Application of explainable machine learning based on Catboost in credit scoring

Ji Qi¹, Ruicheng Yang²* and Pucong Wang³

¹ Finance School, Inner Mongolia University of Finance and Economics, Hohhot, Inner Mongolia, 010000, China
² Finance School, Inner Mongolia University of Finance and Economics, Hohhot, Inner Mongolia, 010000, China
³ Finance School, Inner Mongolia University of Finance and Economics, Hohhot, Inner Mongolia, 010000, China

*Corresponding author’s e-mail: yrc@imufe.edu.cn

Abstract. Credit scoring is the core part of an institution's lending. As artificial intelligence is used in various fields, credit rating is also under the same topic of accepting technological changes. Combining credit evaluation and machine learning can incorporate relatively comprehensive features into the credit evaluation process. Through the excellent performance of Catboost, while ensuring accuracy, it demonstrates the explainability of the model as much as possible, avoiding the traditional trust problem of the black-box model. Explainability is proposed to the machine learning model, which reduces the difficulty of processing large amounts of data and the threshold for non-professionals to understand the model. In this article, the dataset is the personal loan data of LendingClub obtained through python. By analyzing the data through Catboost, we can derive excellent results in applying the explainability of machine learning in personal credit evaluation.

1. Introduction

This Credit evaluation is the process of making decisions after evaluating a series of borrower information[1]. Due to the development of the information society, more and more person-related information can be recorded. For borrowing institutions, even in the next decade years, it may be possible to analyze and evaluate personal holographic information records. But for today, there is already a lot of personal-related information that needs to be sorted out and judged urgently. It can’t even be done manually[2]. Besides, the money suppliers also face a new stage in how to manage the risk of a loan in individuals and companies. In the current situation, there are so many applications in processing, but a few bank workers cannot effectively deal with the constantly emerging audit needs[3]. Nowadays, money-lender prefer to use AI as a proper way to help to deal with this dilemma[4]. However, they do not have enough knowledge about computer technology, so they cannot trust the Ai as the fundamental way to audit the money-need[5]. Catboost's superior performance in predictive accuracy and the need for tuning to reduce model adjustment events are all reflected in previous papers[6], so this article will be based on the advantages of Catboost in related work.
2. Methodology
Catboost is an unbiased boosting algorithm with categorical features, also is a powerful machine-learning technique that achieves high accuracy in an array of the area including computer vision, natural language processing, and data mining. Yandex proposed this algorithm in 2017 and compared it with the main two power-boosting algorithms, XGBoost and LightGBM, all of which are the same Gradient Boosting Decision Tree (GBDT) open-source learning algorithm. Given Yandex’s empirical results, Catboost has a tremendous advantage over current in the boosting algorithms. Because Catboost has two new methods that support its state-of-the-art performance, the first is the algorithm to processing classification features, and another is sort lifting algorithm – ordered boosting.

2.1. Catboost Features
Generally boosting just uses one-hot encoding to dealing the categorical features, but if there are high cardinality features, it’s an infeasibly large number in encoding. So Catboost takes an effective and efficient way to cope with this in dividing substitute the category into several numeric features. Usually, it estimates the expected value as the average after sorting. The process is as follows:
- Randomly sort the input sample set, and generate multiple sets of random permutations;
- Given a sequence, for each example, calculate the average sample value for examples of the same category;
- Transform all categorical feature values into numerical results to obtain minimum information loss.

2.2. Ordered Boosting
The traditional GBDT calculates the gradient of the loss function to the current model for the same data set at each iteration and then trains based on this gradient to obtain the base learner, but this method causes the point-by-point gradient to produce an estimation bias, which eventually leads to Over-fitting, the Catboost algorithm proposes to use the Ordered boosting method to change the gradient estimation method in the traditional algorithm, to obtain an unbiased estimation of the gradient, reduce the influence of the estimation deviation, and improve the generalization ability of the model.

To obtain an unbiased estimate of the gradient value on each sample point, Ordered Boosting utilization sorting to generate a random arrangement order and number the data set, and then use all the ordered samples, then modeling estimate the target value, calculating the residual, and adding new data to form a new sequence to continue repeating the above process. But this method needs to train and optimize a model for every sample, so the number of models is the same as the samples, which increases the algorithm's space complexity and memory requirements. At the same time, the different order of the selected samples also has an impact on the prediction results of the model, which improves the instability of the algorithm. Catboost optimizes the algorithm based on the OB and the basic concept of GBDT.

