Credit Risk Assessment from Combined Bank Records using Federated Learning

Deep Kawa[1], Sunaina Punyani[1], Priya Nayak[1], Arpita Karkera[1], Varshapriya Jyotinagar[2]

[1]Students, Dept. of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai, Maharashtra, 400019
[2]Assistant Professor, Dept. of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai, Maharashtra, 400019

---***---

Abstract - *Credit risk management is an essential activity for financial institutions. Various subjective and quantitative factors are used to predict credit risks. The prevailing advancements in AI/ML has led to the advent of machine learning to generate effective predictive models by using customer data. Better models can be created by using data from multiple institutions. But sharing of data from multiple institutions puts the privacy of customers' data at risk and incurs huge communication costs. In this paper, current credit risk assessment practices are discussed. A federated learning based method to enable the institutions to collaboratively learn a shared prediction model while keeping all the data with the respective institutions itself is introduced. This will decouple the ability to do machine learning from the need to store data in the cloud, thus eliminating centralization.*

Key Words: **Credit Risk, Credit Risk Assessment, Federated Learning, Neural Networks**.

1. INTRODUCTION

Granting loans is the core business part of financial institutions. Revenue is generated by lending money at rates higher than the cost of money lent. This brings the institution at risk from defaulting customers. Default is the failure to meet legal obligations of a loan, i.e., when a borrower fails to repay the loan [1]. The biggest private default in history is Lehman Brothers, with over $600 billion when it filed for bankruptcy in 2008. In India, as of March 2018, the gross Non-Performing Assets (NPA), or bad loans, for Indian banks stood at Rs. 10.36 trillion [2].

In order to prevent possible bank capital losses and possible bankruptcy, an effective credit risk management is crucial for the financial institutions. A credit risk is the risk of default on a debt that may arise from a borrower failing to make required payments [1]. Due to non-repayment of the loan the lender incurs losses. The loss may be complete or partial. In an efficient market, higher levels of credit risk is associated with higher borrowing costs.

To reduce the lender's credit risk, the lender may perform credit check on the prospective borrower. Credit risk assessment is the method by which one calculates the creditworthiness of an individual, business or organization. It helps banks to evaluate if a loan applicant can be defaulter at a later stage.

In the traditional approach, financial institutions consider various quantitative and subjective factors to generate a rating for credit risk assessment. This approach is reactive rather than predictive. There is a requirement, therefore, to develop fairly accurate quantitative prediction models. Prediction models are created considering various internal and external factors.

A machine learning model can yield much better insights from the data than a human analyst. The risk context predominantly determines which model to be used for the analysis. There is no prescriptive method entirely tied to a set of algorithms. Also, the spending patterns of customers are changing and increasing.

It would be desirable for financial institutions to be able to share their data to build better models, since modern machine learning models such as neural networks require huge amounts of data that a single institution might not be able to provide.

To send all data to a central server and apply a machine learning algorithm to it would be a naive approach.This approach is neither often desirable nor feasible.

Thus we propose to use federated learning which is elaborated in the following sections. The section gives a review of the current credit risk assessment techniques starting from the early approaches used to the machine learning approaches widespread in the banking industry. Further, a brief introduction to federated learning is provided, followed by proposal of the mentioned technology for the credit risk analysis system.

2. REVIEW OF CREDIT RISK ASSESSMENT IN FINANCIAL INSTITUTIONS

Prior to making a lending decision the financial institutions evaluate the five C's of credit about a potential borrower. The five C's of credit are Character, Capacity, Capital, Collateral and Conditions [3]. Character is the personal and business reputation of the borrower. Capacity refers to the means in which the borrower will repay the debt. Capital refers to the risk borrower is willing to take. Collateral is the property used to secure the loan. Conditions refers to financial conditions at the time of the loan.

2.1 Early Approaches

The decision of granting credit to a borrower is based on the financial health of the borrower indicated by a rating assigned to them. The rating may be calculated internally or provided by a third party like Moody's, Standard & Poor's for a fee. The drawback of this approach is the subjective aspect which leads to inconsistent estimates.

