Scientific Research Publishing

Open Journal of Applied Sciences, 2019, 9, 761-773
https://www.scirp.org/journal/ojapps
ISSN Online: 2165-3925
ISSN Print: 2165-3917

New Trend in Fintech: Research on Artificial Intelligence Model Interpretability in Financial Fields

Han Yan[1,2], **Sheng Lin**[1,2,3]

[1]International Institute of China Construction Bank, Beijing, China
[2]Postdoctoral Workstation of China Construction Bank, Beijing, China
[3]Guangzhou Digital Finance R&D Business Group of China Construction Bank Fintech, Guangzhou, China
Email: yanhan.zh@ccb.com, linsheng3.zh@ccb.com

How to cite this paper: Yan, H. and Lin, S. (2019) New Trend in Fintech: Research on Artificial Intelligence Model Interpretability in Financial Fields. *Open Journal of Applied Sciences*, **9**, 761-773.
https://doi.org/10.4236/ojapps.2019.910062

Received: September 27, 2019
Accepted: October 21, 2019
Published: October 24, 2019

Abstract

With the development of Fintech, applying artificial intelligence (AI) technologies to the financial field is a general trend. However, there are some inappropriate conditions, for instance, the AI model is always treated as a *black box* and cannot be interpreted. This paper studies the AI model interpretability when the models are applied in the financial field. We analyze the reasons of *black box* problem and explore the effective solutions. We propose a new kind of automatic Regtech tool—LIMER, and put forward policy suggestions, thereby continuously promoting the development of Fintech to a higher level.

Keywords

Fintech, Regtech, AI, Model Interpretability, LIMER

1. Introduction

In recent years, the rapid development of innovative technologies, for instance, artificial intelligence (AI) has had a great influence on the global financial industry. However, the *black box* phenomenon of AI models has also attracted the attention of many international government agencies and financial regulatory authorities. The *black box* phenomenon of AI models is that AI models are extremely complex and cannot be interpreted, which are always treated as a *black box*. Some institutions have emphasized the importance of model interpretability when AI is applied in the financial field.

For example, in report *Big Data Meets Artificial Intelligence—Challenges and Implications for the Regulation of Financial Services* (July 2018) [1], BaFin pointed out that the precondition of applying AI to the financial field is that financial institutions, such as banks, have some methods providing how AI models work and why decisions are made (*i.e.*, model interpretability), thus preventing models from being treated as a pure *black box*.

Moreover, the Financial Stability Board (FSB) issued the report *Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications* in November 2017 and noted that AI and machine learning are likely to bring many challenges to the financial stability [2]. In particular, AI models and machine learning algorithms are extremely complex and generally lacking in interpretability. It is difficult for users to know how these applications will affect the market. They may bring unexpected shocks to the financial stability, even causing systematic risks. At the present stage, on the basis of well and truly evaluating the risks of AI and machine learning with respect to data privacy, network security, etc., AI model interpretability should be constantly improved, and the supervision of applications of AI and machine learning in the financial field should be strengthened.

Financial regulatory authorities in China made similar requests. In April 2018, the People's Bank of China (PBOC), China Banking and Insurance Regulatory Commission (CBIRC), China Securities Regulatory Commission (CSRC) and State Administration of Foreign Exchange (SAFE) jointly issued *Guidance on Regulating the Asset Management Business of Financial Institutions* [3]. It noted that financial institutions should report to the financial regulatory authorities the main parameters of AI models and the logic of asset allocation. In addition, financial institutions should not only establish specific intelligent investment management accounts for investors, but also fully prompt inherent defects and risks of AI algorithms. Moreover, they should know clearly the process of transactions and monitor the trading positions of intelligent investment management accounts.

