Machine Learning-based Empirical Investigation For Credit Scoring In Vietnam's Banking

Khanh Quoc Tran[1,2], Binh Van Duong[1,2], Linh Quang Tran[1,2],
An Le-Hoai Tran[1,2], An Trong Nguyen[1,2], and Kiet Van Nguyen[1,2,⋆]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{18520908, 18520505, 18520997, 18520426, 18520434}@gm.uit.edu.vn,
kietnv@uit.edu.vn

Abstract. In this paper, we aim to develop novel and effective solutions for credit scoring in Vietnam with machine learning models based on our submissions for the Kalapa Credit Score Challenge. We conduct experiments with modern machine learning methods based on ensemble learning models: LightGBM, CatBoost, and Random Forest. Our experimental results are better than single-model algorithms such as Support Vector Machine (SVM) or Logistic Regression. As a result, we achieve the F1-Score of 0.83 (Random Forest) with the sixth place on the leaderboard. Subsequently, we analyze the advantages and disadvantages of the used models, propose suitable measures to use for similar problems in the future, and evaluate the results to select the best model. To the best of our knowledge, this is the first work of the field in Vietnamese banking.

Keywords: Credit scoring · Prediction · Machine Learning · Ensemble Models · Data mining

1 Introduction

Machine learning plays an essential role in all areas of human lives in Industry 4.0. The finance-banking sector is potential, having many aspects to apply machine learning such as: predicting the stock market, classifying customers for banks. In particular, credit scoring is a real problem, which machine learning can effectively solve it.

The latest Fitch Ratings report said "The Covid-19 pandemic led to an increase in overdue debts. These debts threaten Vietnamese banks" income and capital growth - leading many banks face a shortage of capital if economic conditions continue to weaken. In the banking and finance sector, bad debt is always a problem. Bad debt affects financial resources and reduces profits, and consequently, it hinders economic growth. Therefore, minimizing the bad debt ratio is an urgent problem for banks and credit institutions, especially the potential risk of bad debt increased due to the effects of the Covid-19 epidemic.

Credit scoring is a procedure that every credit institution or bank always conducts when customers want to open a new credit card or loan. This is an indicator of whether a borrower can receive a loan or open a credit card. This is an essential step in opening

[⋆] Corresponding author

a credit card or loan because it affects the lender's ability to recover the debt and helps reduce risks of bad debt [1].

Nearly all banks have their method of credit scoring. However, thanks to data science and machine learning development, measuring credit scores of consumers by applying artificial intelligence is an efficient method. This method is a big step in helping banks to evaluate customers accurately, effectively and economically.

Machine learning is a subfield of artificial intelligence involved in researching and building techniques that allow systems to learn from data to solve problems. In credit scoring, the specific issue is to predict credit score of customers. A research dataset which we use is realistic data from Kalapa Credit Scoring Challenge For Students (a competition organized by a technical company), so this dataset is valuable for credit scoring in Vietnam.

Our research aims to build an optimal solution for credit scoring in Vietnam and find an appropriate metric to evaluate our results. With the provided dataset, we conduct processing, building, and evaluating a machine learning tool for labeling good/bad corresponding to the customer's credit score. From the obtained results, we can provide suggestions to assist banks in deciding whether or not to open a credit card or loan for a customer.

This paper focuses on introducing related information about credit scoring problems in the Kalapa Credit Score dataset. In Section 2, we represent some related works. We describe the processing of the dataset in detail in Section 3. In Section 4, the solutions and models are represented. Our experiments and results are given in Section 5. Finally, we conclude the paper in Section 6.

2 Related works

In 2007, Cheng-Lung Huang et al. [8] proposed a solution to use the Support Vector Machines (SVM) model to evaluate a customer's credit score based on two datasets of Australian and German Credit . SVM achieved relatively positive classification accuracy (accuracy of GP, BPN and C4.5 is 88.27%, 87.93% and 87.06% for Australian Credit dataset; 77.34%, 75.51% and 73.17% for the German Credit). Experimental results showed that SVM is a promising addition to existing data mining methods. Although SVM has proven to perform well in the classification process, some still have some inductive deviation.

