Proceedings of 6th International Congress on Information and Communication Technology
(ICICT 2021), February 26-27 2021, London, UK, pp. 609-618
DOI:10.1007/978-981-16-1781-2_54

Using Big Data to Compare Classification Models for Household Credit Rating in Kuwait

Najla Albarrak[1,2][0000-0003-1502-7268], Hessa Alsanousi[1,2][0000-0001-5820-3781], Irene Moulitsas[2][0000-0003-0947-9495] and Salvatore Filippone[2,3][0000-0002-5859-7538]

[1] Central Bank of Kuwait, P. O. Box: 526, Safat 13006, Kuwait
[2] Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK
[3] University of Rome Tor Vergata, Via Cracovia, 50, 00133 Roma RM, Italy
n.albarrak@cranfield.ac.uk
h.alsanousi@cranfield.ac.uk
i.moulitsas@cranfield.ac.uk
salvatore.filippone@cranfield.ac.uk

Abstract. Credit rating risks have become the backbone of bank performance. They are the reflection of the current status of the bank and the milestone for future planning. A good credit assessment can better anticipate expected losses and will minimize unexpected losses from accumulating. Given advancements in technology as well as the big data available within banks about customers in an oil country such as Kuwait, a built-in model to help in-household credit scoring is at management's decision. Compared to the current 'black box' rating models, we did a comparison between different classification models for two types of banking: conventional and Islamic. The classification models are as follows: Logistic Regression, Fine Decision Tree, Linear Support Vector Machines, Kernel Naïve Bayes, and RUSBoosted. Sufficiently, the last could be used to classify banks household customers and determine their default cases.

Keywords: Credit Rating Model, Credit Risk, Technology, Conventional Banking, Islamic Banking, Classification Models, Household Customers, Machine Learning, Logistic Regression, Fine Decision Tree, Linear Support Vector Machines, Kernel Naïve Bayes, RUSBoosted.

1 Introduction

Financial crisis is not a new phenomenon or term. Rather, it is an ongoing bubble from the early stages of financial world. The early stages left us with lessons in order to avoid problems before the next crisis appear. The main obstacle is how much this lesson costs and who is in charge of paying the bill. Banks in Kuwait have gone through several crises including the 2008 crisis and the oil prices drop. Kuwait stood strong without the need for any support from regulators buddy for backup. This is due to the stringent regulations in place and the timely adoption of international regulations with additional safe buffers than internationally accepted benchmarks. Although we satisfactory

overcame the crisis, there is always space for improvement to insure better climate for the next crisis arrival. Crises cannot be avoided given the globalization in the business world; however, we could be better prepared for it. From the international 2008 financial crisis, Basel committee has emphasized on the minimum capital available to cover the riskiness of bank's assets and investment decisions. Banking industry mainly operates by utilizing raised capital and borrowed funds to lend money and profit from the difference rates.

The remainder of our paper is organized as follows. Section 2 provides a review of relevant literature. Section 3 describes our methodology. Section 4 describes our data collection. Our results are presented and analyzed in Section 5. Section 6 concludes the article.

2 Literature Review

Recent events shed importance on credit risks, which eventually drew the attention of bankers and regulators with regards to managing the credit portfolio efficiently. There are several machine learning and deep learning options to examine credit probability of default [1]. In a very recent International Monetary Fund working paper [2], the advantages in financial technology are discussed and how machine learning solutions could reduce the cost of credit and to provide much clearer solutions than the 'black box' templates for nontechnical audiences. From corporate to household customers, the recent literature recognized the importance of programming in loan granting process. Advancements in technology have created collaboration between fields, computer science and finance, in order to benefit from the witnessed technological efficiency. In the context of using classification models for calculating the probability of customers loan default, there is a wide range of options to be implemented. Providing a credit rating model is fruitful for regulators and banks decision making evenly due to the advancement in the technology and the outperformance of the models developed and tested [3]. A study compared 17 different methods for credit classes showing that it is suitable to conduct classification methods for credit rating [4]. However, they did not use heavily imbalanced data as the case in their study. Imbalanced data in our context is data that is not well distributed between the two classes. In our data, almost 85% of the data is considered as good customers and the remaining 15% are prone to default. Given that when comparing machine learning methods to classify credit default customers, several studies showed that RUSBoost is the most significant method [5]. In research, it is evidenced that that RUSBoost performance is significantly better for imbalanced data [6].

