Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 6
ISET 2024

Analysis of machine learning methods applied to financial problems

Jiayi Zhang

Ji Nan Li Cheng NO.2 High School, Jinan ,China

mumei@ldy.edu.rs

Abstract. With machine learning being widely used in many fields. In this paper, we study how machine learning methods can be used to make predictions in financial problems and analyze the applications.Machine learning is a branch of artificial intelligence that enables computer systems to learn from data and make decisions or predictions without being explicitly programmed at every step. Machine learning models are trained on large amounts of data to discover patterns and relationships in the data and use these patterns to make predictions or perform tasks. The purpose of this post is to illustrate the role machine learning plays in dissertation research.Machine learning can also be used for smart urbanization, industrial automation, fintech, healthcare, agricultural modernization, cybersecurity, etc. For example, extracting valuable information and knowledge from large amounts of data to help businesses and organizations make better decisions. By analyzing historical data, machine learning can predict future trends and behaviors, such as stock prices, house prices, weather, etc.

Keywords: Regression Analysis, Artificial Neural Networks, Decision tree, Random Forest, Deep learning, Ensemble Method.

1. Introduction

With the advancement of technology and increase in data volume, machine learning plays an increasingly important role in modern society. Among them, machine learning includes regression analysis, neural networks, decision trees, random forests, deep learning and combinatorial methods.

At the heart of machine learning is the creation and training of models that recognize patterns and relationships from data and use the information to make predictions or make decisions.

With increasing amounts of data, computing power, and algorithmic advances, machine learning is becoming more powerful and pervasive, with far-reaching impacts across industries.Machine learning is becoming more and more popular because of the explosion in the amount of data. With the proliferation of the Internet, IoT and mobile devices, the amount of data we generate is growing exponentially. This content data provides rich training material for machine learning to learn better and improve .Machine learning technologies have proven their business value across multiple industries, helping organizations improve efficiency, reduce costs, enhance user experience, and create new business models and revenue streams.

In recent years, researchers in the field of machine learning have proposed many new algorithms and models, such as deep learning, reinforcement learning, etc., and these methods have achieved breakthrough results on many tasks and attracted widespread attention.Machine learning is capable of processing and analyzing complex, unstructured data and solving problems that are difficult to handle with traditional methods, such as natural language understanding and image recognition.

2. Regression Analysis

Regression analysis is used in many fields, including economics, biostatistics, marketing, engineering, social sciences, and more. It is a powerful data analysis tool that helps people understand the relationships between variables and make predictions based on those relationships.

Depending on the number and type of independent variables, regression analysis can be classified into several types:

Simple linear regression: this involves one independent variable and one dependent variable, with the aim of finding a straight line that best describes the relationship between these two variables. Mathematically, this straight line is denoted as y = mx + b, where y is the dependent variable, x is the independent variable, m is the slope and b is the intercept.

Multiple Linear Regression: Use multiple linear regression when there are multiple independent variables. The model takes the form y = b0 + b1x1 + b2x2 + ... + bnxn, where b0 is the intercept and b1, b2, ..., bn are the coefficients of each independent variable.

Logistic regression: This is a regression model for classification problems, especially when the dependent variable is binary (yes/no, success/failure). Logistic regression estimates the probability of an event occurring by using a logistic function.

From Machine Learning Based Research on Stock Price Crash Risk Prediction.[1]

The author construct a more accurate system of indicators and research models for predicting the risk of stock price crashes through linear modeling, logistic regression, etc.

In fact, stock price fluctuations are nonlinear, and traditional linear models are difficult to accurately explain and predict stock market changes. Using the Abu quantization system; Stock prediction; Logic linear regression; Quantitative trading; Machine learning to go for prediction.In view of the scholars' research to create a solid theoretical foundation for designing quantitative investment strategy models. In this context, the article tries to explore the linear regression method in supervised machine learning in stock market quantitative trading.Simulation experiments are performed and analyzed. First, a price model is constructed by making appropriate assumptions about the trading environment mentioned in the article. Then, stock price forecasts are made in the assumed price model and the forecast results are derived.