One effective way to reduce predictive bias is to use ensemble models. Ligang Zhou et al. built synthetic models based on the Least Squares Support Vector Machines (LSSVM) method with an AUC score of 63.95% on UK Credit [13] dataset. The experiment proved to be no significant difference in accuracy and different measurements. Banks can mainly use this model to determine the value of a customer's credit.

Currently, along with the vibrant development of the consumer finance market is the introduction of many personal credit scoring services. There are many factors to consider an individual's credit score, such as income (financial score), loan history (debt score), personal reputation (social score), and identity (who the borrower is, identity card, household registration). Moreover, one of the specific tools built to support solving the above problem is FPT.AI Credit Scoring[1]. FPT.AI Credit Scoring integrates big data and machine learning technology, it can analyze and evaluate credit scores based

[1] FPT.AI Credit Scoring - https://fpt.ai/vi/fptai-credit-scoring-dich-\vu-danh-gia-diem-tin-dung-khach-hang-ca-nhan

on data sources on social networks in Vietnam, with more than 60 million accounts. Another organization that is also very interested in credit scoring in Vietnam, with high applicability topics, is Kalapa. Kalapa has launched the contest Credit Scoring Challenge For Students to create a competition for students to solve real problems. At the same time, this is also an opportunity for Kalapa to find the most feasible credit scoring model to support bank partners and credit institutions.

3 Dataset

This section presents the basics infomation of the dataset and the challenges we faced on the Kalapa Credit Score dataset[2].

3.1 Overview

The original dataset contains customer information including 73,411 data points. There are two labels represent the credit score of customer: label “0” - low credit, label “1” - high credit. Labels are design to facilitate the study of the corelation between information fileds of a user with there credit score

Table 1: Overview statistics of the Kalapa Credit Score Challenge dataset

	Sample size	**#Goods**	**#Bads**	**#Features**
Training	53,030	36, 834	16,196	195
Test	20,381	10,508	9,873	193

3.2 The challenges

This personal credit scoring topic based on this Kalapa Credit dataset is difficult in the initial data processing. The fact that data fields tend to be encrypted to secure customer information is one of the significant challenges to understand and process on the dataset. Therefore, we carry out detailed observations on the dataset to find important information, the basic rules, and the relationship between independent and dependent attributes.

By conducting a survey on the data, we conclude that the challenges posed in the dataset is to find the best solution to this credit scoring problem.

- Imbalanced data: 53,030 data used for training, but only 16,196 bad labels.
- Missing data: up to 117 attributes have a missing data rate$>$ 50%.
- Noise data: several attributes are not normalized (Examples maCv and diaChi), the content contains ambiguous characters/strings (Field _45, 49, 68), the value None accounts for the majority.
- In addition, we also face specific challenges of credit data in Vietnam, such as data is inconsistent, has not been given adequate attention, and database systems are limited, leading to the data is not really big and quality enough. Various properties are encrypted for security reasons, leading to a situation that can make it difficult to understand for our experiment.

[2] Kalapa Credit Scoring Challenge - https://challenge.kalapa.vn/home#gioi-thieu

4 The methodologies

4.1 Preprocessing

We work with a relatively complex dataset with many properties, many data types, and many missing data fields, causing a lack of meaning noise. Therefore, we implement separate data preprocessing methods for each of the above challenges.

First, we remove the attributes with low impact because, after the research, we assess that these are insignificant components in the process of solving the problem, such as: calculating only one uniform value on all data points (Field_13, Field_14, Field_16) has no categorical significance. Some independent attributes have high correlation coefficients, keeping only one attribute in each pair of high correlation attributes (correlation > 0.8 or correlation < -0.8).