This paper aims to cover the gap of small size samples for studies with long-big data samples [1]. As recommended in the literature, ensemble learning and gradient boosting decision trees, are a solution for solving the disadvantages of decision trees specially if the data is large and has long history which is the case of our research [2]. The research done on similar work was on a period of three years while our aim is to expand it up to 11 years [3]. Another important aspect is that most studies have relayed on same set of data gathered from the customer disregarding the data available with banks. Our study

relied on data available with the banks currently and made use of it to estimate default customers [7]. The paper will also provide a multi-period observation along historical events due to that the type of loan picked which has a term of 15 years. For the parameters, we chose customer characteristics already used along with new variables have not been used before (monthly number of transactions done in bank accounts and monthly average cash flow in bank accounts). Moreover, in order to evaluate our results properly, we will use the standard measures in the field of credit scoring [1][8]. Specifically, the standard measures are average accuracy, type I error, and type II error.

3 Methodology

The aim of this study is to come up with the most appropriate credit rating model to predict households default rate. In order to do so, we made a comparison between different classification model: Logistic Regression, Decision Tree, Linear Support Vector Machines, Bayesian Network, and RUSBoosted.

3.1 Logistic regression

In this paper we will be distinguishing between two classes of creditors, good or bad [9]. For this binary response model, the response variable Y can take one of two set of values Y = 0 if the customer is good, non-defaulter, or Y = 1 if the customer is bad, defaulter. Xs are the columns vector of M explanatory variables, $\pi = P(Y = 1|X)$ is the responses probability and N is the number of observations

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T x \qquad (1)$$

Where α is the intercept and βT is the coefficients

3.2 Decision Tree

In classification decision trees, it starts with a single node then through a binary differences (1,0) results in the most information about the class [10]. Then we repeat the process with the resulting new node until we reach a position to stop. Usually the tree is too large, so we back test it through a cross-validation. The dependent variable Y is categorical, so, by using information theory in measuring how much we know about it from knowing the value of another separate variable A

$$I[Y; A] = \sum_a \Pr(A = a)\, I[Y; A = a] \qquad (2)$$

I [Y ; A = a] is the value of the uncertainty about Y decreases from knowing that A = a given that we go from full population to sample where A = a. Therefore, I [Y ; A] is how much our doubt about Y reduces on average from knowing the value of A.

3.3 Support Vector Machines

From assuming a training set of N {(Xi, Yi)}Ni=1 with input data Xi ε Rn and consistent binary class labels Yi ε {-1+1}, the SVM classifier in Vapnik's theory satisfies the following

$$yi[w^T\varphi(xi) + b] \geq 1, \quad i = 1, \dots, N \quad (3)$$

The non-liner function of φ (.) plots the input space to a high dimensional feature space [4]. In which, the mentioned variations construct a hyperplane WT φ (X) + b = 0 discriminating between two classes. In original weighted space, the following equation is used for the classifier

$$y(x) = sign[w^T\varphi(x) + b] \quad (4)$$

But it is never evaluated in this form where the curved optimization problem could be defined as

$$\min_{w,b,\xi} j\,(w, b, \xi) = \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi i \quad (5)$$

Subject to

$$\begin{cases} yi[w^T\varphi(x_i) + b] \geq 1 - \xi_i, & i = 1, \dots, N \\ \xi_i \geq 0, & i = 1, \dots, N \end{cases} \quad (6)$$

The variables used in ξi are loose variables which are needed to allow the misclassifications to occur in the set of inequalities due to overlying distribution. The first section of the objective function is set to maximize the margin between two classes in the feature space. The second part is set to minimize the misclassification error.