What's more, Li *et al.* in China Banking and Insurance Regulatory Commission (CBIRC) also pointed out that when financial institutions use intelligent systems to provide intelligent investment management advices, similar risk indicators and trading strategies may lead to the phenomenon of *buy and sell at the same time*, so that *rise and fall at the same time*, thus exacerbating the market fluctuations [4]. According to the classification of Fintech by the Basel Committee on Banking Supervision, *i.e.*, payment, deposit and loan, investment management and market facilities, the regulatory authorities should focus on information disclosure and investor protection in intelligent investment management, *i.e.*, model interpretability.

To sum up, from the perspective of regulatory authorities, mastering the internal mechanism of the AI models applied by financial institutions can better protect the rights and interests of consumers, which is helpful to remove the discriminatory factors in the model design. In addition, the interpretation of the AI models applied by financial institutions enables regulators to control financial risks and maintain the financial market stability.

When AI models are applied in the financial field, if the models are not fully understood while the models make the business decisions, the users may gradually become indifferent to the risks, thus accumulating financial risks. For example, when AI models are applied to the risk management in the pre-loan credit evaluation, if the probability of default of the customer is predicted without understanding the internal mechanism of the models, improper credit scores may be given. In the asset management business, through the traditional way, for instance, either technical analysis or fundamental analysis, investors will be able to know every detail of decision-making. However, when using AI models, *i.e.*, the intelligent investment, the models may provide similar advices to large number of investors. If investors do not understand the reasons behind the models' recommendations, a *buy and sell at the same time* phenomenon may occur, which magnifies the single financial risk. Therefore, the model interpretability has become a major obstacle for the application of AI in the financial field.

Research by the International Institute of China Construction Bank in 2018 found that the application of AI in the financial field has a large imbalance [5]. It is far more applied in loan and asset management than in information provision, payment and other businesses. We believe that one of many factors restricting the balanced application of AI in the financial field is that when the AI model is applied, the internal mechanism of the model is not understood, *i.e.*, the model cannot be interpreted.

The rest of this paper is organized as follows. We first give the reason why the AI models cannot be interpreted in Section 2. In Section3, we formally define the model interpretability. We review the related work of interpreting the AI models in Section 4 and present our solutions in Section 5. In Section 6, we give some policy suggestions and conclude the paper in Section 7.

2. The Reason Why the AI Models Cannot be Interpreted

We believe that the reason why AI models applied in the financial field cannot be interpreted lies in technology precedes the rules. On the one hand, with the rapid development of AI, the models become more and more complex, which leads to the fact that the models cannot be interpreted. On the other hand, the regulatory rules remain unchanged, that is, the regulatory authorities have not introduced relevant policies in time to adapt to the development of technologies. Both factors result in the situation of technology precedes the rules. In this paper, we take the Risk-weighted Assets (RWA) calculation of commercial banks as an example to illustrate the lag of regulatory rules in adapting to the application of AI models.

2.1. AI Is Developing Rapidly

With the rapid development of technologies, the performance of AI models is continuously enhanced, the accuracy of various prediction tasks is constantly improved, and the complexity of the model is also increasingly high. For example, the original perceptron model only had a few parameters, but now the number of parameters in deep neural network (DNN) model can be as high as one million.

When a complex AI model makes a decision, it will not tell the user the logic of the decision-making process. Moreover, taking the neural network as an example, the super-parameters that can be controlled by human, such as learning rate, coefficients of regularization items and the number of hidden layers, cannot interpret the internal mechanism of the model, but only affect the quality of the model's output. As a result, the model is often treated as a *black box*, where we only know the input and output of the model, but not the process running inside the model. This will cause the user of the model is not able to grasp what knowledge the model has learned from the data so as to make the final decision, thus leading to the user's distrust of the model. Model users' end up being forced to abandon more accurate models (such as neural networks) for critical tasks in favor of traditional, simple machine learning or statistical models (such as linear regression and decision trees). Take the AI model for cancer detection as an example, although the deep neural network (DNN) model invented by the AI laboratory in Stanford University can diagnose whether a patient has skin cancer with an accuracy of 91%, doctors using this model dare not diagnose a patient with this terminal disease just based on the judgment result of the model.