Next, we deal with the missing data, divide the attributes of the dataset into three sub-sets based on its data type: categorical, numerical, datetime and replace the missing data such as 'NaN', 'None', 'NULL' equal to mean values (for numerical fields) or by mode values (for categorical, datetime fields).

For attributes with a datetime data type, we progress to: standardize attributes Field_34, ngaySinH and do delta calculation of attributes x_startDate, x_endDate (where x: A, C, E, F, G).

For attributes with a categorical data type, we advance to standardize data in some attributes: Field_38, Field_47, Field_62, and try Count Encoding method.

For attributes with a numerical data type, the Maximum Normalization method is applied to normalize data. Then, K-mean clustering is used to cluster consumer data into different groups. Once the data has been grouped in K-zoning, analytic hierarchy is used to assign credit ratings. Using these credit ratings, employees are classified as very important, important, normal, or bad customers. We calculate delta for some property pairs and compute the mean and standard deviation of the partner_X attributes. Moreover, we create and compare some 'auto columns' from original numerical fields.

We create some meaningful new attributes from the original ones. By researching and extracting information from the original attributes, we create new attributes to serve more information for the classification process. For instance, we create an age attribute from the available birth attribute, the gender attribute from the gioiTinh and info_social_sex attributes,...

As a result, after the pre-processing, we obtain a new dataset with 173 properties (22 properties removed). We find out that while the tuple complexity is reduced, we can still deal with the lack of context created by this normalization. Although the extraneous attributes and confusing data points are most removed to avoid confusing the prediction, the rest may still noise the data.

4.2 Models

In this paper, we choose to experiment on the dataset provided with machine learning methods LightGBM Classifier, CatBoost Classifier, and Random Forest. These models have the advantages in training time (very suitable for the competitive challenge where we have to deliver the result fast), they are easy to apply and also the SOTA and popular models. Then, statistically, we compare results on models to make conclusions and choose the most suitable machine learning model for the Kalapa Credit Score dataset.

4.2.1 LightGBM Classifier:

LightGBM stands for Light Gradient Boosting Machine [6], it is a free and open-source distributed gradient boosting system for AI at first created by Microsoft. It depends on choice tree calculations and is utilized for positioning, order, and other AI assignments. Its features are the presentation and adaptability.

LGBM was used extensively in many winning solutions in machine learning competitions. Comparative tests on public datasets show that LGBM outperforms existing gradient boosting frameworks in both efficiency and accuracy, with significantly lower memory consumption [10].

4.2.2 CatBoost Classifier:

is YanDex's open source and machine learning algorithm[3]. It can operate on various data types such as audio, text, video. The strength of the algorithm is that it produces good results without the need for large amounts of data and the strong support for descriptive data types that lead to business problems.

CatBoost controls an automatic classification of features based on various statistics. We can use CatBoost without explicit preprocessors to convert categories to numbers.

Besides, CatBoost also supports the ability to fine-tune the hyper-parameters to reduce the overfitting risk which makes the model more general [5].

4.2.3 Random Forest:

Random Forest is a machine learning algorithm built on multiple sets of **Decision Tree**. The model's output is based on the aggregate decision on the decision trees it generates with the voting method. **Random Forest** is a **Supervised Learning** method to handle classification and regression problems. Random Forest gives us a very accurate result with such a mechanism, but the trade-off is that we cannot understand how this algorithm works due to the complicated structure of this model. This is one of the Black Box methods - that is, we put our hands inside and get the results, but cannot explain the mechanism of the model [11], [7].

4.3 Features Selection

In this works, we apply a procedure for feature selecture using target permutation [3]. Feature selection process using target permutation tests actual importance significance against the distribution of feature importances when fitted to noise (shuffled target). We implements the following steps:

1. We create the null importances distributions: these are created to fit the model over several runs on a shuffled version of the target. This shows how the model can make sense of a feature irrespective of the target.
2. We fit the model on the original target and gather the feature importances. This gives us a benchmark whose significance can be tested against the Null Importances Distribution.
3. For each feature test, the actual importance:
   - We compute the probabability of the actual importance with the null distribution. We use a very simple estimation using occurences while the article proposes to fit known distribution to the gathered data. In fact that we compute 1 - the proba so that things are in the right order.