Machine learning is the science of how to empirically improve the performance of computers.The research results and significance of machine learning are multidimensional, and it not only pushes the boundaries of technology, but also brings profound changes to society. With the deepening of research and the expansion of application areas, machine learning will continue to play an important role in promoting social progress and improving the quality of human life. At the same time, there is a need to pay attention to the challenges that machine learning may bring, such as data privacy, algorithmic bias and employment impact, and to seek reasonable solutions.

3. Artificial Neural Networks

Artificial Neural Networks (abbreviated as ANNs), also referred to as Neural Networks (NNs) or called Connection Model, is an algorithmic mathematical model for distributed parallel information processing that mimics the behavioral characteristics of animal neural networks. This type of network relies on the complexity of the system to process information by adjusting the interconnections between a large number of internal nodes.

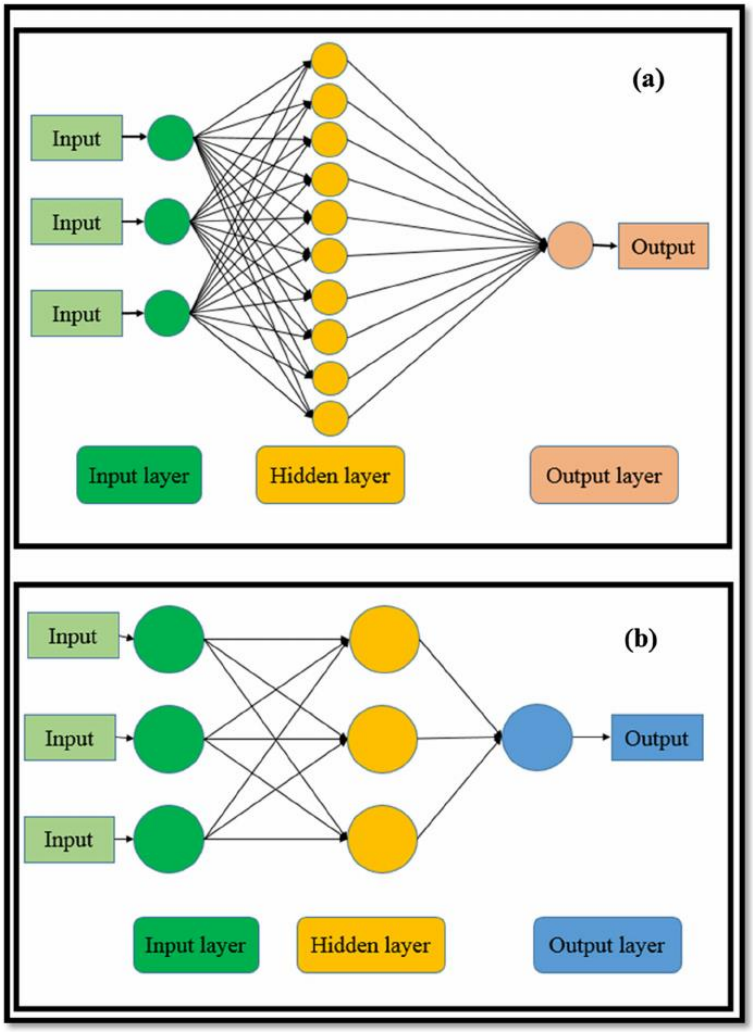

Figure 1 Artificial neural network model diagram afeed forward neural network bradial basis network.

