Reyes Ceballos

Anna Jadonath

CUS 615 Nikhil Yadav

Term Paper Draft

**Predicting the Default of Taiwanese Credit Card Holders**



**Abstract:**

The main aim of this study is to compare and contrast several data mining models in their ability to predict the accuracy of default amongst Taiwanese credit card holders. The time frame examined was a six-month period in 2005. To do this, we compare seven data mining techniques in depth from a risk management perspective and drew comparisons using our case study research. The techniques used are Knearest neighbors, Kmeans, support vector machines, neural networks, random forest, logistic regression, and Gaussian Naïve Bayes.

**Introduction**:

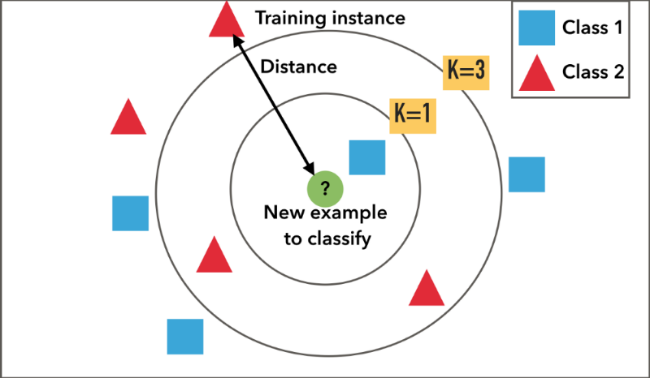
In Taiwan during February 2006, the debt from credit cards and cash cards reached $268 billion dollars.  More than half a million people became credit card slaves and were unable to make their credit card payments. The term “credit card slaves”, was coined in Taiwan referring to people who were unable to pay the minimum balance on their credit card debt every month. [1] Credit officers are faced with the problem of trying to increase credit volume without excessively increasing their exposure to defaults. [2]) In addition, the increased competition in the market and growing burdens on banks have forced the credit markets to explore alternative and effective methods of attracting new customers and controlling losses by reducing the default of loans. [3] This research paper provides an analysis of payment data and predicts whether a Taiwanese credit card holder will likely default on their monthly credit card payment. This analysis consists of the comparison of the data mining models: K-neareast neighbor (KNN), support vector machines (SVM), logistic regression, Naïve Bayes, random forest, Kmeans, and neural networks. When analyzing credit card risk, one must consider the probability of a late payments. [4] The probability that an applicant will default is typically estimated from the information which the applicant provided at the time of their initial application. When deciding whether to grant an individual credit, banks will analyze job information, credit history and macro and micro economic conditions, among other things. As time progresses, a credit card holders situation and economic environment is likely to change, and banks should reevaluate their exposure to default risk. For the banks, accurate classification methods are a benefit for increasing profits and reducing losses. For the applicant, avoiding over commitment with the crediting bank is of the utmost importance and priority. [5] Defaulting on a credit card can negatively impact an individual’s credit score and ability to receive new or additional credit and forestall major life decisions. To dissect whether Taiwanese credit card owners have over committed themselves, we followed the suggestion of researchers Rosenberg and Gliet to test various static and dynamic models which specifically make decisions for consumer credit. [6] Using seven data mining models and techniques in this research paper, we were able to predict and compare the accuracy percentage of credit card owners who would potentially default on their payments. Results should be taken into consideration when delinquent accounts are initiated leading to what action should be taken. [6]

Literary Review

K-nearest neighbor

K-nearest neighbor classifiers is a pattern recognition model that learns based on the k closest training samples in the featured area. Closeness is measured by various distance functions such as; Minkowski, Euclidean and Manhattan. A test sample is assigned the most common class among its K as shown in Figure 1. The disadvantage of KNN however, is it cannot produce a simple probability equation. This method of predictive accuracy can be biased by the distance measure used and the value of K chosen. In the Yeh and Lien paper, K-nearest neighbors had the lowest error rate and when compared to other models in the study, K- nearest neighbors performed better. [11]

**Figure 1**



Logistic Regression has been used by researchers in classifying creditors. Attributes are associated with the probability of default in binary form. This model can be expressed in following equation:

**Log [p/(1-p)] = α + β1X1+ β2X2 + ...+ βnXn**

In the above, p is the probability of the result, α is the constant, **β** is the coefficient and X is the value of the attribute. [3] In the Soric, Vlah and Rosenzweig study logistic regression was used to estimate credit risk by extracting the variables and deciphering which variables are important in credit risk prediction. They analyzed the logistic coefficient which showed that with an increase of the net cash flow, the other attributes such as business experience, profit margins etc. also increased. [13]

Random Forest

Random Forest is a set of randomized classification trees used in an ensemble method to increase predictive performance. One of the methods tested in our research was decision trees. [11] It was used as a decision tool and utilized as a tree-like graph model to make decisions. Decision trees are great for visualizing and interpreting results. However, they are prone to overfitting and face controlling error bias. In this study we selected random forest because it is more robust in nature in that it accumulates the results of many decision trees while simultaneously limiting overfitting and limiting bias error. [12]

K-means

K-means is a clustering method that partitions observations into K number of groups. This algorithm analyzes data and categorizes it based on locations and distances between various points. The K- Means algorithm process in steps is as follows: [14]

**Kmeans (dataset, K, dissimilarity\_measure, replicate)** [12]

1. **Randomly select a K number of initial centroids.**
2. **Repeat the first step and construct K clusters, as per the objective function**
3. **Re-compute the centroid of each cluster**

**until the number of duplicates produced or centroids do not change**

1. **End**

In our study, the K-means algorithm was used as a comparison to the other models and was found to be the lowest performer.

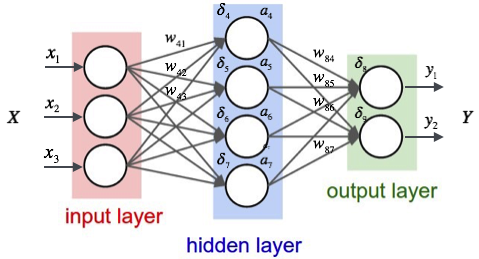
Support Vector Models

According to a classification study on credit scoring, Baesens found that both the LS-SVM and neural network classifiers yield a very good performance, but also simple classifiers such as logistic regression and linear discriminant analysis perform very well for credit scoring. [10] Similar to this study, research done by author Baesens used different types of classifiers. These classifiers were evaluated and compared to a myriad of popular classification algorithms such as logistic regression, k-nearest neighbor, neural networks and decision trees. Baesen's study is similar in that it explores and tests the support vector machines model as well. In both studies, classification accuracy was assessed.

Neural Networks

Neural Networks is a model that processes information based on the human brain. It consists of the input, hidden and output layers of connected neurons as seen in figure 2. It uses various activation functions and assigns weights to each node and adjusts them as it learns. In general, it was found that neural networks is a better model when compared to other models for learning and storing associations. In the Dutta and Shekhar study, Neural Networks were used in making credit decisions and fraud detection. To do this, they reviewed past approaches and found that they were only accurate 60% of the time. They attributed their limited success to the inability to decipher and define a mathematical model. Neural Network does not require a specific model and instead inherently learns a model from the raw data provided. [6] Neural networks was included in our study because a pre-specification of the model is not required. Neural networks are one of the most used models in credit processes because of its associated memory characteristic and generalization capability. [8] The neural network learns the relationships inherent within the data presented to it. It is an excellent solution to solving the problem at hand. [9]

Figure 2



Gaussian Naive Bayes

The Naïve Bayesian classifier, which is a binary classifier was used to get yes/no from the data. Some thinkers believe it can be a very primitive method of finding true or false classification from the dataset, but the algorithm is used as a predictive models in machine learning and data-mining. [7] In the comparative study done by Paolo in 2001, Bayesian classifiers were successfully employed to localize model specification with various inferences. This allowed for a considerable gain in flexibility when modeling the efficiency of computations. [4]

**Methodology**:

The dataset was acquired from UCI.edu. It consisted of a table of 25 columns by 30,002 rows. The data came with two rows of headers which are X1, X2, X3, etc. It was very simple and meant to make the attribute names short and easy to write. The second header was a little more descriptive of the attributes. The attributes are shown in Table 1.



