The Advancements of Machine Learning Algorithms in Lending Systems for Predicting Lending Behavior

Ruoxuan Zhao[a]
Department of Economics, University of California, Davis, U.S.A.

Keywords: Machine Learning, Deep Learning, Lending Behavior, Financial Systems.

Abstract: Machine learning is indispensable for people to predict the data to make decisions wisely. Especially in the lending system, predicting the lending behavior of customers will affect the decision making of lending platforms or individuals, which in turn will affect the development of the whole economy, so it is worthwhile to pay attention to the application of machine learning to train models to predict the lending behavior. Machine learning workflows typically include data collection and cleaning, model selection and training, evaluation, parameter tuning, and final prediction. Deep learning further enhances this process by using artificial neural networks that employ a hierarchical structure. This paper works on traditional ML-based prediction methods, including decision trees and logistic regression, and compares them to deep learning-based methods such as Long Short-Term Memory (LSTM) networks and Back Propagation (BP) neural networks. Decision trees effectively classify loan applicants by evaluating attributes, while logistic regression models are relatively easy and fast to train. LSTM networks with enhanced attentional mechanisms handle long-term data flexibly, while BP neural networks excel in complex data processing. This comparative study highlights the advantages and applications of various machine learning and deep learning techniques in predictive modeling.

1 INTRODUCTION

The lending industry is a critical component of the global economy, facilitating the flow of capital within the financial system. The ability to predict whether a loan will be good or bad has become an essential aspect of lending operations. Accurate predictions support risk management, financial stability, and operational efficiency while also helping to avoid issues related to asymmetric information. The implementation of advanced predictive models for loans empowers financial institutions to make more informed decisions, ultimately contributing to a stronger and more resilient financial system.

Every enterprise needs capital to assist in its various operational and non-operational activities, and utilizing external funds to solve financial problems is an integral part of an enterprise. Loans from financial institutions such as banks are a key source among the main external sources for enterprises (Imran & Nishat, 2013). The popularity of online lending loans is on the rise, with an increasing number of personal and corporate credit systems available on the internet. In this highly competitive market, some platforms have streamlined their lending requirements to make borrowing money easier for users (Zhu et al., 2023). However, this simplification of procedures has led to an increase in interest rates and late repayment fees, consequently raising the risk of borrowers defaulting on their loans. In order to assist decision-makers in mitigating financial risks, it is crucial to identify the factors that could impact loan repayment (Zhu et al., 2023).

Traditionally, lending decisions have relied heavily on credit scores and historical financial data. However, these methods often fail to capture the complete picture of a borrower's creditworthiness, leading to unadvisable lending decisions. In this situation, the advent of machine learning provides new opportunities to improve predictive accuracy and streamline the lending process. Machine learning algorithms solve the dimensionality problem in empirical studies and can effectively handle large-scale data sets (Ozgur et al., 2021). Moreover, machine learning techniques allow flexible choice of the functional form of the model, simplifying the

[a] https://orcid.org/0009-0002-2317-9819

Zhao, R.
The Advancements of Machine Learning Algorithms in Lending Systems for Predicting Lending Behavior.
DOI: 10.5220/0013263800004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 387-392
ISBN: 978-989-758-726-9

process of variable selection and improving the accuracy of the final predictive model (Ozgur et al., 2021). By utilizing enormous amounts of data and sophisticated algorithms, machine learning models can uncover patterns and insights that traditional methods might miss, enhancing the ability to predict loan behaviors more accurately. In the last decade, many studies have built divergent machine learning models to predict lending behavior. For instance, COSER et al. have presented an analysis of a database of a lending platform, in which a portion of customers were unable to repay their loans and went into default. The authors utilized the methodology of data mining and machine learning algorithms to develop a series of predictive models based on various algorithms such as LightGBM, XGBoost, Logistic Regression, and Random Forest, to assess the customer's loan default probability (COSER et al., 2019). Three sampling schemes were created to compare the classification between unbalanced and balanced datasets, and a comparative analysis of the models was conducted, revealing some interesting insights: customers in default had higher average loan interest rates, slightly lower annual incomes (COSER et al., 2019). Understanding the characteristics of defaulting customers and building predictive models can be a good way to enable lending platforms or banks to make decisions to avoid bad loans.