It shows the minimum, maximum and average of sediment and flow in training set and testing set. In training data, the minimum, maximum and average values for sediment are 3353, 5.6 and 192.56 (ton/day) respectively, and for testing are 474.6, 25.3 and 111.4 (ton/day) respectively. Also the table shows the minimum,maximum and average values for flow (FromANN Based Sediment Prediction Model Utilizing Different Input Scenarios by Haitham Abdulmohsin Afan &Ahmed El-Shafie &

Zaher Mundher Yaseen &Mohammed Majeed Hameed &Wan Hanna Melini Wan Mohtar &Aini Hussain).[2]

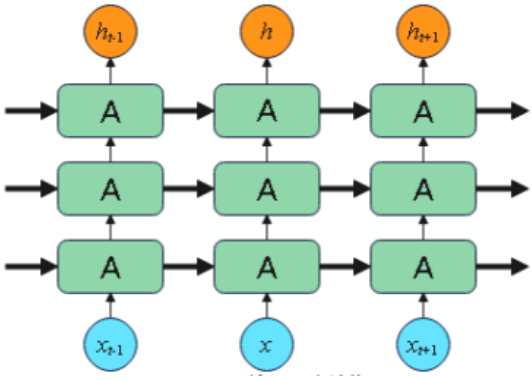

Figure 2 LSTM neural network structure diagram from LSTM neural network based bridge displacement missing data reconstruction method[3]

Online margin trading strategy based on LSTM prediction information .Using several technical indicators as input variables, the LSTM neural network model predicts the upward and downward trends of stock prices. Secondly, the experts who invest in a single stock are considered, and the buying and selling strategies of each expert are constructed based on the prediction results of the LSTM model. [4]Then, a weight optimization model based on the performance of experts is proposed, and the weights of each expert are determined by solving the model. Finally, to illustrate the effectiveness of the constructed strategies, an empirical analysis is conducted using historical stock market trading data. The results show that the constructed strategies are able to outperform the benchmark strategies and maintain their superiority when transaction costs are taken into account.

In order to improve the competitive performance of online portfolio strategies, the stock price prediction technique in quantitative finance is used to construct online portfolio strategies. Li et al. constructed an OLMAR (On-line Moving Average Reversion) strategy based on mean reversion forecasting technique. Huang et al. proposed a Robust Median Reversion (RMR) online portfolio strategy based on robust median regression to improve the stock price prediction technique. Machine learning techniques are able to deal with the nonlinear relationships between complex financial data, and have received extensive attention in the field of financial forecasting in recent years.

4. Decision tree

Decision tree is a commonly used machine learning algorithm for classification and regression tasks. It is a supervised learning algorithm that makes decisions in a tree structure. Here are some basic concepts of decision tree: a decision tree consists of nodes, including root nodes, internal nodes and leaf nodes.Root node is the starting point of the decision tree. Internal node is represents a feature or attribute, which is used to divide the data. Leaf node is represents the final decision result.

When constructing a decision tree, the algorithm automatically selects the best features and division points in order to maximize information gain or minimize impurity. In this way, decision trees can form a hierarchical structure for classifying or regressing new data for prediction.

5. Random Forest

Random Forest is an integrated learning algorithm that consists of multiple decision trees (called "weak learners") and is used for classification and regression tasks. Random Forest improves the generalization and stability of decision trees by introducing randomness in the construction of each tree.

The core idea of Random Forests is "brainstorming", i.e. voting or averaging across multiple trees to improve the accuracy of predictions.

The following are some of the key features of Random Forests.When constructing each tree, Random Forests randomly select a number of features from all the features to be considered for splitting. This number is usually determined by some parameter (e.g. square root or logarithm).

In addition to randomly selecting features, Random Forests will also construct each tree by randomly selecting samples (usually with put-back sampling) from the original dataset.

For classification problems, Random Forests determine the final classification result by majority voting. That is, each tree makes predictions about the data and then chooses the category with the most votes as the final prediction. For regression problems, Random Forest takes the average of the predictions from all trees to get the final prediction. Due to the introduction of randomness, Random Forest can reduce the risk of overfitting and improve the performance of the model on unknown data.

Random Forests can provide an assessment of feature importance, which is useful for understanding the data and controlling model complexit. Although Random Forests consist of multiple trees, they are usually computationally efficient in real-world applications because they can be computed in parallel.

Random forests are widely used in various fields, including data analytics, recommender systems, financial forecasting, etc., due to their strong predictive ability and robustness. It is one of the very popular and effective algorithms in machine learning.

