JIST Journal of Information Science and Technology
Volume 13, NO 1 | JAN – JUN 2023 | 38-44
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

An Investigation of Machine Learning Techniques for Loan Default Payments Prediction

*Jesada Kajornrit, Wilawan Inchamnam and Waraporn Jirapanthong**
Collage of Creative Design and Entertainment Technology
Dhurakij Pundit University, Bangkok, Thailand

Received: June 10, 2023; Revised: June 11, 2023; Accepted: June 24, 2023; Published: June 30, 2023

ABSTRACT – In banking business, loan default payments of individual customers are counted as risks that result in the loss of the business. Thus, some assessment mechanisms are needed to assess the risks of individual customers who apply for personal loan products. This paper presents an investigation of machine learning techniques to predict loan default payments based on individual customers information backgrounds. The paper emphasis on the ensemble techniques that mostly used in banking business. Besides the ensemble prediction models, the principal component analysis is also used for further investigation. The experimental results showed that all prediction models provided acceptable prediction of non-defaulting payment class, but provided unacceptable prediction of default payment class. That is because the imbalance nature of the data and the features used are not specific enough for the prediction models to classify the minor class from the major class. This paper acts as an initial study of the credit default payment analysis.

KEYWORDS: Loan Default Payments, Imbalance Data, Machine Learning, Ensemble Techniques, Dimensionality Reduction

1. Introduction

In banking business, individual customer loan evaluation is an important process to reduce the impact of default payments and to mitigate risks into an acceptable level. Presently, the process of loan evaluation is not only limit to human expert decision, but also additional modern analytics techniques. Loan default payments prediction model makes use of the current and historical information related to the customers to make a prediction about the customer ability to pay back on time [1]. Accurate prediction model is able to enhances the decision making of human experts with higher confidence. Thus, developing accurate loan default payments prediction system is an important task for the bank profitability and sustainability.

Presently, machine learning techniques have been taking over several businesses and, of course, banking business is no exception. By using machine learning prediction models, banks are able to predict the probability of loan default payments in advance, and thus helping them in mitigating risks. It is indisputable that the success of machine learning model depends on the high quality of training data. Unfortunately, most loan default payments data (or other related credit default payments data) are usually imbalance [2]. Number of default payments data are less than non-default payments data significantly. Furthermore, features of the dataset are varied among organizations. One way to assess the feasibility of developing a prediction system is to test the limitation of prediction techniques toward the dataset. This paper reports our feasibility study.

The paper is organized as follows; Sections 2 presents some related literatures and machine learning techniques. Section 3 describes the statistical views of the loan default payments dataset. Section 4 present

*Corresponding Author: waraporn.jir@dpu.ac.th

experimental results and some discussions. Section 5 is our conclusion.

2. Related Backgrounds

2.1 Machine Learning Techniques

Decision Trees (DT) are a non-parametric supervised learning method used for classification (and regression) tasks. A DT model is generated from labeled dataset by means of learning algorithms. The popular algorithms include ID3, C4.5, and CART. Some advantages of DT are that it is simple to understand, interpret, and visualize. The DT requires little data preparation [3].

Random Forest (RF) is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting. This is called ensemble method. The goal of ensemble methods is to combine the predictions of several base estimators built with a given learning algorithm in order to improve generalizability and robustness over a single estimator [4].

Extremely Randomized Trees (ERT) is one that randomness goes one step further in the way splits are computed. As in Random Forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias [5].

Both RF and ERT fall into an averaging method, that is, the driving principle is to build several estimators independently and then to average their predictions. On average, the combined estimators are usually better than any of the single base estimator because its variance is reduced.

The core principle of AdaBoost Tree (ADT) is to fit a sequence of weak learners on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote to produce the final prediction. Each subsequent weak learner is forced to concentrate on the examples that are missed by the previous ones in the sequence [6].

Gradient Boosted Decision Trees (GBT) is a generalization of boosting to arbitrary differentiable loss functions. GBT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems in a variety of areas including web search ranking and ecology [7].

Both ADT and GBT fall into a boosting method, base estimators are built sequentially and one tries to reduce the bias of the combined estimators. The motivation is to combine several weak models to produce a powerful ensemble.