2.2. The Regulatory Rules Remain Unchanged

Since the 1990s, with the continuous progress of computer technology, communication technology, financial engineering and the development of the global financial market, the risk modeling methods of international large commercial banks have become increasingly mature and have been applied in credit management, risk pricing, capital allocation and other aspects in a wide range. In order to measure the credit risk more accurately, the Basel Committee issued Basel II in 2006. While maintaining two key elements, *i.e.*, definition of capital and capital adequacy ratio in Basel I unchanged, a comprehensive model method is introduced in Basel II. Especially it permitted commercial banks to use internal ratings-based approach (IRB) to calculate Risk-weighted Assets (RWA), so as to significantly increase the risk sensitivity of capital requirements (Basel III continued similar requirements, just improving the calculation details). Using the model methods to calculate the Risk-weighted Assets (RWA), the internal risk assessment of commercial banks plays a decisive role in the setting of capital requirements. Under the latest Basel regulation framework, capital adequacy ratio regulation includes three basic elements—definition of capital, Risk-weighted Assets (RWA) and capital adequacy ratio requirements. Among them, the calculation of Risk-weighted Assets (RWA) is the technical core of the supervision of capital adequacy ratio and the basis of the whole regulation framework.

According to the rules of calculating Risk-weighted Assets (RWA) in Basel III, for the exposures of sovereign, financial institutions, corporations and retails, the bank should first calculate the correlation (R) and time adjustment factor (b) separately for each single asset in four types of exposures. Here the mainly used indicator for calculation is the probability of default (PD) of assets. Next, the capital requirements (K) of single non-retail risk exposure and retail risk exposure are calculated. Here, the probability of default (PD) of assets should be utilized, and another indicator to be used is the default loss rate (LGD). Finally, the Risk-weighted Assets (RWA) of a single credit risk exposure is calculated, and the indicator to be used here is default risk exposure (EAD). When calculating RWA, the main used indicators for commercial banks using the IRB advanced method, *i.e.*, probability of default (PD), default loss rate (LGD) and default risk exposure (EAD), need to be estimated by the corresponding models developed by commercial banks themselves.

However, in terms of the models used to estimate the above three indicators, currently, the regulators reject the models generated by AI due to the high complexity of these models (they cannot be interpreted), which will bring greater obstacles to the regulators' supervision. Although currently, the regulators require that the model parameters and internal mechanism used to estimate these indicators should be easily understood, the regulators have not given any suggestions on how to make the models be interpretable, nor have they issued any relevant policies to be implemented to adapt to the application of new technologies.

3. What is the Model Interpretability?

At present, there is no unified definition of model interpretability in both academia and industry fields. Therefore, we give its definition based on the relevant research of model interpretability.

3.1. Why Does the Model Need to be Interpreted?

From a broad sense, the necessity for interpretability comes from the fact that human beings do not know enough about a certain problem or task. With respect to the field of AI, although complex AI model, such as deep neural network (DNN) has high expression ability, cooperating with some parameter tuning technologies that can be called as modern alchemy, can achieve high accuracy in many specific tasks. However, for humans, the trained model is just a nonlinear function formula with a pile of seemingly parameters and its results have very high accuracy. We believe the model itself also means knowledge. When using the results of the model, people also want to know what knowledge the model has learned from the data and what insights (expressed in a way that can be understood by human) are behind the final decision.

3.2. The Definition of Model Interpretability

AI model interpretability refers to the interpretation of reasons behind the model decision in a way that human can understand (via images or text), which makes human beings have a full understanding about the model logic (*i.e.*, acquiring the new knowledge), and eliminates the anxiety of uncertainty when using the models.

4. Find Technical Methods for Interpreting the AI Models

In order to solve the problem that AI models applied in the financial fields cannot be interpreted (*i.e.*, the *black box* phenomenon), we try to find and select technical methods—model interpretable methods and discuss which method is more suitable for the financial field. In the report released in July 2018, BaFin pointed out that the problem could be solved, but it did not give specific methods or relevant policy suggestions [1]. Therefore, this paper carries out further research on the model interpretable methods and analyzes the specific methods applicable to the financial field that can be used to interpret the AI models.