Many financial experts have built divergent machine learning models to predict lending behavior. This paper will study and compare different professional forecasting models, provide a comprehensive review of the application of machine learning algorithms to build predictive models for lending loan behavior, explore the various factors that influence loan repayment, and discuss the challenges associated with modeling these factors. The report analyzes the machine learning techniques in lending practices, ultimately aiming to find a more efficient model to offer a more secure lending environment.

2 METHOD

2.1 The Workflow of Machine Learning Algorithms

“Machine Learning” was first coined by computer scientist Arthur Samuel in 1959, who defined it as the ability of a computer to learn without being programmed precisely, can process data and experience to detect patterns and learn how to make predictions and recommendations, and also adapt to new data and experiences to improve over time. Machine learning, while encompassing a wide variety of different algorithms, is usually divided into a few clear steps (McKinsey, 2024). First comes the data part, which includes collecting the data and cleaning it (Banoula, 2023). Then comes the model part, including choosing an algorithm, training the model and evaluating it (Banoula, 2023). Finally, there is tuning the parameters and then making predictions (Banoula, 2023). Additionally, some machine learning algorithms specialize in training themselves to detect patterns; this is called deep learning. Deep learning is a more advanced version of machine learning that is particularly good at working with a wider range of data sources, uses a hierarchical algorithmic structure called an artificial neural network, which is based on the way neurons interact in the human brain to ingest and process data through multiple layers of neurons that recognize increasingly complex data features (Pedamkar, 2019). To use deep learning for analytics, it is also divided into several steps similar to traditional machine learning: identifying the data, selecting an algorithm, training the model and testing the model. Finally, a valid predictive model is derived and ready to be applied in daily life (Pedamkar, 2019).

2.2 Traditional Machine Learning-Based Prediction

2.2.1 Prediction Based on Decision Tree

In order to assist lending platforms with customer selection, Sivasree et al. suggest that categorization techniques be used in lending systems to more effectively filter good and poor loans (Sivasree M S & Rekha Sunny T, 2015). Classification methods based on decision trees are frequently used in this manner. A structure with a root node, branch nodes, and leaf nodes is called a decision tree (Sivasree M S & Rekha Sunny T, 2015). Every internal node symbolizes an attribute test, every branch denotes the test's outcome, and every leaf node has a class label. The root node is the highest node in the tree. The authors gather and acquaint themselves with the bank dataset's customer details, which are necessary for data mining (Sivasree M S & Rekha Sunny T, 2015). They next filter the dataset's attributes and choose the pertinent attributes needed for prediction, building an effective decision tree through the use of the decision tree induction algorithm (Sivasree M S & Rekha Sunny T, 2015). The rank of the attributes can be determined by manually adding, using Ranker as a search, and using Information Gain as an attribute evaluator (Sivasree M S & Rekha Sunny T, 2015). It

creates a model with the six most important characteristics. The decision tree's root node is the attribute with rank 1, while the intermediate nodes are made up of attributes with ranks 2 through 6 (Sivasree M S & Rekha Sunny T, 2015). Every node makes a decision, and the leaf nodes give us the outcome in the end. The main advantage of using data mining is that, if a customer meets the minimum loan payback requirement, future risks can be avoided. People can always depend on the algorithm's results to approve or deny a loan application.