[3] https://catboost.ai/

- We simply compare the actual importance to the mean and max of the null importances. This gives sort of a feature importance that allows to see major features in the dataset. Indeed, the previous method may give us lots of ones.

Finally, we decide to select the 33 features that have the most impact (Information Values > 0.2) on prediction. We use them as key attributes to training the machine learning models.The results obtained from the experimental process are presented in the Figue 1, Table 2 and Section 5.3.

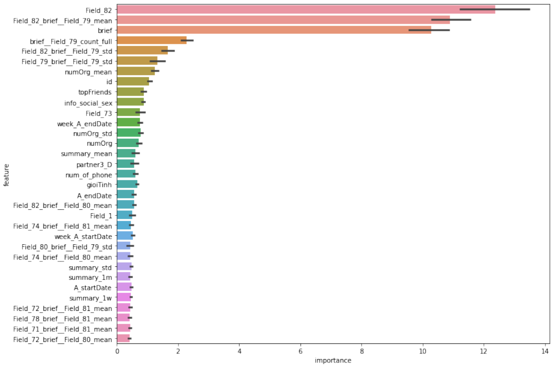


Fig. 1: Top 33 features most influence to dependent variable.

Table 2: Experimental results on Kalapa Credit dataset.

Model	**Original**		**Processed**	
	AUC	**F1-score**	**AUC**	**F1-score**
LightGBM Classifier	0.71	0.69	0.75	0.71
CatBoost Classifier	**0.77**	0.73	**0.82**	0.78
Random Forest	0.74	**0.78**	0.81	**0.83**

5 Experiments

5.1 Data preparation

We implement data preprocessing, as mentioned in section 4. Subsequently, the dataset is splitted into training and test sets. We also take advantage of the Cross-validation method on the training set while training the models. This technical method helps us to train the models on some subsets from training set, so we can directly evaluate the models on the training phase to find out the best models to apply on test set.

5.2 Models implementation and the parameters refinement

The models are all trained on the training set and assessed on the test set. In this paper, we evaluate experiments through three given models: LightGBMClassifier, CatBoostClassifier, and Random Forest. Then, we can also make comparisons of their performances on the dataset.

– **LightGBM**: We refine 3 parameters with specific values num_leaves = 128, learning_rate = 0.02, and max_depth = 8.

– **CatBoost**: With Catboost model we use these parameters: iteractions = 1000, learning_rate = 0.1, and random_seed = 42.

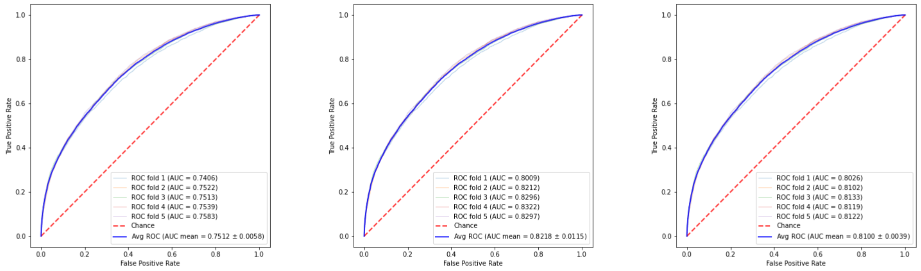
– **Random Forest**: We implement a Random Forest model with max_depth = 17, max_features = ‘auto’, seed = 2020, and n_trees = 767.