Principle Component Analysis (PCA) is linear dimensionality reduction using Singular Value Decomposition (SVD) of the data to project it to a lower dimensional space. The input data is centered but not scaled for each feature before applying the SVD. PCA technique falls into unsupervised machine learning type. PCA is practically used to visualize high-dimensional data. Also, it is used as a pre-processing step to reduce the dimensions of input vector before feeding to the prediction model [8].

2.2 Literatures Reviews

Many literatures related to loan default payments prediction prefer using interpretable machine learning models liked Decision Trees and its ensemble techniques. That is because such models provide good prediction results, relatively fast training time, require little data preprocessing, and provide human-understandable prediction mechanism. Some literatures are presented as follows.

Soni and Shankar [9] adopted Random Forest classification to predict bank loan defaulting. They claimed that the ensemble technique outperforms single model such as logistic regression, k-nearest neighbours, support vector machine, and decision tree classification.

Shaheen and ElFakharany [11] showed that Random Forest and Gradient Boosting Tree outperform single technique in prediction accuracy when apply these models to predict load default dataset

Fan [10] compared LightGBM to Random Forest algorithms to predict personal loan defaulting. He claimed that LightGBM showed the better prediction;

Lai [12] confirmed the performance of AdaBoost that is show better performance than those of XGBoost, Random Forest and K-Nearest Neighbors, and Neural Network in order to predict loan defaulting from real-world dataset of a prestigious international bank.

Barua et al. [13] explored the use of CatBoost algorithm for loan default prediction. CatBoost is a fast-learning algorithm which is capable of handling categorical data type. They compared to the Random Forest and Gradient Boosting Tree. They claimed that CatBoost achieved the highest accuracy amongst all other algorithms.

Al-qerem et al. [14] presented some different classification methods including Naïve Bayes, Decision Tree, and Random Forest for loan defaulting prediction. Furthermore, comprehensive different pre-processing techniques are being applied on the dataset, and three different feature extractions algorithms are used to enhance the accuracy and performance.

Patel et al. [15] used Logistic Regression, Gradient boosting, CatBoost Classifier, and Random Forest to forecast loan default. They claimed that Gradient boosting and CatBoost Classifier provide the equivalent accuracy and slightly better than Random Forest. While Logistic Regression provided unacceptable result.

Up to this point, however, accuracy of prediction is subject to dataset characteristics, especially, features of the dataset. Furthermore, such problem becomes difficult when it exhibits a profile of imbalanced data, because classifier may misclassify the rare samples from the minority class. To find out an appropriate solution of a specific dataset, some preliminary study is necessary.

3. Loan Default Payments Data

The dataset in this paper is bank loan default payments of individual customers. As the original dataset are confidential, the dataset is anonymized. The features of the dataset are personal information of the customers. The original dataset consists of sixteen features as shown in the Table 1. However, for practical reasons, some features such as LOAN_ DATE or ZIP are removed since they are not appropriate to use. The records that contained *NaN* values are removed. Table 2 shows some statistic of numeric features and Table 3 shows some statistics and numeric code of category features.

The dataset contains 42767 records of customers details which are labeled as default payment (1) or non-default payment (0). The dataset is divided into cross- validation data and final-validation data. Number of cross-validation data is 32075 records and number of final-validation data is 10692 records. The imbalance ratio of cross-validation data is 1:4.5 and 1:4.8 for final-validation data.

4. Experiment and Discussion

The entire loan defaulting payment dataset are divided into cross-validation dataset and final-validation dataset. The former dataset is used to determine the optimal hyper parameters of the models, and is also used to train the models for final validation. The later dataset is used to validate performance of the trained models. The proportion of the data division is about 75 to 25 percent.
The machine learning techniques in this experiment include Decision Tree Classifier (DT), Random Forest (RF), Extremely Randomized Tree (ERT), AdaBoost Tree (ABT), Gradient Boosting Tree (GBT). All algorithms are developed based-on Scikit-Learn library (*https://scikit-learn.org*). The validation measures include accuracy, precision, recall and F1-scores. The number of folds in cross-validation process is set to 5.

The results from the cross-validation process pointed out that “gini” is the good selection criterion for all selected models. The DT model selected 30 for the “max_depth” parameters. The RF model selected 20 for the “n_estimator” parameter and do not set limitation of the “max_depth” parameter. The ERT model was set the same as the RF model, except select 15 for the “n_estimator” parameter. The ABT model set “n_estimator” to 20, and the rest parameters are the same as the RF model. Finally, the GBT model sets “n_estimator” to 50. All optimized models are tested with final-validation dataset and the results are shown in the Table 4.