The importance of different influences on the predictor variables varies, but the Random Forest regression model not only predicts safe evacuation times, but also gives importance scores for each variable, and the extent to which inputs influence outputs, so the Random Forest importance analysis can be used to measure the extent to which various influences affect evacuation times. From Crowd evacuation prediction and influencing factors analysis based on random forest.

Random forests can use multiple decision trees to aggregate predictions. By averaging or voting on the predictions of these trees, random forests reduce the relative variation of individual trees, resulting in more accurate predictions (ravinderkamatw .15 Feb, 2024) .[5]

Although random forests are more efficient to train, their pre-test volumes are longer than some other algorithms (ravinderkamatw .15 Feb, 2024).[6]

6. Deep learning

Deep learning is a branch of machine learning that uses artificial neural networks (ANNs) to model and solve complex problems. It is based on the idea of building multi-layered artificial neural networks (called deep neural networks) that can learn hierarchical representations of data.

Here are some key features of deep learning: Deep learning models usually have a multi-layer structure where the number of hidden layers can be large. This structure allows the model to learn more abstract and complex features. Unlike traditional machine learning algorithms, deep learning models are able to automatically learn features of the data without the need for manual feature engineering. Deep learning models typically have a large number of parameters (i.e., weights and biases) that are tuned in a data-driven manner. And also require significant computational resources for training, especially when dealing with large-scale datasets and complex models.

Deep learning has achieved remarkable success in a variety of areas such as image recognition, speech recognition, natural language processing, medical

There are somg challenges .Deep learning models require large amounts of labelled data and the training process can take a long time. In addition, deep learning models are also prone to overfitting and the models are not as explanatory as some traditional machine learning algorithms.

Deep learning is one of the most active and promising directions in the field of artificial intelligence today, and its development provides effective solutions to many practical problems.

Deep learning algorithms do not require hand-crafted features; they can automatically learn features from data, which is particularly useful for tasks where features are difficult to define, such as image recognition（akashmomale 30 Jan, 2023）.[7]