3. Explainability
Artificial intelligence has developed rapidly and has made great progress in various fields, such as medical, law, finance. But although artificial intelligence can provide extremely high accuracy, the meanings behind this accuracy are still unknown. So even with so many advantages, some key steps still cannot replace manual work, is because we don't understand the reason for machine learning predictions, and we don't know why the result generates. Therefore, artificial intelligence applied in those key areas still cannot be fully trusted. Especially in the field of credit rating, the application of artificial intelligence has an extremely high demand for explainability. And how to understand these models to build user's trust in artificial intelligence has become a hot topic in financial technology.

Fintech introduced artificial intelligence very early. But the flaw of the traditional black-box machine-learning model is that it cannot give an explanation behind the prediction results, which leads to the trust problem of the prediction model. So, the most needed machine-learning not only to provide the most accurate predictions of possible probability but also needs to give the reasons for the prediction results. Explainable machine learning is to show the reasons for the prediction results obtained by
machine learning in a way that humans can understand. Moreover, it can lower the threshold for users, and it can also make explainable machine-learning be used in a wider range. Credit rating is an important part of loan issuance. In the field of credit rating, machine learning is a relatively new and relatively important new application. Given the current strict supervision of the loan market and fierce business competition, banks and related financial institutions will take machine learning technology that can accurately analyze large amounts of data as a crucial competitive advantage. Users can intuitively and comprehensively obtain the results of machine learning, which is very important for practitioners.

3.1. SHAP value
Because of this, this paper uses SHAP (SHapley Additive exPlanations) inspired by cooperative game theory to construct an additive explanatory model, and all features are regarded as "contributors". For each prediction sample, the model produces a prediction value, and the SHAP value is the value assigned to each feature in the dataset. Shapley value was first created by Lloyd Shapley, a professor at the University of California, Los Angeles (UCLA), and was mainly used to solve the distribution equilibrium problem in cooperative game theory.

Our existing data set will contain many characteristic variables. From the perspective of game theory, each characteristic variable can be regarded as a player. The prediction results obtained by using this data set to train the model can be seen as the benefits of many players cooperating to complete a project. Shapley value, by considering the contributions made by each player, to fairly distribute the benefits of cooperation.

The traditional feature importance only shows which feature is important, but we don't know how this feature affects the prediction results. The significant advantage of the SHAP value is that SHAP can reflect the influence of the characteristics in each sample, and it also shows the positive and negative effects of the impact.

This article focuses on using SHAP technology to solve the following problems:
- Measure importance through global variables
- Correlation between different characteristics
- Explainable explanations for different results.

4. Experiments
The article uses the historical transaction data on the Lending Club platform from 2017 to 2019, 235,575 pieces of data have been obtained, including 193,836 samples of loans that have been paid on time, and 41,739 samples have defaulted. And the number of these data features is very large, which is relatively complete customer information. To facilitate understanding, this article divides the mentioned features into three categories, namely: the borrower’s basic information, credit information, and borrowing information.

In the dataset, some samples have obvious missing features, such as lack of personal information or the status of holding collateral, as well as duplicate and incorrect data, which are all filtered out. And some redundant features have been eliminated by comparing the characteristics of correlation. After a series of analysis and manual processing, a total of 235,573 data are valid, of which 41,738 are in default, accounting for 15% of the dataset, and there are 193,835 non-default data, accounting for 85% of the dataset. The proportions of the two types are shown in figure 1.

In simulating the model, 80% of the data is used as the training set of the training model, and the remaining data is used as the test set to evaluate the performance of the model.

Based on Catboost's superior data processing capabilities, we retain as much holographic information about a person's borrowing as possible. So, there are a lot of features. Thus, we only list the features that have an obvious influence on the prediction generated by the model. Others are not mentioned features and will not be shown in Table I. below.
4.1. Global Explanations
SHAP provides a perspective to grasp each feature from a global perspective. Figure 2 shows the magnitude of influence of each feature on the output result.