5.3 Experimental results

In this Session, the achievements that we obtained from the Kalapa Credit Challenge For Students 2020 contest are visualized. The contest results are based on the Gini score (with Gini = 2 * AUC - 1). Besides, we add one more measures: the F1-score with an effort to find out the appropriate measure for classification problems in general and credit scoring problems in particular [9]. The experiment results on the Kalapa Credit Score dataset are revealed in Table 3. We draw a conclusion that the Random Forest model has the best performance.

Table 3: Assessment values on the Kalapa Credit Score dataset.

Model	**AUC-score**	**F1-score**
LightGBM Classifier	0.75	0.71
CatBoost Classifier	**0.82**	0.78
Random Forest	0.81	**0.83**



(a) LightGBM.

(b) CatBoost.

(c) Random Forest.

Fig. 2: Plot of ROC Curve - AUC.

Table 4 shows the top 5/760 high-performance solutions on public test data from the Kalapa Credit Challenge For Students 2020 contest. As a result, we ranked 5th on the individual standings with a Gini score 0.50013. The score was not far different from the rest. Nevertheless, in the final ranking table, we were ranked 6th place (the_cook_of_the_king) on the team standings.

Table 4: The results of the top 5 on private-test set - Kalapa Credit Challenge[4].

Rank	Team	Gini-score	F1-score
1	bker_team_-_khanh_vu_duy	0.46028	0.75
2	iu_boys	0.45295	0.74
3	ai_beginner	0.44732	0.74
4	carl_friedrich_gauss	0.44455	0.73
5	cainaychoisao	0.44250	0.72
6	**the_cook_of_the_king**	0.44183	0.83

5.4 Results analysis and discussion

5.4.1 ROC Curve and AUC ROC Curve and AUC in this particular problem (or binary classification problems) are usually the crucial measures chosen as a criterion for finding out the best model [12].

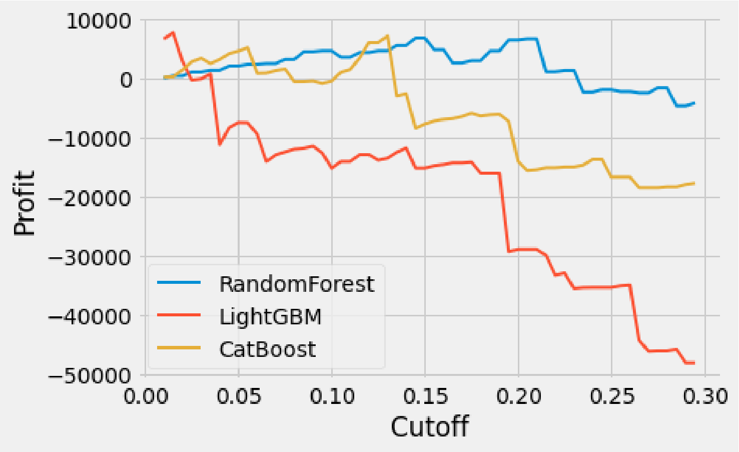


Fig. 3: Comparison of profits predicted by three models by threshold.

Despite this, it is still hard to find out the answer to the practical problem:

[4] Kalapa Credit Score Challenge - https://challenge.kalapa.vn/#bang-xep-hang

- The thing that is concerned by credit agencies and banks is the profit. However, no one can make sure that the model with a high AUC will be suitable in practice. As shown in Table 3, the CatBoost model and Random Forest model's AUC scores are 0.82 and 0.81, respectively. Despite having a higher AUC score, the interest predicted by the Catboost model is less than that of the Random Forest model (Figure 3).
- Regardless of profit increment, the banks also take care of financial risks or profit volatility [4]. Figure 3 shows that the revenue predicted by the CatBoost model starts to fall dramatically (the orange line) with a little bit of different amount from the optimal threshold, even though it has the highest predicted profit. Meanwhile, the value predicted by the Random Forest model is less than that of the CatBoost model, but there is no similar volatility around the optimal threshold. In other words, profit volatility from the CatBoost model is larger, so that the usage of this model raises potential financial risks.