Preprocessing

Some Preprocessing was done to the data. The first header provided no assistance when attempting to differentiate between the different attributes. This row was deleted as it added very little to the research. The second header was better but needed to be modified to be easy to use. The 18 columns of monthly data had been given as numbers. These were renamed to show the appropriate month year name. Having done this, it became evident that the columns had been listed in reverse order. This proved useful to know when looking at things such as whether clients credit card bills where increasing or decreasing and in determining the rate at which they paid off their debts. The first column labeled ID was deleted as it should not bias the different analytical techniques.

Normalization/standardization

Our first task was to create three different versions of the dataset. The first dataset had the above-mentioned preprocessing but nothing else. We will refer to it as the non-normalized dataset.

The second dataset was normalized. Due to the nature of the data, the normalization process was unusual. When observing the first attribute, credit limit, one notices a wide range of $10,000 to $1,000,000. This attributed was also heavily skewed right. Credit limits are typically given by banks depending on what they perceive a client’s ability to pay. Therefore, an individual A who has a bill of $10,000 dollars and has hit his/her maximum credit limit is very different than individual B who has the same bill but a credit limit of $1,000,000 dollars. Individual A would have a difficult time paying of their balance especially after the interest rates which have been charged are taken into consideration. Naturally, they would likely to default. Individual B on the other hand, would have an easier time paying off that balance and may even attribute that quantity as a rounding error in their checking account. For this reason, attributes 12 – 23 were normalized as a percentage of the total credit limit. Attributes 2-4 were left as is since they were discrete variables and age was also left as is. Attributes 6-11 represented how many months the client was late on their payments. This was also left as is. This data set will be referred to as our normalized data set.

Our third data set consisted of feature selection. Three features, late payments, amount of bill and payments amounts were looked at. Each of these features consisted of 6 months of data. To simplify the process, we took the average of each of the 6 months, thus reducing the number of attributes from 18 to 3. Except the times mentioned above all other features were used as it was monthly data. This dataset will be referred to as our feature selection dataset.

Training / testing sets

Our data was randomized and split into training and testing sets. 10% was used for testing and the rest was used for training.

In addition to the cross validation method ten-fold was used to detect overfitting. For every model, a cross validation score was calculated and compared to the test the accuracy.