Figure 3 is a summary of the SHAP value of each feature in the entire dataset, and where each dot corresponds to a person in the study. The set of points density represents the SHAP value of a certain feature. The X-axis is the positive or negative impact of each feature's SHAP value on the model's prediction results, while the y-axis is the ranking of the absolute value of the feature values. A positive SHAP value increases the likelihood that the sample will be marked as a customer with a low probability of default, and vice versa. The positive high characteristic value on the X-axis is positively correlated with the dependent variable Y, which is, the SHAP value has a positive influence on the prediction, and it is predicted that there is no default. For example, in the fourth feature (total_rec_late_fee), the higher the sample value, the greater the absolute SHAP value, and therefore the higher the probability of a bad individual. And the sixth feature (annual_inc) is anti-correlation, the higher the value, the less likely it is for default to occur.
Figure 2. Bar chart of the average SHAP value magnitude.

From the figure above, the most relevant information is the loan term personal credit rating, and the next influencing factor is the loan grade.

Figure 3. A set of beeswarm plots about SHAP value for some features.

Figure 4. Dependence plot of the SHAP values of the grade against term.

The dependence plot is also a heuristic to understand the global performance of a specific account in the SHAP value distribution. In figure 4, we can see the relationship between this feature and its SHAP value. To outline which features are most important to the model, we can plot the SHAP value of each feature of each individual. The following figure sorts the elements by the sum of the magnitudes of the SHAP values on all samples and uses the SHAP value to show the distribution of the influence of each element on the model output. The color represents the characteristic value (red means high, blue means low). It can be seen that since the distribution of credit ratings is discrete, the points on the graph are also discretely distributed. Taking the first level as an example, the longer the period, the higher the probability of default, and the blue part corresponds to the low probability of default. Moreover, in longitudinal comparison, the higher the level, the higher the SHAP value is, and it can be seen that the default probability of the high level is also smaller. It can be seen that in the default prediction model, some of the features that we don't think have an impact on the default are shown in the model. However, some characteristics considered to be influential, such as gender, did not affect the judgment of the probability of default.

Figure 5. A waterfall plot about a not-default account by the model.
4.2. Local Explanations
In this subsection, the main discussion is the accuracy of the model and the relationship between each feature. We use some specific customer examples to demonstrate.

Figure 5 is a model process for judging a good customer. We can see that this customer has a probability of 0.924 to be predicted as a good customer with a low probability of default. Although the first relatively large feature tends to move to the right, that is, the customer does not have a fixed job, the trend of combining other characteristics of this customer allows this customer to show good performance.

5. Conclusion
In this article, we introduce explainability into the detection of loan defaults by machine learning. Through a series of sorting and research findings, we simulated this method and apply it to solve practical problems in credit scoring in lending. The empirical results prove that Catboost’s excellent performance has also allowed the application of artificial intelligence in this field to gain wider attention. And in the above experiment, we can also see that some features are not as closely related to the impact of default as people intuitive.

In the part of using explainable machine learning, we focused on using SHAP as a way to explain the model. By linking different features, even global features, we have come up with a reasonable and interpretable model that can be applied to the lending process of banks. And because Catboost does not need complicated tuning to perform well, this also makes the application threshold very low, which is conducive to the wide application of this technology. The relationship between lending institutions and lenders has been optimized by explanations and can help satisfy the transparency requirements in loans.

Acknowledgements
This research was supported by the National Natural Science Foundation of China (No. 71761029), Natural Science Foundation of Inner Mongolia Autonomous Region (No. 2017MS717), Program for Innovative Research Team in Universities of Inner Mongolia Autonomous Region (No. NMGIT1405).

References
[1] Myers, J. H., and Forgy, W. (1963) The development of numerical credit evaluation system . Journal of the American Statistical Association, 58: 799–806.
[2] Abdou, H. A., and Pointon, J. (2011) Credit scoring, statistical techniques and evaluation criteria: a review of the literature. Intell. Sys. Acc. Fin. Mgmt., 18: 59–88.
[3] Yan, J., Yu, W. and Zhao, J. L. (2019) FinBrain: when finance meets AI 2.0. Frontiers Inf Technol Electronic Eng, 20: 914–924.
[4] Zheng, X., Zhu, M., Li, Q., Chen, C. and Yan, Y. (2015) How signaling and search costs affect information asymmetry in P2P lending: the economics of big data. Financial Innovation, 19–30.
[5] Adadi, A. and Berrada, M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6: 52138-52160.
[6] Dorogush, A., Ershov, V. and Gulin, A. (2017) Catboost: gradient boosting with categorical features support. NPIS, 1-7.