3.4 Bayesian Network

Bayesian Network is a simple and high performance classifier [4]. This classification model works through learning the class condition probability p (Xi|Y) from each input variable Xi I = 1 … n given the class label Y. Then, a new observation is classified by Bayes' rule to calculate the following probability of each class of Y given the vector of observed feature values

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \quad (7)$$

To make things easier, an assumption behind the naïve Bayes classifier is that the features are in theory independent given the class label, therefore,

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y) \quad (8)$$

The probabilities p (Xi|Y) are then estimated through using the frequency counts for the discrete features and a normal based method for the continues features.

3.5 Gradient boosting

Gradient boosting technique, which is an ensemble algorithm founded by [11], is used to calculate the probability of loan default. It relies on incremental minimization of the error term, in which, improves the precision of the prediction function [9]. In trees, and after setting the learner base, every tree calculated then is fit to the 'pseudo residuals', which is the deviation from the median and not from the expectation, from the earlier predictions in order to lower the error in general. Therefore, the following model is used

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + \cdots + \beta_n T_n(x) \qquad (9)$$

G0 is the initial value for the set. T1 … Tn are the trees and β1 … βn are the coefficients for particular tree nodes calculated by the algorithm. To conduct the gradient boosting classifier, a maximum branch size needs to be set. We chose 30 learning cycles and 0.1 learn rate.

4 Data

Our study is based on household customers in Kuwait with a sample of two banks, one conventional and one Islamic. The period chosen is 11 years, from 2008 to 2018, on monthly basis. There are two types of household loans: consumer loans and installment loans.

Consumer Loans: loans for the purpose of personal needs and durable goods with a limit of 15,000 KWD or 15 times the salary (whichever is less).
Installment Loans: loans for the purpose of maintenance or purchase of private residents with a limit of 70,000 KWD.

We chose installment loans exposures given the higher amount granted which gives a higher impact to the economy. For bank 1, the conventional bank, we took a sample of 100,000 customer base, out of which 37,488 have loans (installment and consumer). The number of customers with installment loans was 28,033 with 996 default cases. When we calculate them in terms of observations for machine learning classification, we have 347,977 transactions given that each customer could have more than one Installment loan. For bank 2, the Islamic bank, we took a sample of 100,000 customer base, out of which 21,559 have loans (installment and consumer). The number of customers with installment loans are 15,108 with 1,394 default cases. When we calculate them in term of observations for machine learning classification, we have 249,567 transactions given that each customer could have more than one installment loan.

To calculate the probability of household loans default, we gathered data of the loan's portfolio including outstanding balance, which is the amount left from the loan granted to be paid. The principal amount, which is a part of the monthly total payment to be paid against the principal amount of the loan. The remaining part of the monthly payment is the interest charge of the loan. Defining the default cases are the case when the customer fails to meet his monthly total obligations for three consecutive months

while there is a remaining outstanding balance. This means that the customer is defaulting (bad customers). Hence, the ongoing payments of monthly obligations are considered non-defaulting (good customers).

The parameters, or independent variables, chosen for this study are:

- Credit card exposure
- Income
- Age
- Gender
- Education
- Nationality
- Relationship duration
- Monthly number of transactions done in bank accounts
- Monthly average cash flow in bank accounts
- Number of loans

A limitation with bank 2 is that they did not provide us with nationality information.

5 Results

After running the classification models to predict the probability of default, a comparison of the area under curve (AUC) is done to facilitate which model to choose (Table 1).

Table 1. AUC results

Bank 1	AUC	Bank 2	AUC
Logistic Regression	0.71	Logistic Regression	0.72
Decision Tree	0.7	Decision Tree	0.69
Linear Support Vector Machines	0.45	Linear Support Vector Machines	0.66
Bayesian Network	0.67	Bayesian Network	0.70
RUSBoosted	0.86	RUSBoosted	0.80

Also, Table 2 and 3 is a summary of the confusion matrix output and the performance indicators.