At present, most institutions employ their own prediction models to grant credit to clients. [4] shows that there are two approaches to create these models - structural and empirical. In structural approach, the modelling is based on the dynamics of the borrower's characteristics such as a firm's assets and balance sheets. Under structural models, a default event is deemed to occur for a borrower when the assets reach a sufficiently low level compared to its liabilities. The main advantage of these models is that they provide an intuitive picture and understanding the reasons behind a default is simple. The limitation of these models is that they are based on assumptions of the borrower's characteristics. In empirical approach, the modelling of relationship between the borrower's characteristics and default is learned from the data. It compares borrower's characteristics with past borrowers' and predicts the outcome. Because empirical models do not explicitly try to capture the economics of default, they are much simpler. But unlike structural models, empirical models are not intuitive.

The following are some borrowers' characteristics considered for credit risk assessment:

- Credit Score
- Age
- Income
- Debts
- Assets

Credit score determines the credit worthiness of an individual. The credit reports and scores of the borrowers are provided to the banks by the credit bureaus like Credit Information Bureau (India) Limited also known as CIBIL, partnered with TransUnion [5]. CIBIL collects and analyses information on credit history of both the existing and prospective borrowers. CIBIL provides a CIBIL report and CIBIL-TransUnion score that takes credit utilization, defaulting/delinquency and trade attributes into account [6].

2.2 Machine Learning Approaches

In recent years, machine learning as become popular in the financial sector. Machine learning can be used in credit risk analysis as with increasing amount of data ML can give better insights compared to humans. Also, it is faster than traditional approaches.

Several models of machine learning have been tried on the given problem, centering one bank, including linear, logistic and multinomial regression by using elastic net approach. Random forest and gradient boosting algorithms have also been successfully tested [7]. Experimentally the random forests work better because they are not constrained to predict linear or continuous relationships. A machine learning model can yield much better insights from the data than a human analyst. The risk context predominantly determines which model to be used for the analysis. There is no prescriptive method entirely tied to a set of algorithms. A report by McKinsey states that machine learning can reduce credit loses by 10% and credit decision times by 20-25% [9] [8]. Also, the spending patterns of customers are changing and increasing. Machine learning helps the lending institutions by decreasing guesswork. AI based scoring models combines customers credit history and the power of big data to improve credit decisions.

Using predictive models make it difficult for institutions to explain the scores to the customers. Several types of risk models have a greater level of transparency as the traditional methods. Gradient Boosting Machines (GBM) are predictive models built from sequence of several decision tree sub-models. The nature of GBM makes it easier than deep learning or neural network algorithms to explain the logic behind the model's predictive behavior. This is because GBMs are represented as sets of decision trees that can be explained as opposed to the neural networks that are represented as cryptic numbers that are much harder to understand.

3. FEDERATED LEARNING

Machine Learning models can be improved by using more data by collecting data across different banks. Centralized collection of data is not a feasible solution because of following reasons:

- Banks would be reluctant to share their privacy sensitive data with others due to business reasons.
- Security and management issues of training data at the center.
- Transfer of large amount of data over expensive or unreliable network is difficult.

With increasing privacy concerns in today's world, centralized storage of data is considered vulnerable for the above cited reasons. Federated Learning is a new approach to machine learning where the training data does not leave the client at all. Weight updates are computed by the client using their locally available data, instead of the data being shared. It is a collaborative form of machine learning where the training process is distributed among many clients. It enables devices to collaboratively learn a shared model by keeping all the training data on the device, without the need of centralization [10][11][12].

Federated learning differs from the standard machine learning techniques in that, the latter has more independent and identically distributed data which is not the case for the former. Devices used for federated learning approach may be intermittently available, may have higher latency, lower throughput. Work has been done to improve the communication efficiency of the technique. The result is the introduction of the Federated Averaging Algorithm which requires 10-100x less communication than the naive federated version of Stochastic Gradient Descent. Also, the uploading costs are reduced by another 100x by compressing the model updates by random rotation and quantization. Federated learning can significantly reduce privacy and security risks by limiting the attack surface to only device rather than device and the cloud.

4. USAGE OF FEDERATED LEARNING IN THE CREDIT RISK ASSESSMENT SCENARIO

One approach to utilizing the data from all the banks for training purposes would be to send the data from all banks to a central server and apply algorithms to it. This is not a viable solution because of the data breach vulnerabilities and other business issues. Collecting data at one place can increase the possibility and extent of attacks.