2.2.2 Prediction Based on Logistic Regression

Chen et al. developed a model to measure client behavior on Chinese online lending platforms using logistic regression (Chen, 2017). The researchers gathered primary data by collecting information from 3,650 borrowers on a highly regarded Chinese online P2P lending platform (Chen, 2017). They then organized the data into different categories and identified a set of variables that had a strong ability to explain the outcomes of interest using the information gain technique (Chen, 2017). The logistic regression model is the optimal approach for modeling the attributes of explanatory and predictor variables, making it the preferred choice for constructing the predictive model for lending behavior. Variable A was represented as the dependent variable, with a higher information gain value indicating a stronger explanatory power of variable A (Chen, 2017). The variables were further filtered based on the criterion of 0.001 to get the ultimate indicator system. He attributed a distinct value, WOO (weight of evidence), to each group, which might symbolize the impact of each independent variable on the default rate (Chen, 2017). Chen built a regression model using the STATA program and assessed the reliability and precision of the model by comparing its predictions with the observed values of the test samples (Chen, 2017). The findings indicate that the coefficients for the average monthly income, loan interest rate, and average payback rate are all negative and statistically significant at the 0.1 level (Chen, 2017). Conversely, there is a strong and statistically significant correlation between the time it takes for a serious maturity to occur and the rate at which defaults happen, at a significance level of 0.1 (Chen, 2017). Ultimately, the precision of the forecasting model is exhibited by the comparative outcomes. In general, the study presents a robust loan pre-qualification approach for online peer-to-peer lending in China.

2.3 Deep Learning-Based Prediction

2.3.1 Long Short-Term Memory Network

Experts are increasingly using deep learning alongside traditional machine learning techniques to create more precise predictive models. Wang et al. proposed the utilization of the attention mechanism, which is based on the LSTM neural network, for the prediction of loan behaviors (Wang et al., 2019). Their proposed methodology was assessed using authentic datasets (Wang et al., 2019). Given the susceptibility of RNNs (Recurrent Neural Network) to experiencing either a significant increase or decrease in gradient values during training, as well as their difficulty in capturing long-term relationships, the use of an RNN-based LSTM network is proposed in this context (Wang et al., 2019). The LSTM network incorporates a distinctive structural unit and three distinct "gate" configurations. Substances entering or leaving the cell are selectively added or eliminated (Wang et al., 2019). The "gate" structure is established by employing the sigmoid function, which yields values between 0 and 1 to indicate the extent to which information can be let to flow (Wang et al., 2019). Initially, the LSTM unit analyzes the data from the previous memory state using the forgetting gate to ascertain which information should be discarded (Wang et al., 2019). Subsequently, the LSTM unit determines the information to be stored by means of two mechanisms: firstly, the input gate selects the information to be updated; secondly, the candidate vectors are updated through the tanh layer (Wang et al., 2019). The LSTM unit thereafter integrates the aforementioned two components to modify the memory state (Wang et al., 2019). The LSTM unit employs output gates to regulate the memory state that should be produced as output. To thoroughly assess the model's effectiveness and account for the need to evaluate the default probability of its outputs, three commonly used metrics in credit scoring are selected: the ROC (receiver operating characteristic) curve, the AUC (area under the curve), and the KS (Kolmogorov-Smirnov) statistic (Wang et al., 2019). The findings demonstrate that the LSTM-based deep learning model surpasses the conventional manual feature extraction method. Furthermore, the consumer credit scoring system, which incorporates the attention mechanism LSTM, exhibits a substantial enhancement.

2.3.2 Back Propagation (BP) Neural Network

Ma et al. developed a lending behavior evaluation system for online lending platforms, taking into account the peculiarities of online lending and past research (Ma et al., 2021). They then designed a BP neural network using the BP algorithm (Ma et al., 2021). A BP neural network, also known as a multilayer neural network, consists of three or more layers, which include many hidden layers in addition to the input and output layers. The neural network employs an error backpropagation method to facilitate learning, enabling it to extract additional information from input samples and effectively perform intricate data processing tasks (Ma et al., 2021). During the training process of a neural network, the weights of the neurons are adjusted (Ma et al., 2021). This adjustment occurs by propagating the signals in the opposite direction of error reduction (Ma et al., 2021). The weights of the connections between neurons in each layer are modified starting from the output layer and moving through the hidden layer (Ma et al., 2021). The output values, which have been altered through the process of backward propagation of mistakes, are once again linked to the input neurons as inputs for the subsequent computation (Ma et al., 2021). During successive iterations, the neural network's output value progressively diminishes until it reaches a state of stability (Ma et al., 2021). Upon comparing the performance curves of neural networks utilizing the Levenberg-Marquardt (LM) algorithm, Scaled Conjugate Gradient, and Bayesian Regularization as training functions, it is evident that the network model based on the LM algorithm exhibits superior performance (Ma et al., 2021). This is due to the fact that both the number of iterations and the best test error are smaller in comparison to the models based on scaled conjugate gradient and Bayesian regularization. Hence, the LM algorithm is selected as the training technique for the neural network model.