Table 2. Confusion matrix results and model; performance results for bank 1

Bank 1	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	100%	0%	0%	100%	50%	50%	0%
Decision Tree	99%	0%	1%	100%	50%	50%	1%
Linear Support Vector Machines	100%	0%	0%	100%	50%	50%	0%
Bayesian Network	99%	0%	1%	100%	50%	50%	1%
RUSBoosted	84%	74%	16%	26%	79%	24%	16%

Table 3. Confusion matrix results and model; performance results for bank 2

Bank 2	True Positive	True Negative	False Positive	False Negative	Average Accuracy	Type I Error	Type II Error
Logistic Regression	99%	0%	1%	100%	50%	50%	1%
Decision Tree	99%	2%	1%	98%	51%	50%	1%
Linear Support Vector Machines	100%	0%	0%	100%	50%	50%	0%
Bayesian Network	99%	7%	93%	1%	53%	1%	48%
RUSBoosted	67%	78%	22%	33%	73%	33%	25%

From the illustration provided earlier, the RUSBoosted can be the most efficient model for calculating the probability of default due to, as discussed earlier, that our sample set is very large in bank 1 and runs for long period 11 years. This huge data sets shows the lack of existing classification methods and enhance the importance of ensemble models. From an AUC of 0.86 to average accuracy of 79% with lowest value

in type I error (24%), the RUSBoosted is the ideal solution for big data classifications especially in imbalanced data. The increase rate in type II error of 16% could be justified by that RUSBoosted is the only model working for the data set on hand indicating false cases while the rest models are overfitted with zero or 1% values in false positive. An important clarification is that this model is used to estimate the credit default for household's customers in order to classify the customers in to stages for determining the unexpected loss of good customers. Given that we defined the default rate for three consecutive non-payments, as per the provisioning scheme, the high false-negative rates are covered risk wise through the provision charges.

In bank 2, an AUC of 0.80 and average accuracy of 73% with lowest value in type II error, 25%, the RUSBoosted is the ideal solution for big data classifications especially in imbalanced data. The increase rate in type I error of 33% could be justified by, as stated, that RUSBoosted is the only model working for the data set on hand indicating false cases while the rest models are overfitted with zero or 1% values in false positive.

To evaluate our model, a training subsample for bank 1 of 70% (243,584 observations) from total observations of 347,977 has been tested and the following AUC has been calculated to show the accuracy of the model (see Fig. 1). Nevertheless, a training subsample for bank 2 of 70% (174,697 observations) from total observations of 249,567 has been tested and the following AUC has been calculated to show the accuracy of the model (see Fig. 2).