According to the Table 4, overall accuracy measures of the models are fall between 70 to 80 percent. The model performance can be ranked as RF > ERT > ABT > GBT > DT respectively. This results obviously show that ensemble techniques yield better accuracy than single model, in turn, Bagging technique yields better accuracy than Boosting techniques. However, the dataset is likely to imbalance, accuracy measure may not give the substantial performance of the prediction

In term of class_0 prediction (non-defaulting payment), overall result is acceptable. The precision measures are about 83 percent and the recall measures range between 79 to 95 percent. The model performance based on this measure are the same as one describing by the accuracy measures. The Bagging ensemble technique show the best premising prediction and all ensemble techniques provide better prediction than single model outstandingly.

Table 1. The dataset features

No	Features	Type	Use	Value Range
1	AGE	Integer	Y	[18, 71]
2	COMPANY_TYPE	Category	Y	5 Unique Values
3	CUSTID	Identifier	Y	
4	CUSTOMER_DOB	Category	N	-
5	EDUCATION	Category	Y	4 Unique Values
6	LOAN_AMOUNT	Integer	Y	[19415, 100513]
7	LOAN_DATE	Date	N	-
8	MARITAL_STATUS	Category	Y	4 Unique Values
9	NO_OF_DEPENDENT	Integer	Y	[0, 44]
10	SEX	Category	Y	2 Unique Values
11	STATE_NAME	Category	Y	21 Unique Values
12	TOTAL_MONTHLY_INCOME	Integer	Y	[0, 1500000]
13	YEARS_OF_EXPERIENCE	Integer	Y	[0, 70]
14	YRS_IN_PRESENT_JOB	Integer	Y	[0, 60]
15	ZIP	Category	N	-
16	LABEL	Category	Y	2 Unique Values

Table 2. The dataset statistics of numeric features

Statistics	Age	Loan Amount	Number of Dependent	TotalMonthly Income	Year of Experience	Year in Present Job
mean	33.371	49700.842	1.048	16784.980	6.450	6.209
std	9.589	6462.786	1.417	14917.690	6.614	6.156
min	18.000	19415.000	0.000	0.000	0.000	0.000
max	71.000	100513.000	44.000	1500000.000	70.000	60.000
skewness	0.674	-0.132	2.648	55.410	2.189	2.115
kurtosis	-0.311	0.482	36.034	5056.051	6.074	4.932

Table 3 The dataset statistic of category features

Code	Name	Counts
COMPANY_TYPE		
0	Government	6883
1	Individual	7047
2	Private limited company	18817
3	Public limited company	1856
4	Others	8162
EDUCATION		
0	High school	15610
1	Graduate	18787
2	Post-graduate	970
3	Others	7400

Code	Name	Counts
MARITAL_STATUS		
0	married	31365
1	single	11292
2	widowed	76
3	divorced	34
SEX		
0	male	4240
1	female	38527

Table 4. Experimental Results without PCA transformation

Models	Class 0			Class 1			Accuracy
	Precision	Recall	F1	Precision	Recall	F1	
DT	0.833	0.799	0.816	0.194	0.233	0.212	0.701
RF	0.832	0.959	0.891	0.268	0.072	0.113	0.806
ERT	0.832	0.924	0.876	0.225	0.106	0.144	0.783
ABT	0.834	0.862	0.848	0.208	0.174	0.189	0.744
GBT	0.834	0.820	0.827	0.202	0.218	0.210	0.716

Table 5. Experimental Results with PCA transformation

Models	Class 0			Class 1			Accuracy
	Precision	Recall	F1	Precision	Recall	F1	
DT	0.831	0.818	0.824	0.188	0.203	0.195	0.712
RF	0.830	0.943	0.883	0.212	0.074	0.109	0.793
ERT	0.831	0.919	0.873	0.210	0.103	0.138	0.779
ABT	0.829	0.871	0.850	0.182	0.137	0.156	0.745
GBT	0.832	0.832	0.832	0.193	0.193	0.193	0.722

However, in term of class_1 prediction (defaulting payment), overall result is unacceptable. All classifiers provide very low performance in both precision and recall measures. Precision measures range between 19 to 26 percent. and the recall measures fall under 23 percent. Considering the recall measures, the model performance is contrast to class_0 prediction results. The higher value comes from single classifier (DT) and the lower values come from ensemble technique, especially, Bagging algorithms.