3 DISCUSSIONS

Traditional Machine Learning emphasizes the selection of suitable algorithms to train models based on the features of a given dataset, extending its applicability to a broad spectrum of data predictions. This domain encompasses a variety of algorithms, including Linear Regression, Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, among others. Each algorithm targets specific tasks. For instance, Decision Trees are particularly effective for classification purposes. Machine Learning facilitates predictions and informed decision-making on new, unseen data using existing datasets, playing a crucial role in bolstering the economic sector. However, the rapid advancement of digitalization has highlighted the limitations of traditional Machine Learning, raising concerns among experts. One significant drawback is its limited flexibility in processing high-dimensional data, its generalization capabilities are constrained, making it susceptible to issues like overfitting and underfitting, which can result in inaccurate predictions (Grieve, 2023). In complex environments, such as large-scale lending scenarios, traditional Machine Learning falls short in effectively analyzing and predicting outcomes. In response to these challenges, Deep Learning has emerged as an advanced subset of Machine Learning, gaining widespread adoption among experts. Deep Learning utilizes sophisticated architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to handle high dimensional data sets more effectively (Grieve, 2023). By incorporating multi-layered neural networks and complex nonlinear activation functions, Deep Learning enhances the model's generalization ability and prediction accuracy, thereby addressing the limitations of traditional Machine Learning techniques.

Various techniques have been investigated to improve the usefulness, comprehensibility, and confidentiality of machine learning models in order to maximize the effectiveness of artificial intelligence. An exemplary instance is the advancement of expert systems, which are intricate information systems engineered to emulate the decision-making procedures of human experts in well-defined domains (Waterman, 1985). These systems incorporate expert knowledge and experiential learning to replicate the decision-making processes of professionals. Machine learning poses a considerable obstacle because to its demanding computing requirements, necessitating extensive time and top-of-the-line technology. Expert systems, in contrast, provide a cost-effective solution by diminishing the time and resources required for machine learning activities. Furthermore, the incorporation of expert systems into machine learning can enhance the comprehensibility of models, thereby aligning artificial intelligence more closely with human knowledge systems (Malekipirbazari & Aksakalli, 2015). In addition to expert systems, the SHapley Additive exPlanations (SHAP) method has

been created to tackle the issue of interpretability in machine learning. SHAP values offer a systematic approach to understanding individual predictions by measuring the impact of each attribute on the model's output (Fukas et al., 2022). This technique enables the representation of feature contributions and dependencies, hence enhancing the transparency and comprehensibility of the model's judgments. The transparency of machine learning models is essential for improving their interpretability and comprehensibility. Federated Learning is a possible way to address privacy concerns. Federated Learning, which is based on the theoretical principles of Distributed Database association rule mining techniques, allows models to be trained on numerous decentralized devices or servers without the requirement of sharing raw data (Long et al., 2020). This strategy effectively reduces the likelihood of data leakage or theft, so ensuring the protection of privacy. To tackle the difficulties in the field of machine learning, a comprehensive strategy is needed. This strategy should involve progress in developing algorithms, implementing effective data management procedures, refining interpretability techniques, upgrading computational infrastructure, and enforcing strict ethical and privacy norms. The full realization of machine learning's potential, along with the mitigation of its inherent risks and limitations, can only be achieved through comprehensive efforts.

Despite the continuous evolution of AI technology, transitioning from traditional machine learning to deep learning, the field of machine learning still encounters several significant limitations and challenges. Firstly, the strict requirement of data quality remains a critical concern, lending models require extraordinarily high requirements for data accuracy, generalizability and completeness; subpar data quality such as missing data can result in inaccurate models, while biases in training data can lead to models that perpetuate stereotypes or make unfair decisions. Besides, interpretability is still one of the main puzzles in ML and cannot be fully addressed temporarily. As the algorithmic models become more complex, explaining these models keeps increasingly difficult. Although these models can outperform humans in many specific tasks, understanding their internal workings has become more challenging, posing a significant problem. Many lending behavior predictive models, while giving clear predictions, make it difficult for the average person, other than economics experts, to understand the results given by AI. Furthermore, ethics and privacy represent perhaps the most controversial issues in AI. In particular, the database of a lending system may contain a lot of private information and sensitive information that will affect the whole economic system. Machine learning, which involves the transfer of data, can be intercepted in insecure network environments, leading to data breaches or malicious tampering by hackers. These security and privacy concerns have raised significant questions about the trustworthiness of machine learning systems. In order to utilize machine learning to predict lending behavior better, in addition to developing AI technology, focusing on and addressing the challenges posed by technological developments is inevitable in the future.