Since 2009, in the field of AI, due to the awareness of the importance of model interpretability, many scholars in statistics, informatics and computer science have carried out a wealth of research on this problem. The model interpretable methods proposed in existing studies can be divided into three types, *i.e.*, Hidden Neuron Analysis Methods, Model Mimicking Methods and Local Interpretation Methods.

4.1. Hidden Neuron Analysis Methods

The hidden neuron analysis methods interpret a pre-trained deep neural network by visualizing, revert-mapping or labeling the features that are learned by the hidden neurons. A neural network consists of hierarchical neurons and edges connecting each pair of neurons. According to different inputs, each neuron will get corresponding output through specific activation function, and then in couple with the weight associated with the edge next to this neuron, we can compute the input of the neuron in next layer. Neurons in a neural network can be divided into input layer, output layer and hidden layers according to their different locations.

Yosinski *et al.* (2015) [6] visualized the live activations of the hidden neurons of a ConvNet, and proposed a regularized optimization to produce a qualitatively better visualization. Erhan *et al.* (2009) [7] proposed an activation maximization method and a unit sampling method to visualize the features learned by hidden neurons. Cao *et al.* (2015) [8] visualized a neural network's attention on its target objects by a feedback loop that infers the activation status of the hidden neurons. To understand the features learned by the hidden neurons, Mahendran *et al.* (2015) [9] proposed a general framework that revert-maps the features learned from an image to reconstruct the image. Dosovitskiy *et al.* (2016) [10] performed the same task as Mahendran *et al.* (2015) [9] did by training an up-convolutional neural network. Zhou *et al.* (2017) [11] interpreted a CNN by labeling each hidden neuron with a best-aligned human-understandable semantic concept. However, it is hard to get a golden dataset with accurate and complete labels of all human semantic concepts.

Based on the above studies, the hidden neuron analysis methods provide useful qualitative insights into the properties of each hidden neuron. However, qualitatively analyzing every neuron does not provide much actionable and quantitative interpretation about the overall mechanism of the entire neural network. More importantly, the visualization method has a better interpretable effect on the image data as input, especially the convolutional neural network (CNN). In the financial field, AI models are mostly applied to risk management or asset management business. In relevant scenarios, the application of AI models on image data is not too much. Therefore, this model interpretable method will not show obvious effects in the financial field.

4.2. Model Mimicking Methods

By imitating the classification function of a neural network, the model mimicking methods build a transparent model that is easy to interpret and achieves a high classification accuracy.

Ba *et al.* (2014) [12] proposed a model compression method to train a shallow mimic network using the training instances labeled by one or more deep neural networks. Hinton *et al.* (2015) [13] proposed a distillation method that distills the knowledge of a large neural network by training a relatively smaller network to mimic the prediction probabilities of the original large network. To improve the interpretability of distilled knowledge, Frosst and Hinton (2017) [14] extended the distillation method by training a soft decision tree to mimic the prediction probabilities of a deep neural network. Che *et al.* (2015) [15] proposed a mimic learning method to learn interpretable phenotype features. Wu *et al.* (2018) [16] proposed a tree regularization method that uses a binary decision tree to mimic and regularize the classification function of a deep time-series model.

Based on the above studies, the mimic models built by model mimicking methods are much simpler to interpret than deep neural networks. However, due to the reduced model complexity of a mimic model, there is no guarantee that a simpler shallow model can successfully imitate a deep neural network with a large VC-dimension (Vapnik Chervonenkis Dimension). Thus, there is always a gap between the interpretation of a mimic model and the actual overall mechanism of the target deep neural network.

4.3. Local Interpretation Methods

The local interpretation methods compute and visualize the important features for an input instance by analyzing the predictions of its local perturbations.

