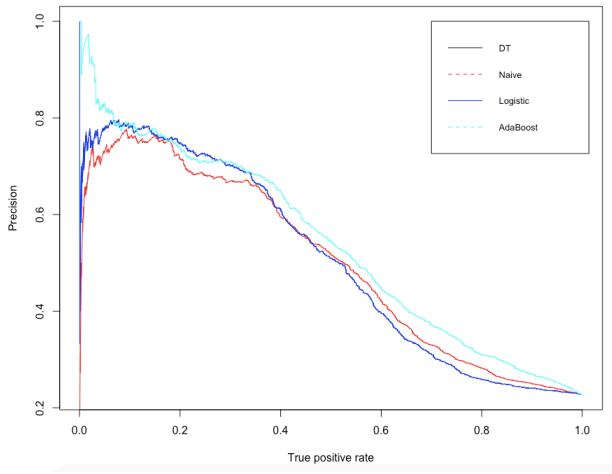
Baseline					
Methods	**Train Accuracy**	**CV accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.9945	0.7232	0.8246	0.8154	0.82
Naïve Bayes	0.7469	0.7376	0.8688	0.778	0.8209
Logistic Regression	0.8117	0.808	0.8142	0.974	0.8869
Adaptive Boosting	0.8226	0.8146	0.8339	0.9492	0.8878

Final Model						
Methods	**Train Accuracy**	**CV accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
Decision Tree	0.8236	0.815	0.832	0.9532	0.8885	0.6997
Naïve Bayes	0.8128	0.8091	0.8403	0.9299	0.8828	0.7173
Logistic Regression	0.8107	0.8055	0.812	0.974	0.8856	0.7002
Adaptive Boosting	0.8232	0.8166	0.8353	0.9501	0.889	0.7496

ROC curve:



PR curve:



Below are the findings from various model techniques:

1. Almost all the model performs similar in Accuracy rate.
2. The decision tree model has low AUC of ROC curve values as compared to other models.
3. Naïve Bayes and Logistic Regression have low accuracy and AUC of ROC curve values compared to Adaptive Boosting.
4. As we can see the precision value of AdaBoost model for low true positive rates is very high which indicates the correct predictions by the model even at low TPR levels.
5. The AUC of ROC curve for AdaBoost model is highest among all models which can also be seen from the plot above.

Thus, from the above observation, we conclude that Adaptive Boosting works the best for our data.

Discussion and Further Analysis

Adaptive Boosting is a linear combination of weak learners. This becomes easy to tune the models while improving the performance. AdaBoost is a powerful classification algorithm and require less tweaking of parameters. Thus, this is a good method to use when there are not many features in the model. We can see that this method

Decision trees produce rules that are relatively easy to use, understand and implement. However, they are sensitive to training data. Even a small change in the training data can cause large changes in the tree. Thus, care needs to be taken while changing the features or altering the dataset.

Naive Bayes Classifier as a method is computationally fast and simple to implement. However, it relies more on the independence assumption which may not give accurate results. Also, this method is susceptible to a single zero probability if Laplace estimator is not used.

Logistic Regression is a traditional approach used in building probability of default models. It is easily interpretable and computational. However, it might lead to high bias at times. Also, it works better for large sample size.

There are certain examples in our dataset that were predicted correctly by one model but incorrectly by other. No model is perfect but there are several reasons for this to happen as each model gives importance to certain characteristics of the explanatory variables while prediction. The reason for these changes needs to be further analyzed to get an idea about the unique features of different models.

Conclusion

This paper used 4 data mining techniques to analyze the probability of default by the credit card customers in Taiwan by comparing the accuracy of each. As a traditional approach, logistic regression had been used for a credit risk model classification as it is not only easy to interpret but also gives the probabilities of default computation easily. The aim of this paper was thus to come up with other better methods that can be used in machine learning. The 4 models showed little differences in accuracy rates. However, the AUC metric was the decision maker for this study where Ada Boost turned out to be better performing than others. Therefore, industry must start to explore methods like Ada Boost that goes beyond the logistic regression model in classifying the defaulters. Further study needs to be done in order to implement various machine learning algorithms in not only classifying the defaulters but also to take it a step further in predicting the probability of default which will be more appropriate for banks in taking decisions while formulating various policies.

References

1. Berry, M.,&Linoff, G. (2000). Mastering data mining: The art and science of customer relationship management. New York: John Wiley & Sons, Inc.

2. Paolo, G. (2001). Bayesian data mining, with application to benchmarking and credit scoring. Applied Stochastic Models in Business and Society,17, 69–81.

3. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

4. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.