**Results**:

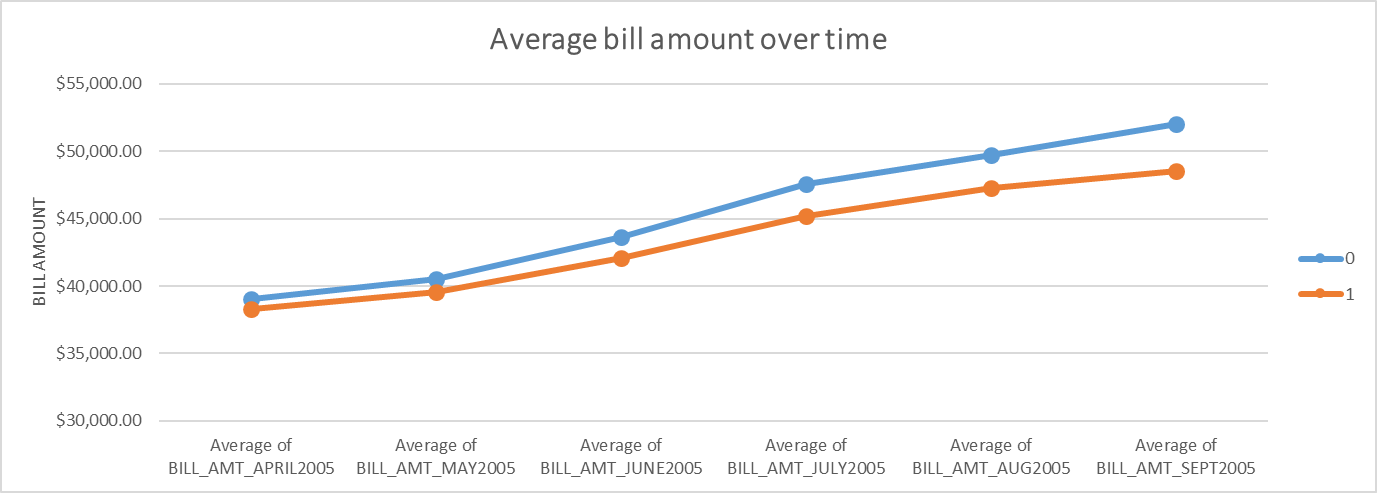
Descriptive statistics / Visualizations

Our first analytical technique consisted of some basic descriptive statistics which allowed us to get a better understanding of the data. The mean credit limit was $167,484 and skewed right. Table 2 lists out the credit limits by groups of 100,000. The default rate is highest for those with the lowest credit limit and lowest for those with the highest credit limit.



Table 3 has the default rates for the 4 demographic variables. Females made up 60% of the observations and were 3.4% less likely to default than males. When the education attribute was tested and observed, 16.4% had a high school degree, 46.8% had a bachelor’s degree, 35.3% had a graduate degree and the remaining 1.5% was not specified. As luck would have it the 1.5% of the observations that were not specified had the lowest default rates. Those labeled 0 did not default. If they were better identified, credit card companies should seek to increase their marketing to this demographic. Those with graduate degrees had default rates of 19%, bachelor degrees defaulted 24% of the time and those with only a high school degree had the highest default rate at 25%. Moving on to marital status 45.5% were married and 53.2% were single with 1.3% not specified. Married couples were 2.5% more likely to default than those who were single. Age turned out to have the clearest difference in default rates. Twenty year olds had the lowest default rate at 22.44% and it progressively increased to 33.33% default for those in their eighties. The mean age was 35. The next 6 variables describing how late clients were on their payments showed at least 75% were on time. Those who defaulted were on average late on their payments more often than those who did not. The average bill payment amount steadily increased for everyone from April to September but it surprisingly increased more so for those who did not default as seen in figure 3. At the same time those who did not default were also making significantly higher monthly payments. Finally, 22% of the observations defaulted in October 2005 and 78% did not.

Figure 3





The data was modeled with several different algorithms. Accuracy, mean square error (mse) and precision were calculated. Ten-fold cross validation was also done to test for overfitting. The main performance measure we will be focusing on is precision. Precision will tell us our accuracy rate in correctly predicting the default rates. Accuracy tells us our prediction overall. It is possible that precision can be a lot for one model and accuracy high for another. In case of differing accuracy and precision we will be used. Table 4 is presented as a summary of the best results among all three data sets. All algorithms were run from the Sklearn library. You may refer to Figure 4 to better visualize what was done.

KNN

The K-nearest neighbors (KNN) algorithm was run for various K’s for all three versions of the data. K of 1, 2, 3, 5, 20, 50, 100, 500, 1,500, and 3,000 were used. For all 3 versions, KNN proved to have the highest precision out of all the models. The K of 3,000 was found to have the highest precision of 77.42% for the non-normalized dataset. K < 5 was less than 50% and K = 5 – 1,500 was in the 60% range. The K of 3,000 increased by 7% in performance when compared to the previous K’s and was therefore the best performer.

Neural Networks

The MLP classifier was used for neural networks. In our study, various activation functions (identity, logistic, tanh, relu) and solvers (ibfgs, sgd, adam) were used. The best performer was activation “identity” and solver “ibfgs” with 71.6% precision. It performed best with the non-normalized dataset. Neural nets was the second highest performer, but it proved to be 6% less accurate than KNN.



Logistic Regression

Logistic regression had the third highest precision accuracy with 71.5%. In our testing using the non-normalized dataset it was found that Logistic Regression was significantly lower than KNN by 6%. In comparison with neural networks, there was only a 0.1% difference.

Random Forest

Random forest was run for different number of leaves (3, 10, 50) and trees (10, 30 50). The best performer was 50 leaves and 50 trees with a precision of 67% using the normalized data set. It was the fourth best performer.