Simonyan *et al.* (2013) [17] generated a class-representative image and a class-saliency map for each class of images by computing the gradient of the class score with respect to an input image. Ribeiro *et al.* (2016) [18] proposed LIME to interpret the predictions of any classifier by learning an interpretable model in the local region around the input instance. Zhou *et al.* (2016) [19] [20] proposed CAM to identify discriminative image regions for each class of images using the global average pooling in CNNs. Koh *et al.* (2017) [21] used influence functions to trace a model's prediction and identify the training instances that are the most responsible for the prediction.

Based on the above studies, the local interpretation methods generate an insightful individual interpretation for each input instance. Compared with hidden neuron analysis methods and model mimicking methods, the local interpretation methods have no obvious shortcomings.

4.4. A Brief Summary

The hidden neuron analysis methods, the model mimicking methods and the local interpretation methods have their own advantages and disadvantages (as shown in **Table 1**). In terms of the applicability in the financial field, we believe that the local interpretation methods are most applicable, while the model mimicking methods are not as applicable as local interpretation methods, and the hidden neuron analysis methods are least applicable for the financial field. When AI models are applied in the financial field, the data is not image data in most cases. Therefore, the hidden neuron analysis methods are not suitable for the financial field. If we only want to have a general understanding of the internal mechanism of the model, we can adopt the model mimicking methods, but the interpretation effect is not ideal, because the shallow model cannot completely represent the complex model. Therefore, the model mimicking methods are generally applicable in the financial field. If we want to get the reason behind the corresponding prediction of the model for a specific input instance, we can adopt the local interpretation methods, which can interpret the actual internal mechanism of the complex model with high accuracy and thus is most suitable for the financial field. Therefore, regulators can interpret the AI models applied by commercial banks in their relevant businesses by using the model interpretable scheme based on the local interpretation methods.

5. Solutions to the Black Box Problem of AI Models in Financial Fields

5.1. Model Interpretable Methods that Meet the Existing Regulatory Rules to the Maximum Extent—Taking the Risk-Weighted Asset (RWA) Calculation Process as an Example

Based on the selection of technical means to solve the problem that AI models in financial field cannot be interpreted, we believe that there are feasible technical solutions to enable regulators to accept commercial banks to use AI models for indicators estimation (probability of default (PD), default loss rate (LGD) and default risk exposure (EAD)). Specifically, regulators first build a set of factors that cautiously used when the models are built. All factors in this set are factors

Table 1. Analysis of the applicability of three types of model interpretable methods in the financial field.

Methods	Advantages	Disadvantages	Applicability
Hidden Neuron Analysis Methods	It is straightforward to interpret the details of complex models in visualization forms.	1) Suitable for image data only, not for the financial field. 2) It can only interpret the details of the model, but cannot interpret the overall behavior of the model.	Low
Model Mimicking Methods	The complex models are imitated by shallow model, and the overall behavior of the model is easy to be interpreted.	1) There is a gap between the interpretation of the shallow model and the actual overall mechanism of the complex model. 2) The reasons behind the models make predictions are not given.	Middle
Local Interpretation Methods	The interpretation of a single prediction can interpret the actual internal mechanism of a complex model and give the reason for the prediction of the model with high accuracy.	Compared with above two types of methods, local interpretation methods have no obvious disadvantages.	High

that regulators do not allow commercial banks to adopt when building AI models. Regulators then select specific individual assets from the set of assets that commercial banks use in calculating Risk-weighted Assets (RWA) (sovereign, financial institution, corporate and retail assets). Using the local interpretation methods, regulators can interpret the AI models used by commercial banks to estimate the above three indicators of the asset. If there is no factor cautiously used in the interpretation, the model can be regarded as compliance; otherwise, the commercial bank needs to rectify the model.

5.2. A New Automated Regtech Tool—LIMER

According to the above scheme, we propose a new automatic Regtech tool—LIMER (Local Interpretable Model-agnostic Explanations Regtech) based on LIME algorithm in the local interpretation methods (Ribeiro *et al.* (2016) [18]), and effectively avoid raising the regulatory costs when just adopting the local interpretation methods. The pseudocode and flowchart of LIMER are shown in Figure 1 and Figure 2 respectively.