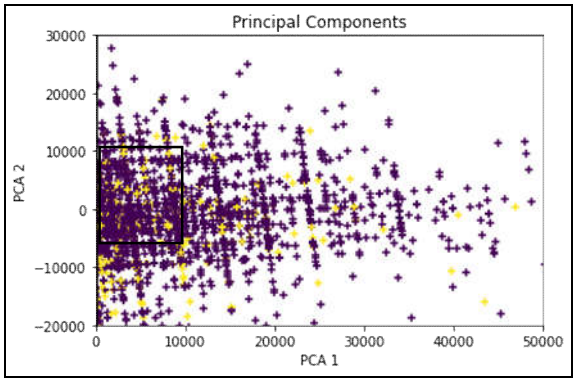
Some further analysis is conducted by unsupervised technique because of the unacceptable prediction of the class_1. This experiment used PCA technique to map high-dimensional default payments data into two dimensions in order to observe homogeneity of the data. The Figure 1 (a) shows the scatter plot between the first and second principal components of the data. The Figure 1 (b) shows the expansion view of the principal components in the rectangle area.

According to the Figure 1, it is possibly that the data points of the both classes are too close. Such data are not in clustered regions. This heterogeneous data is very difficult for classifier to create boundary decision between them. Further experiment is conducted a bit more on the hypothesis that “is it possible to use PCA as preprocessing step before prediction by machine lineaging”. After observed the variance of principle components (Figure 2), the data are transformed by PCA into three-dimensional data

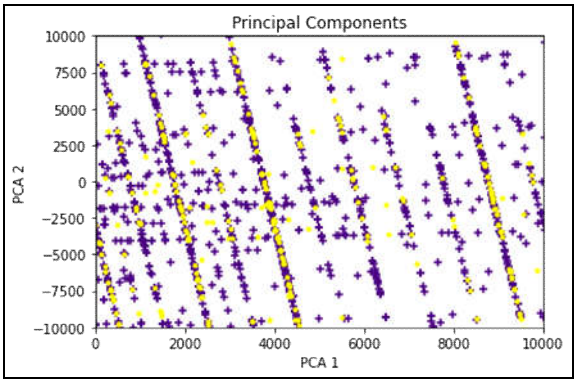
(dimensional reduction). The experimental result show in Table 5. Unfortunately, it seems that PCA techniques cannot improve the performance of class_1 prediction.

The experimental results aforementioned could be summarized as:

- The ensemble models provide better overall accuracy and prediction of the major class than the single model.



(a)



(b)

Figure 1. The first two PCA components of default payment data

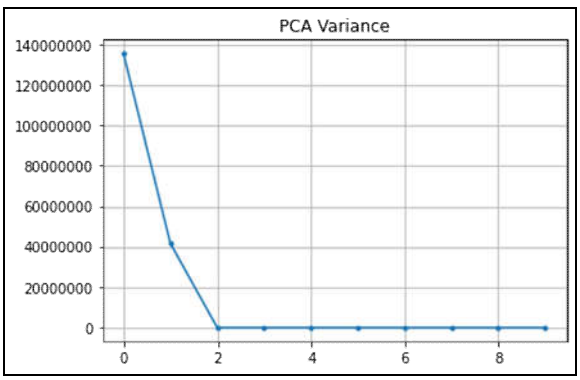


Figure 2. The variants of the PCA transformation.

- In turn, the Bagging techniques provide better overall accuracy and prediction of major class than Boosting techniques
- The ensemble techniques tend to increase the recall measure of the major class, especially, the Bagging technique, but do not improve the precision measure of the major class.
- The ensemble techniques tend to increase precision measure and decrease recall measure of the minor class, especially, Bagging technique.
- PCA transformation (dimensional reduction) may not be an appropriate for preprocessing step as the technique cannot separate the major class form the minor class.

So, the next step to achieve this challenge are as follows.