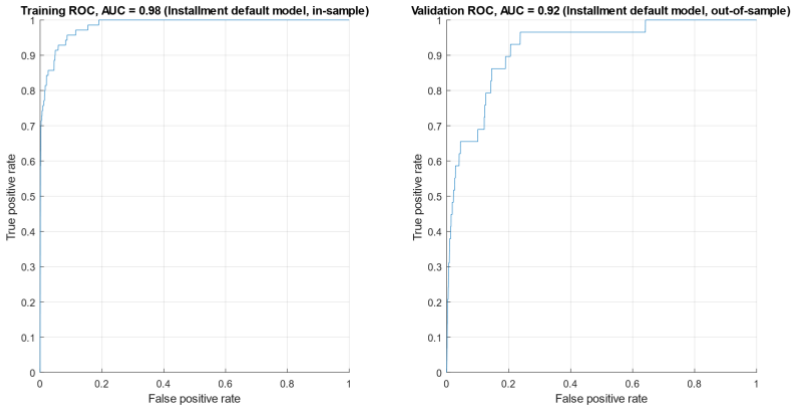

Fig. 1. AUC results for training and validation bank 1

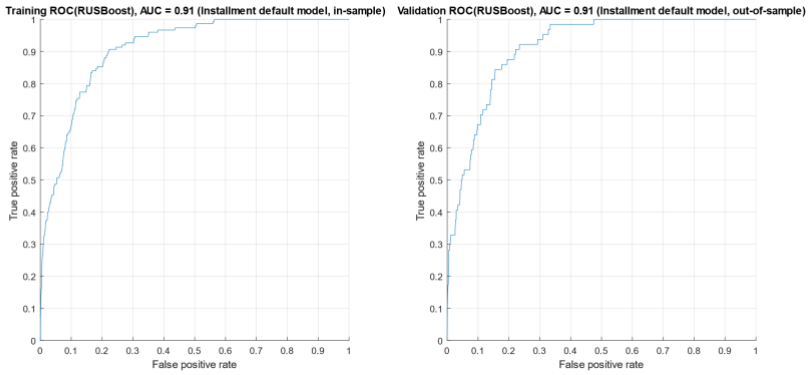

Fig. 2. AUC results for training and validation bank 2

The high number of AUC, from the test and training samples' AUC for both figures, indicate that the True Positive Rate (TPR) is greater than the False Positive Rate (FPR). The TPR corresponds to the proportion of positive data points that are correctly considered as positive while FPR corresponds to the proportion of negative data points that are mistakenly considered as positive. The higher AUC, the fewer positive data points we will miss while the less negative data points will be miss classified as positive which specifies that the model is sufficient.

6 Conclusion

In general, classification methods are increasingly implemented in other fields than computer science. The literature review is full of studies evidencing the efficiency of such models in knowing the expected resulting different classes. Nevertheless, classification methods have been used to categorize credit default classes, good or bad, in order to benefit regulators and bankers to better anticipate risks. This paper compared the different methods of classifications (Logistic Regression, Decision Tree, Linear Support Vector Machines, Bayesian Network and RUSBoosted) in order to examine the credit default cases. Our study relied on big data from Kuwaiti bank for 11 years to tackle the gap of not lengthy data. The parameters also have included new items -- more than what has been used currently. Moreover, those data came from banks indicating the importance of data existing within banks data bases. From the AUC, average accuracy, type I error, and type II error RUSBoosted were chosen as the outperforming method. A supporting result for training sets have indicated the efficiency of the model selected.

Acknowledgement

This work is partially sponsored by the Central Bank of Kuwait

References

1. Wang, G., Ma, J., Huang, L., Xu, K.: Two credit scoring models based on dual strategy ensemble trees. Knowledge-Based Syst. 26, 61–68 (2012). https://doi.org/10.1016/j.knosys.2011.06.020.
2. Bazarbash, M.: FinTech in Financial Inclusion: Machine Learning Applications in Assessing Credit Risk. IMF Work. Pap. 19, 1 (2019). https://doi.org/10.5089/9781498314428.001.
3. Petropoulos, A., Siakoulis, V., Stavroulakis, E., Klamargias, A.: A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. use big data Anal. Artif. Intell. Cent. Bank. 50, 30–31 (2018).
4. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. J. Oper. Res. Soc. 54, 627–635 (2003). https://doi.org/10.1057/palgrave.jors.2601545.
5. Munkhdalai, L., Munkhdalai, T., Namsrai, O.E., Lee, J.Y., Ryu, K.H.: An empirical comparison of machine-learning methods on bank client credit assessments. Sustain. 11, 1–23 (2019). https://doi.org/10.3390/su11030699.
6. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Trans. Syst. Man, Cybern. Part ASystems Humans. 40, 185–197 (2010). https://doi.org/10.1109/TSMCA.2009.2029559.
7. Nyathi, K., Ndlovu, S., Moyo, S., Nyathi, T.: Optimisation of the Linear Probability Model for Credit Risk Management. Int. J. Comput. Inf. Technol. 03, 1340–1345 (2014).
8. Samreen, A., Zaidi, F.B., Sarwar, A.: Design and development of credit scoring model for the commercial banks of Pakistan: forecasting creditworthiness of individual borrowers. Int. J. Bus. Soc. Sci. 2, 1–26 (2013).
9. Brown, I., Mues, C.: An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Syst. Appl. 39, 3446–3453 (2012). https://doi.org/10.1016/j.eswa.2011.09.033.
10. Speybroeck, N.: Classification and regression trees. Int. J. Public Health. 57, 243–246 (2012). https://doi.org/10.1007/s00038-011-0315-z.
11. Friedman, J.H.: Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378 (2002). https://doi.org/10.1016/S0167-9473(01)00065-2.

Using Big Data to compare classification models for household credit rating in Kuwait

Albarrak, Najla

2021-09-10

Albarrak N, Alsanousi H, Moulitsas I, Filippone S. (2021) Using Big Data to compare classification models for household credit rating in Kuwait. In: 6th International Congress on Information and Communication Technology, 25-26 February 2021, London, UK

https://doi.org/10.1007/978-981-16-1781-2_54