The time complexity of LIMER is $O(n)$, where n is the number of assets that need to be inspected. Also, LIMER has a good stability and scalability.

6. Policy Suggestions

6.1. The Proposal for Basel III—Explicitly Allowing Commercial Banks to Use AI Models in IRB

Basel III improves the details of risk exposure calculation using the internal

```
Input: the model $f$ to be interpreted (can be any complex model); The collection of assets $R$ when commercial banks calculate risk-weighted assets (RWA)
Output: $\forall x \in R$, when commercial banks use $f$ to make prediction of default probability (PD) of $x$, the interpretation $w$ of $f$
1. $Q \leftarrow build()$; // regulators build a set of factors $Q$ that cautiously used when the models are built
2. $x \leftarrow select(q \in Q)$; // according to some factor $q$ in $Q$, regulators select an individual asset $x$ to be inspected from $R$
3. $x' \leftarrow transform(x)$; // transform $x$ to an interpretable representation $x'$
  for $i \in \{1,2,\dots,N\}$ do
4.     $z_i' \leftarrow sample\_around(x')$; // sample around $x'$, the number of samples is $N$, and the sample after sampling is $z_i'$
5.     $Z \leftarrow Z \cup < z_i', f(z_i), \pi_x(z_i) >$; // restore $z_i'$ to the space of $x$, the representation after restoring is $z_i$, use model $f$ to compute $f(z_i)$, use similarity function $\pi_x$ to compute the similarity $\pi_x(z_i)$ between $z_i$ and $x$, and put this tuple into the set $Z$
  end for
6. $w \leftarrow K\_Lasso(Z,K)$; // use set $Z$ and $K$ (length of interpretation) to train regression model $Lasso$ with $z_i'$ as features, $f(z_i)$ as labels, and take the sample with large coefficient $z_i'$ as interpretation $w$
7. return $w$; // $w$ is the interpretation of $f$ when commercial banks use $f$ to make prediction of default probability (PD) of asset $x$
```

Figure 1. LIMER (pseudocode).

ratings-based approach (IRB), but not clearly point out that the models used to estimate probability of default (PD), default loss rate (LGD) and default risk exposure (EAD) can be AI models. As above stated, when using AI models to calculate the Risk-weighted Assets (RWA), we already have the feasible interpretable methods. Therefore, we suggest that the Basel III explicitly allows commercial banks to use AI models in IRB, and at the same time, starts the study of principle of building the set of factors that cautiously used when the models are built.

6.2. The Proposal for Regulators in China—Use Automated Regtech Tool to Interpret the AI Models Used by Commercial Banks

Based on the above model interpretable methods during the process of Risk-Weighted Assets (RWA) calculation, we suggest that regulators in China should adjust their regulatory policies in time and accept the AI models when commercial banks estimate the corresponding factors. In addition, regulators in China should study how to build the set of factors that cautiously used when the models are built according to the national conditions and use automated Regtech tool to interpret the AI models used by commercial banks. Therefore, they can adapt to the changing of the modeling technologies and commercial banks can easily test and deploy new AI models.