We propose use of federated learning to perform credit risk assessment. In the proposed design for this system, a central authorized institution plays the role of a central server, coordinating with the federation of clients i.e. bank servers, where most of the work is performed.

To begin with, an initialized model is present on the server. A certain number of clients are randomly selected to improve the model in Federated Learning. We propose to select all the clients in every iteration to improve the model, as the total number of banks in India is not huge and every bank can be taken into consideration to build a system, homogenous for all the banks. The current model is received by each client from the server. Each client trains the model using its local data. All the model updates are then sent back to the server where they are aggregated. The central model is updated with the aggregate and sent to clients for further rounds. This process is repeated till convergence.

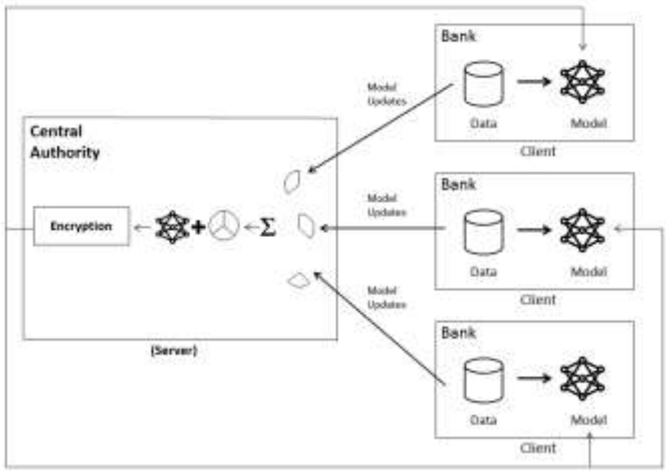

Chart -1: Proposed Model for Loan Default Prediction

3. CONCLUSION

In this work, we have analyzed and presented the current practices of credit risk assessment used by banks. We have also proposed the use of federated learning for learning a credit risk assessment model on data from a large number of banks. The next step towards a real-life implementation is to design a system for prediction of the risk on lines of the idea presented taking into consideration the communication and security aspects.

REFERENCES

1. O'Sullivan, Arthur; Sheffrin, Steven M.: Economics: Principles in Action. Pearson Prentice Hall, Upper Saddle River, New Jersey 07458 (2003)
2. India's bank non-performing loans rise to \$150.2 billion at end-March. Reuters, The Economic Times (2018). https://economictimes.indiatimes.com/markets/stocks/news/indias-bank-non-performing-loans-rise-to-150-2-billion-at-end-march/articleshow/65120728.cms
3. Five C's of Credit. In: Reviewer, Julia Kagan, Investopedia(2018). https://www.investopedia.com/terms/f/five-c-credit.asp
4. Amir F. Atiya, Senior Member, IEEE: Bankruptcy Prediction for Credit Risk Using Neural Networks: A Survey and New Results. IEEE TRANSACTIONS ON NEURAL NETWORKS, Vol-12, Number-4 July (2001)
5. Grant of 'Certificate of Registration' – For carrying on the business of Credit Information – CIBIL, https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=7046\&Mode=0
6. Daisy P.K.: A Study on Credit Information Bureau (India) Limited (CIBIL). ERPA International Journal of Economic and Business Review, Vol-4, Issue-3 March(2016)

7. Thitimanan Damrongsakmethee, Victor-Emil Neagoe: Data Mining and Machine Learning for Financial Analysis. Indian Journal of Science and Technology, Vol-10(39) October (2017)
8. How machine learning can help banks rate credit risk?, https://www.artificialintelligence-news.com/2018/06/13/how-machine-learning-can-help-banks-rate-credit-risk/
9. McKinsey On Risk. McKinsey and Company, Number - 4, January (2018)
10. Federated Learning: Collaborative Machine Learning without Centralized Training Data, https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
11. The Amazing Tech Stack of OpenMined, https://hackernoon.com/the-amazing-tech-stack-of-openmined-a4a0b208d62e
12. Federated-Learning, https://florian.github.io/federated-learning/