A consequence can be obtained from the above: Though CatBoost results in the highest AUC score, banks or Kalapa companies will not deploy this model for credit scoring problems in practice. It is possible that they can take other models that have less AUC score into account.

5.4.2 Classification results on each label Each class from the Kalapa Credit Score dataset (training set) is immensely imbalanced (class 0: 36234/53030 samples, class 1: 16796/53030 samples). In order to handle this threat, Recall and F1-score are recommended as the most common and efficient measures.

Table 5: The classification result in each class on the test set.

Label	Model					
	LightGBM		CatBoost		Random Forest	
	Recall	F1-score	Recall	F1-score	Recall	F1-score
0	0.95	0.86	0.95	0.87	0.97	0.90
1	0.44	0.57	0.57	0.69	0.66	0.77
Average	0.70	0.71	0.76	0.78	0.81	0.83

As can be seen, from the metric records in Table 3, we can figure out that the three models all classified relatively well on the class 0 (good). On the other hand, we see a contrary result in the class 1 (bad). This can be explained by the imbalance of the dataset (the class 1 just accounts for 30.54%).

According to the experiment results above, the classification accuracy is effective in general because the F1-scores are greater than 70%. Moreover, the Random Forest model has slightly higher accuracy than the rest (83%). In conclusion, Random Forest is an acceptable and flexible solution to optimize the features subset and parameters for credit scoring problems in practice.

6 Conclusion and Future works

In this paper, we introduce scoring credit methods in Vietnamese market by applying some machine learning algorithms. These methods are useful in evaluating customers' credit scores and help save time and money compared to traditional methods. We analyzed essential information about credit scoring and evaluated several ensemble learning experimental results on the Kalapa CreditScore dataset. As we can see, the solution of applying machine learning for credit scoring is good. The best algorithm is Random Forest with AUC, Recall, and F1-score is 0.81, 0.81, and 0.83 respectively.

The LightGBM model can process a large amount of data with little memory, parallel computing, and GPU learning, and this obtains better accuracy, less training time, and is more effective. The CatBoost Classifier performs the best AUC performance with 0.82. This is because Catboost effectively handles categorical features and allows to tune hyperparameters. However, when CatBoost is not in its optimal range, it shows large volatility and low-performance [2]. Finally, Random Forest gains a useful model with optical performance and stability, but its drawback is needing much time and memory to train and process.

For credit score datasets in general and Kalapa CreditScore in particular, data preprocessing, exploratory data analysis, and important feature selection are significant for developing models. They are factors that affect the performance of models. The good handling of essential features can increase the model's results and make the model more accurate and reliable.

In the future, we plan to apply Deep Learning algorithms to find out better models for credit scoring. Training time is also a problem that we are going to improve in the future. With techniques for extracting essential features and hardware development, credit scoring aim to become more accurate and efficient by applying machine mearning.

References

1. MS Irfan Ahmed and P Ramila Rajaleximi. An empirical study on credit scoring and credit scorecard for financial institutions. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 8(7), 2019.
2. Essam Al Daoud. Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6–10, 2019.
3. André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.
4. Show All Code and Hide All Code. Deep learning for credit scoring in the era of big data (adapted from a research conducted by mis, banking academy of vietnam).
5. Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support, 2018.
6. Jerome H Friedman. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4):367–378, 2002.
7. Nazeeh Ghatasheh. Business analytics using random forest trees for credit risk prediction: A comparison study. *International Journal of Advanced Science and Technology*, 72(2014):19–30, 2014.
8. Cheng-Lung Huang, Mu-Chen Chen, and Chieh-Jen Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.
9. László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 245–251. IEEE, 2013.

10. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.
11. Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
12. Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.
13. Ligang Zhou, Kin Keung Lai, and Lean Yu. Least squares support vector machines ensemble models for credit scoring. *Expert Systems with Applications*, 37(1):127–133, 2010.