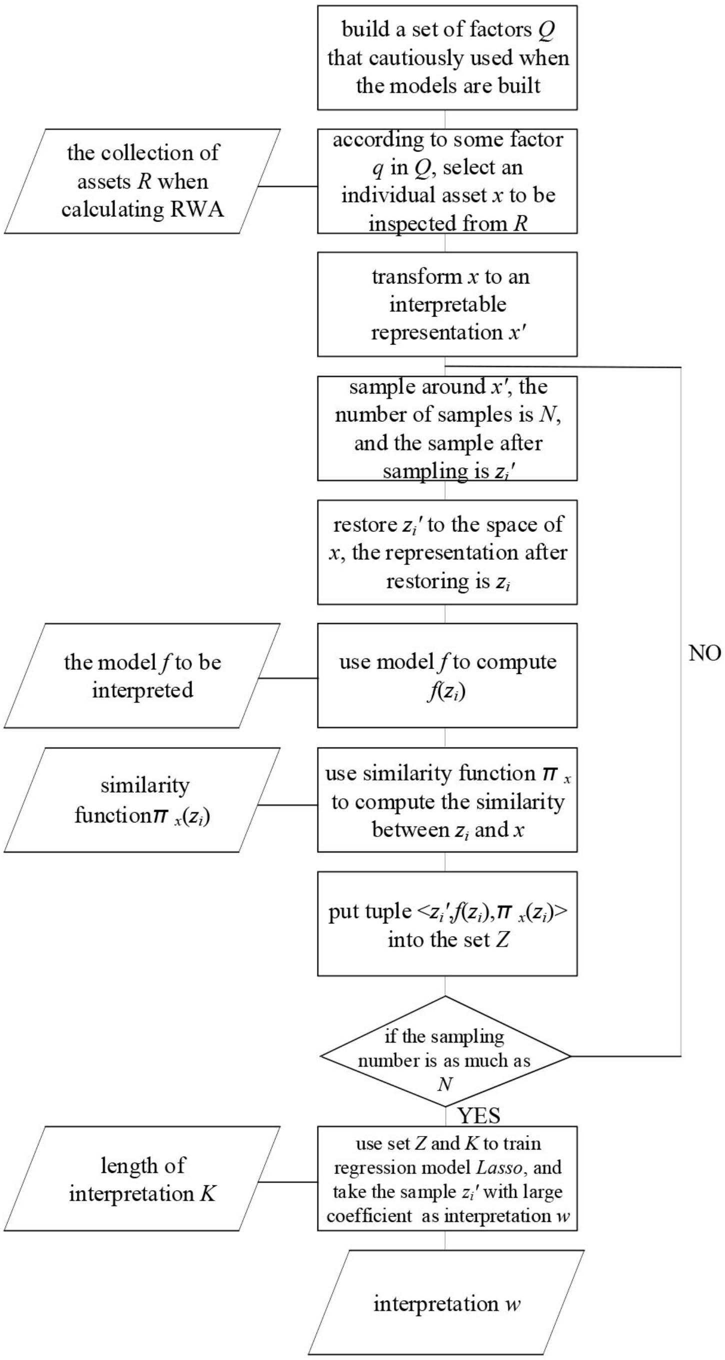
build a set of factors Q that cautiously used when the models are built
the collection of assets R when calculating RWA
according to some factor q in Q, select an individual asset x to be inspected from R
transform x to an interpretable representation x′
sample around x′, the number of samples is N, and the sample after sampling is zi′
restore zi′ to the space of x, the representation after restoring is zi
the model f to be interpreted
use model f to compute f(zi)
NO
similarity functionπ x(zi)
use similarity function π x to compute the similarity between zi and x
put tuple <zi′,f(zi),π x(zi)> into the set Z
if the sampling number is as much as N
YES
length of interpretation K
use set Z and K to train regression model Lasso, and take the sample zi′ with large coefficient as interpretation w
interpretation w

Figure 2. LIMER (flowchart).

6.3. The Proposal for Large Commercial Banks—Adopting Model Interpretable Methods to Continuously Improve the Control of AI Risks

For large commercial banks, we suggest that they should use AI technologies to build new models and use more data sources to evaluate important indicators such as probability of default (PD) and default loss rate (LGD) of assets. At the same time, when large commercial banks design and develop models to be applied in the financial field, the adoption of model interpretable methods is expected to comprehensively improve the control of AI risks. For example, in business areas such as risk control and asset management, model interpretation also has many advantages. First, it can make the model more effective (the process of model interpretation is also the process of knowledge discovery, and commercial banks can use new knowledge to optimize and improve the effects of the model). Second, it can make it easier for commercial banks to meet regulatory requirements. Third, it can protect the practitioners of commercial banks. Fourthly, it can check the model errors caused by mixing data in the data set that will not appear in the actual situation and inconsistency between training data and test data. Therefore, we also suggest that large commercial banks, on the premise of meeting regulatory requirements, actively use model interpretation methods to continuously improve their ability to control risks of AI. Therefore, they can apply AI to business development, and promote the digital transformation of the banking industry with high quality.