SVM

Support Vector Machines (SVM) was run for two different kernels, sigmoid and rbf. Its best performer was rbf at a precision of 66.37% for the normalized data set. It was the fifth best performer. among the other models.

Gaussian Naive Bayes

Gaussian Naive Bayes was the second worst performer. It had a precision of 61.7% the dataset with feature selection.

K-means

The K-means algorithm proved to be the worst performer. Its highest accuracy was a K of 4 with a precision of 36% with the normalized dataset. We tested K of 1 – 10 with decreasing precision for each increasing K.

Conclusion:

This paper examined seven different classification techniques on three different versions of a credit card defaults data set. As evaluated by the accuracy in identifying the default card holders, KNN, neural networks and logistic regression proved to be the most accurate classifiers and performed best on the non-normalized dataset. Of the remaining four methods, three performed best on the dataset which was normalized while only one performed its best on the normalized dataset that had feature selection. The performances listed here in the order of highest to lowest are KNN, neural networks, logistic regression, random forest, SVM, Gaussian Naïve Bayes, and Kmeans. Our results show that the best performer was KNN which performed best on all tested data sets.

Further Study

This study was limited to analyzing customer credit payment data with a few demographic variables. Future studies can look into analyzing more data such as comparative studies in different countries and demographics analyzing their shopping trends, purchase history and re-evaluating initial credit approval requirements such as jobs and salary.

**References**:

Dataset

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Code used in this project is from the Sklearn library

Citations

1. Chou, M. (2006). Cash And Credit Card Crisis In Taiwan Business Weekly, 24–27.
2. Cheng-Lung Huang et al., Credit scoring with a data mining approach based on support vector machines, Expert Systems with Applications (2006), doi:10.1016/j.eswa.2006.07.007
3. Asia Samreen, (2013), “Design and Development of Credit Scoring Model for the Commercial Banks in Pakistan: Forecasting Creditworthiness of Corporate Borrowers” International Journal of Business and Commerce Vol. 2, No.5: Jan 2013[01-26] (ISSN: 2225-2436)
4. Paolo, G. (2001). Bayesian data mining, with application to benchmarking and credit scoring. Applied Stochastic Models in Business and Society,17, 69–81.
5. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society, Series A – Statistics in Society, 160(3), 523–541.
6. Rosenberg, E., & Gleit, A. (1994). Quantitative methods in credit management: A survey. Operations Research, 42 (4), 589–613.
7. NH Niloy, MAI Navid, Naïve Bayesian Classifier and Classification Trees for the Predictive Accuracy of Probability of Default Credit Card Clients, American Journal of Data Mining and Knowledge Discovery. Vol. 3, No. 1, 2018, pp. 1-12. doi: 10.11648/j.ajdmkd.20180301.11
8. Overstreet, Jr., G.A., and Bradley, Jr., E.L. (1994), "Applicability of generic linear scoring models in the USA Credit Union environment: Further analysis", Working Paper, University of Virginia
9. Desai, V. S., Crook, J. N., & Overstreet, G. A. A. (1996). Comparison of neural networks and linear scoring models in the credit union environment. European Journal of Operational Research, 95 (1), 24–37.
10. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-ofthe-art classification algorithms for credit scoring. Journal of the Operational Research Society, 54(6), 627–635.
11. I.-C. Yeh, C.-h. Lien (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, Expert Systems with Applications 36 2473–2480I
12. Gérard Biau. 2012. Analysis of a random forests model. J. Mach. Learn. Res. 13, 1 (April 2012), 1063-1095.
13. Sarlija, Natasa & Soric, Kristina & Jerić, Silvija & Vojvodić Rosenzweig, Višnja. (2009). Logistic regression and multicriteria decision making in credit scoring. Proceedings of the 10th International Symposium on Operational Research in Slovenia, SOR 2009.
14. Sahu, Santosh. (2014). A Study of K-Means and C-Means Clustering Algorithms for Intrusion Detection Product Development. International Journal of Innovation, Management and Technology. 5. 207-213. 10.7763/IJIMT.2014.V5.515.