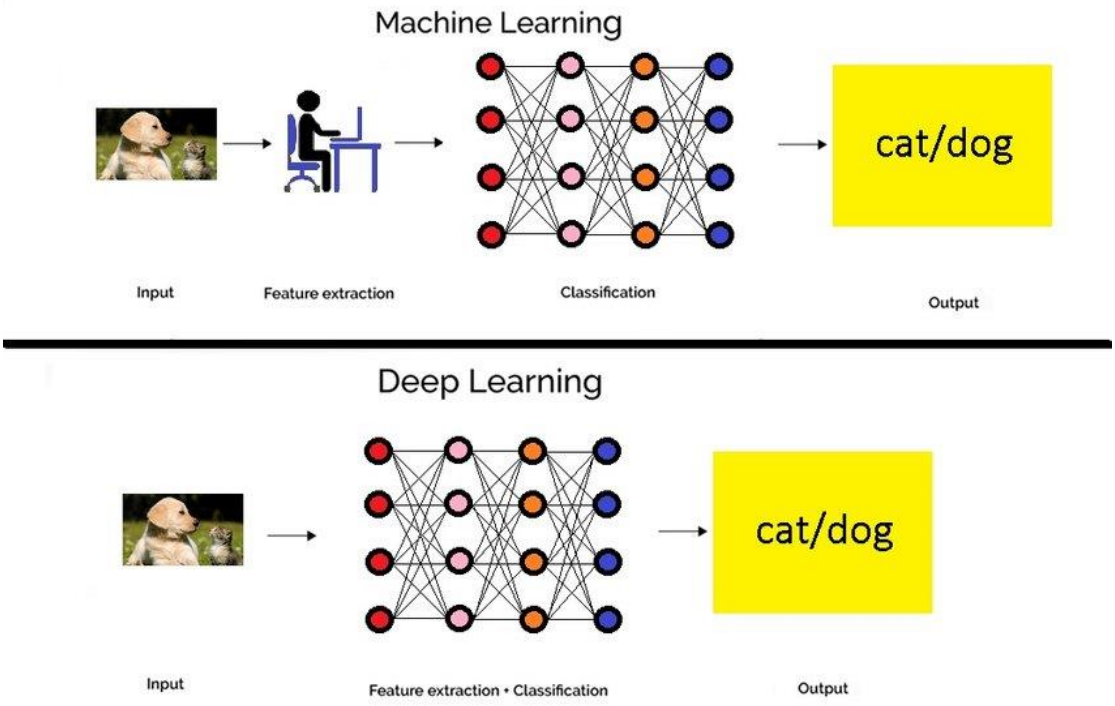

Figure 3 choose the right classifier based on the handcrafted features, shown in these figure. But in DL, we simply train the model end-to-end with the images and their labels. The suitable features are automatically learned by the model (Deep Learning based Smart Attendance Monitoring System by Rohit Haldar, Rajdeep Chatterjee*, Debarshi Kumar Sanyal, and Pradeep Kumar Mallick).[8]

7. Ensemble Methods

Ensemble Methods is a machine learning technique that improves predictive performance and robustness by combining multiple models. The core idea of the ensemble method is that multiple relatively weak models can be combined to form a more powerful and accurate model. This approach is based on several principles:

Diversity: combinatorial methods typically use different algorithms or different training sets to build models to ensure diversity among models. Diversity reduces the bias of individual models and improves the robustness of the overall model.

For classification problems, combinatorial methods typically use a voting mechanism to determine the final prediction. For example, in majority voting, the predicted outcome is the category agreed upon by the majority of the models. For regression problems, combinatorial methods usually obtain the final prediction by averaging the predictions of all models. Combination methods can improve the accuracy of predictions by reducing the risk of overfitting and improving the performance of the model on unknown data. By combining multiple models, combinatorial methods can better capture the complexity and non-linear relationships in the data, thus improving the generalization ability of the model.

Common combination methods include:

Bagging: Random Forest, for example, builds multiple trees by resampling the training data and using different subsets of features, and then voting or averaging to determine the final prediction.

Boosting: such as AdaBoost, XGBoost, and LightGBM, these algorithms train models iteratively, focusing on misclassified samples from the previous round in each round, and then combining the predictions of multiple models through weighted voting or weighted averaging.

Stacking: Stacking is a hierarchical combination method that first trains multiple different models and then uses the outputs of these models as features to train a new model to get the final prediction.

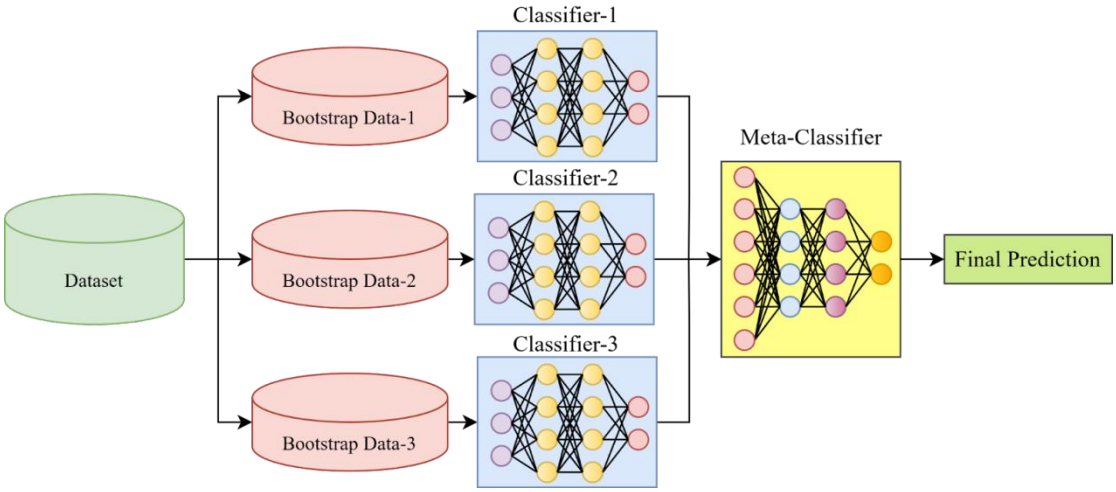

Figure 4 Photo The Complete Guide to Ensemble Learning by Rohit Kundu.[9]