4 CONCLUSIONS

This paper provides a comprehensive view of how different machine learning algorithms work to predict lending behavior using datasets from banks and lending platforms. Today there are many economic experts working on machine learning algorithms to try to build predictive models that are as accurate and efficient as possible. From traditional machine learning algorithms (e.g., decision trees) to deep learning (e.g., BP neural networks), different algorithms are suitable for different kinds of data. Overall, deep learning is more efficient than traditional machine learning and can digest and process high-dimensional data more accurately. However, there are still many challenges in the field of machine learning. For example, understandability, applicability, and privacy. The future needs to focus on overcoming these challenges. For example, developing expert systems to improve efficiency, utilizing SHAP to improve understandability, and applying federated learning to reduce the risk of data breaches. The development of machine learning is becoming more and more complex, the emergence of challenges is inevitable, in the focus on the development of technology, but also need to focus on solving the problems brought about by the development, only in this way can individuals better utilize artificial intelligence to promote the development of human society.

REFERENCES

Banoula, M. 2023. The Complete Guide to Machine Learning Steps. Simplilearn.com. https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps

Chen, Y. 2017. Research on the Credit Risk Assessment of Chinese Online Peer-to-peer Lending Borrower on Logistic Regression Model. DEStech Transactions on Environment, Energy and Earth Science, apees.

Coşer, A., Maer-Matei, M. M., & Albu, C. 2019. Predictive Models for Loan Default Risk Assessment. Economic Computation & Economic Cybernetics Studies & Research, 53(2).

Fukas, P., Rebstadt, J., Menzel, L., & Thomas, O. 2022. Towards explainable artificial intelligence in financial fraud detection: Using shapley additive explanations to explore feature importance. In International Conf. on Advanced Information Systems Engineering (pp. 109-126). Cham: Springer International Publishing.

Grieve, P. 2023. Deep learning vs machine learning. Zendesk. https://www.zendesk.co.uk/blog/machine-learning-and-deep-learning/

Long, G., Tan, Y., Jiang, J., & Zhang, C. 2020. Federated Learning for Open Banking. Lecture Notes in Computer Science, 240–254.

Ma, Z., Hou, W., & Zhang, D. 2021. A credit risk assessment model of borrowers in P2P lending based on BP neural network. PLOS ONE, 16(8), e0255216.

Malekipirbazari, M., & Aksakalli, V. 2015. Risk assessment in social lending via random forests. Expert Systems with Applications, 42(10), 4621–4631.

McKinsey. 2024. What is machine learning? Www.mckinsey.com. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-machine-learning

Ozgur, O., Karagol, E. T., & Ozbugday, F. C. 2021. Machine learning approach to drivers of bank lending: evidence from an emerging economy. Financial Innovation, 7(1).

Pedamkar, P. 2019. Deep Learning Technique |Two Phases of Operations in Deep Learning. EDUCBA. https://www.educba.com/deep-learning-technique/

Sivasree M, S., & Rekha Sunny T. 2015. Loan Credibility Prediction System Based on Decision Tree Algorithm. International Journal of Engineering Research And, V4(09).

Wang, C., Han, D., Liu, Q., & Luo, S. 2019. A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM. IEEE Access, 7, 2161–2168.

Waterman, D. A. 1985. A guide to expert systems. Addison-Wesley Longman Publishing Co., Inc.

Zhu, X., Chu, Q., Song, X., Hu, P., & Peng, L. 2023. Explainable prediction of loan default based on machine learning models. Data Science and Management, 6(3).