7. Conclusion

This paper studies the AI model interpretability when the models are applied in the financial field. We analyze the reasons of *black box* problem and explore the effective solutions. We propose a new kind of automatic Regtech tool—LIMER, and put forward policy suggestions. This work may open up many promising directions for future work. Firstly, it is worth supporting more scenarios when AI models are applied in the financial field. Secondly, it is worth studying more model interpretable methods.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Fin, B. (2018) Big Data Meets Artificial Intelligence—Challenges and Implications for the Supervision and Regulation of Financial Services.

[2] FSB (2017) Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications.

[3] PBOC, CBIRC, CSRC and SAFE (2018) Guidance on Regulating the Asset Management Business of Financial Institutions.

[4] Li, W. and Jiang, Z. (2017) FinTech Development and Regulation: A Regulator's Perspective.

[5] Yan, H. and Bian, P. (2018) The Reviews of AI Applications in Financial Fields. FinTech Research and Evaluation 2018. FinTech Index of Global Systemically Important Banks, 157-176.

[6] Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. and Lipson, H. (2015) Understanding Neural Networks through Deep Visualization. arXiv:1506.06579.

[7] Erhan, D., Bengio, Y., Courville, A. and Vincent, P. (2009) Visualizing Higher-Layer

Features of a Deep Network. University of Montreal, 3.

[8] Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., *et al.* (2015) Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. 2015 *IEEE International Conference on Computer Vision*, Santiago, Chile, 7-13 December 2015, 2956-2964. https://doi.org/10.1109/ICCV.2015.338

[9] Mahendran, A. and Vedaldi, A. (2015) Understanding Deep Image Representations by Inverting Them. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 5188-5196. https://doi.org/10.1109/CVPR.2015.7299155

[10] Dosovitskiy, A. and Brox, T. (2016) Inverting Visual Representations with Convolutional Networks. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27-30 June 2016, 4829-4837. https://doi.org/10.1109/CVPR.2016.522

[11] Zhou, B., Bau, D., Oliva, A. and Torralba, A. (2017) Interpreting Deep Visual Representations via Network Dissection. arXiv:1711.05611.

[12] Ba, J. and Caruana, R. (2014) Do Deep Nets Really Need to Be Deep? NIPS, 2654-2662.

[13] Geoffrey, H., Vinyals, O. and Dean, J. (2015) Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

[14] Frosst, N. and Hinton, G. (2017) Distilling a Neural Network into a Soft Decision Tree. arXiv:1711.09784.

[15] Che, Z., Purushotham, S., Khemani, R. and Liu, Y. (2015) Distilling Knowledge from Deep Networks with Applications to Healthcare Domain. arXiv:1512.03542.

[16] Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V. and Doshi-Velez, F. (2018) Beyond Sparsity: Tree Regularization of Deep Models for Interpretability.

[17] Simonyan, K., Vedaldi, A. and Zisserman, A. (2013) Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034.

[18] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 1135-1144. https://doi.org/10.18653/v1/N16-3020

[19] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. (2016) Learning Deep Features for Discriminative Localization. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27-30 June 2016, 2921-2929. https://doi.org/10.1109/CVPR.2016.319

[20] Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. and Batra, D. (2016) Grad-Cam: Why Did You Say That? Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision*, Venice, Italy, 22-29 October 2017, 618-626. https://doi.org/10.1109/ICCV.2017.74

[21] Koh, P.W. and Liang, P. (2017) Understanding Black-Box Predictions via Influence Functions. arXiv:1703.04730.