The Bagging ensemble technique is the acronym for "bootstrap aggregating "and is one of the earliest ensemble methods proposed. For this method, subsamples from a dataset are created and they are called "bootstrap sampling." To put it simply, random subsets of a dataset are created using replacement, meaning that the same data point may be present in several subsets. The image shown above exemplifies the Bagging ensemble mechanism.

Evaluation of geohazard risk can be effectively carried out using a ensemble methods in Multi-scale Geologic Hazard Risk Evaluation Based on Combined Methods.[10]

In regression analysis, there are some advantages. It is Simple and easy to understand and explain. It also can handle linear relationships and widely used in predictive analytics. However,it also has some disadvantages. Limited ability to handle non-linear relationships and sensitive to outliers. Assumes that the data satisfies linearity, independence, homoscedasticity, etc.

In Neural Networks

The advantage is capable of capturing non-linear relationships.and have strong adaptive learning capability. It is excellent in image recognition, speech recognition, etc.

The disadvantages are training process can take a lot of time and resources. Models are usually "black box" and not very explanatory. Easily over fitted, requires regularisation and a lot of data.

In Decision Trees

The Advantage is models are easy to understand and highly interpretable.Can handle classification and regression problems.Doesn't require much data preprocessing and doesn't need feature scaling.

The disadvantage easily over fitted, requires pruning. May not handle continuous features well. May favour most classes for data with unbalanced classes.

8. Conclusion

The purpose of this paper is to explore how machine learning is used in finance. Machine learning models can analyze historical and market data to help financial institutions better identify and assess risk, including credit risk, market risk and operational risk. By analyzing transaction patterns and user behavior, machine learning systems are able to identify potential fraud in real time, thereby reducing fraud losses. Machine learning algorithms automate trading strategies to capitalize on market opportunities and improve trading efficiency and profitability. Based on a customer's investment history, risk appetite and financial goals, machine learning can provide personalized investment advice and asset allocation solutions. Machine learning can analyze large volumes of financial

statements, news articles, and social media content to provide investors with insights and investment opportunities.

The application of machine learning in the financial industry not only improves efficiency and accuracy, but also helps financial institutions develop new products and services, enhancing competitiveness and customer experience. With the continuous progress of technology, the application of machine learning in the financial field will be more in-depth and extensive.

The effectiveness of machine learning models depends heavily on the quality and quantity of data. In the financial sector, obtaining high-quality, comprehensive data can be a challenge, especially when it comes to privacy and sensitive information.

Most machine learning models currently require large amounts of data for training. Future research may explore ways to allow models to learn and make accurate predictions even with only a small number of samples.

References

[1] Machine Learning Based Research on Stock Price Crash Risk Prediction.

[2] ANN Based Sediment Prediction Model Utilizing Different Input Scenarios by Haitham Abdulmohsin Afan &Ahmed El-Shafie & Zaher Mundher Yaseen &Mohammed Majeed Hameed &Wan Hanna Melini Wan Mohtar &Aini Hussain.

[3] LSTM neural network structure diagram from LSTM neural network based bridge displacement missing data reconstruction method.

[4] Online margin trading strategy based on LSTM prediction information .

[5] ravinderkamatw .15 Feb, 2024（1）

[6] ravinderkamatw .15 Feb, 2024（2）

[7] akashmomale 30 Jan, 2023

[8] Deep Learning based Smart Attendance Monitoring System by Rohit Haldar, Rajdeep Chatterjee*, Debarshi Kumar Sanyal, and Pradeep Kumar Mallick

[9] The Complete Guide to Ensemble Learning by Rohit Kundu.

[10] Multi-scale Geologic Hazard Risk Evaluation Based on Combined Methods.