- Deep down analysis of dataset is need so as to find more relationships in the data. Some data preprocessing may be applied to transform values or to filter out unnecessary information.
- The number of features of the data may not be adequate to classify both classes. More features must be investigated form the data warehouse.
- As the data tend to be imbalance, some techniques such as SMOTE or ADASYN must be investigated (*https://imbalanced-learn.org*).

5. Conclusion

This paper presents an investigation of machine learning technique to predict customer loan default payments. Machine learning techniques include Decision tree, Random Forest, Extremely Randomized Tree, Adaboost Tree, and Gradient Boosted Tree. The experimental results show that prediction of non-default payment class is high accurate. But prediction of default payment class needs to be improved as the poor accuracy. Furthermore, Principal Component Analysis is used to visualize the homogeneity of the data. It showed that minor class data are mix up into major class region. Besides, the dataset is likely to imbalance. This resulted in poor performance of minor class prediction, especially ensemble techniques. Thus, the future study should focus on how to handle the imbalance problem and how to gather and extract more necessary features form the bank's data warehouse**.**

References

[1] A. K. I. Hassan and a. Abraham, "Modeling Consumer Loan Default Prediction Using Ensemble Neural Networks," In Proc. of Int. Conf. on Computing, Electrical and Electronic Engineering (ICCEEE), Khartoum, Sudan, 2013, pp. 719-724.

[2] T. Alam, K. Shaukat, I. A. Hameed, S. Luo, M. U. Sarwar, S. Shabbir, J. Li, and M. Khushi, "An Investigation of Credit Card Default Prediction in the Imbalanced Datasets," IEEE Access, Vol 8. October 2020, pp. 201173-201198.

[3] M. Dumont, R. Maree, L. Wehenkel, and P. Geurts, "Fast multi-class image annotation with random subwindows and multiple output randomized trees,", in Conf. Computer Vision Theory and Applications, Lisboa, Portugal, February 2009.

[4] L. Breiman, "Random Forests, "Machine Language, Vol. 45, No. 1, October 2001, pp 5–32.

[5] P. Geurts, D. Ernst., and L. Wehenkel, "Extremely randomized trees", Machine Learning, Vol. 63, No. 1, 2006, pp. 3-42.

[6] Y. Freund, and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, Vol. 55, No. 1, August 1997, pp. 119-139.

[7] J. H. Friedman. "Greedy function approximation: A gradient boosting machine," Ann. Statist, Vol. 29 No. 5, October 2001, pp. 1189 - 1232.

[8] O. Sornil and J. Kajornrit, "Improving performance of neural network models for intrusion detection using singular value decomposition", Recent Advances in Intelligent Systems and Signal Processing, WSES Press, 2003, pp. 338-342.

[9] A. Soni and K. C. P. Shankar, "Bank Loan Default Prediction Using Ensemble Machine Learning Algorithm" In Proc. of 2nd Int. Conf. on Interdisciplinary Cyber Physical Systems (ICPS), Chennai, India, 2022, pp. 170-175.

[10] S. Fan, "Design and implementation of a personal loan default prediction platform based on LightGBM model," In Proc. on 3rd Int. Conf. International Conference on Power, Electronics and Computer Applications, Shenyang (ICPECA), China, 2023, pp. 1232-1236.

[11] S. K. Shaheen and E. ElFakharany, "Predictive analytics for loan default in banking sector using machine learning techniques," In Prod. Of 28th Int. Conf. on International Conference on Computer Theory and Applications (ICCTA), Alexandria, Egypt, 2018, pp. 66-71.

[12] L. Lai, "Loan Default Prediction with Machine Learning Techniques," In Prod. of Int. Conf. International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020, pp. 5-9.

[13] S. Barua, D. Gavandi, P. Sangle, L. Shinde, and J. Ramteke, "Swindle: Predicting the Probability of Loan Defaults using CatBoost Algorithm," In Proc. of 5th Int. Conf. on Computing Methodologies and Communication (ICCMC), Erode, India, 2021, pp. 1710-1715.

[14] A. Al-qerem, G. Al-Naymat, and M. Alhasan, "Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection," In Proc. of Int. Arab Conference on Information Technology (ACIT), Al Ain, United Arab Emirates, 2019, pp. 235-240.

[15] B. Patel, H. Patil, J. Hembram, and S. Jaswal, "Loan Default Forecasting using Data Mining," In Proc. of Int. Conf. for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4.