How to use machine learning in finance

Mestiri, Sami

Université de Monastir

October 2023

How to use machine learning in finance

Sami Mestiri [1]
Applied Economics and Simulation
Faculty of Management and Economic Sciences of Mahdia,
University of Monastir, Tunisia. Rue Ibn Sina Hiboun, Mahdia Tunisia

Abstract: In the last years, the financial sector has seen an increase in the use of machine learning models in banking and insurance contexts. Advanced analytic teams in the financial community are implementing these models regularly. In this paper, i present the different Machine Learning techniques used, and provide some suggestions on the choice of methods in financial applications. We refer the reader to the R packages that can be used to compute the Machine learning methods
JEL codes: C45, G00
Keywords : Financial applications; Machine learning ; R software.

1 Introduction

Machine learning (ML) is an application of Artificial Intelligence (AI) that allows systems to learn and improve from experience without being explicitly programmed. In effect, it is about developing predictive models that can access data and use it to learn on their own. There are several types of learning, we distinguish:
Supervised learning: is done using a truth, that is, we have prior knowledge of what the output values for our samples should be. Therefore, the goal of this type of learning is to learn a function that given a sample of data and the desired results, in order to best approximate the relationship between observable inputs and outputs. There are two types of supervised learning. Classification algorithms which seek to predict a class/category and Regression algorithms which seek to predict a continuous value.
Unsupervised learning: aims to data structure inference. The two most common subcategories in unsupervised learning are clustering and dimensionality reduction. In clustering observations are grouped in such a method as to produce high intra-group similarity and low inter-group similarity. The different types of clustering methods that have been proposed are entropy-based, density-based and distribution-based methods. Reduction of dimensionality aims to increase the information density of the data by reducing their dimensionality while retaining most of the inherent information. There are different techniques based on principal component analysis (PCA) which derive linear combinations of the original variables to cover as much of the variance in the data as possible. Second, neural network-based methods reduce dimensionality with particular architectures.

AI is increasingly entering our daily lives with impressive applications. This article discusses the use of ML to solve problems in finance research. The contribution of this

[1] https://orcid.org/0000-0002-2060-3242

article is threefold. First, we provide an introduction to Machine Learning. Next we pay particular attention to the different R package implemented (see Mestiri.S (2019) [23]). We build a taxonomy of current and future ML applications in finance. Finally, we study the prospects of ML applications in finance. The research paper is organized as follows: Section 2 presents the different Machine Learning techniques used. In section 3, we present a taxonomy of existing ML applications. The fourth section is devoted to limitation and perspective. Finally, we conclude in section 5.

2 Machine learning techniques

2.1 Linear Discriminant Analysis (LDA)

Ronald Fisher (1933)[10] pioneered work on discriminant analysis. In his work, he developed a statistical technique for defaults prediction, by developing a linear combination of quantitative predictor variables. This linear combination of descriptors is called discriminant function. The output of ADL is a score that is consists of classify a data observation between the good and bad classes.

$$Score = \sum_{i=0}^{p} a_i X_i \tag{1}$$

Where a_i are the weights associated with the quantitative input variables X_i.

The *lda* function from the **MASS** library (Venables and Ripley, 2002)[33] have been used to implement the discriminant analysis as follows:
lda_ mod <- lda(Y∼.,training_ data)

2.2 Logistic Regression (LR)

Logistic regression is a statistical method used for binary classification tasks (e.g. 0 or 1, bad or good, health or default, etc). Corresponding to Ohlson (1980) [26], the outcome of LR model can be written as:

$$P(y = 1|X) = sigmoid(z) = \frac{1}{1 + exp(-z)} \tag{2}$$

where $P(y = 1|X)$ is the probability of y being 1, given the input variables X, z is a linear combination of X: $z = a_0 + a_1X_1 + a_2X_2 + .. + a_pX_p$
where a_0 is the intercept term, $a_1, a_2, ..., a_p$ are the weights, and $X_1, X_2, ..., X_p$ are the input variables.
The *glm* functions from the stats library (R Core Team, 2016) have been utilized for the estimation of the Logit.
logit_ mod <- glm(Y ∼ .,family=binomial, data = Training_ data)

2.3 Decision Trees (DT)

Decision trees are typically not formulated in terms of mathematical equations, but rather as a sequence of logical rules that describe how the input variables are used to predict the output variable. However, the splitting criterion used to select the best split at each decision node can be expressed mathematically. The Gini impurity measures the probability of misclassifying an observation in S if we randomly assign it to a class based on the proportion of observations in each class (Gelfandet al., 1991) [12]. A small value of G(S) indicates that the observations in S are well-separated by the input variables. The split with the smallest value of ΔG is chosen as the best split. The decision tree algorithm proceeds recursively, splitting the data at each decision node based on the best split, until a stopping criterion is met, such as reaching a maximum depth or minimum number of observations at a leaf node.

The following R script runs the *rpart* function from the **rpart** package (Therneau TM, Atkinson EJ, 1997.)[25], used for the Decision Trees model:
DT_ Mod <- rpart(formula = Y∼ . , data = training_ data, method = "class", parms = list(loss = , nrow =)))

2.4 Support Vector Machine (SVM)

Support vector machine (SVM), developed by Vapnik (1998) [32], is a supervised learning algorithm used for classification, regression, and outlier detection. The basic idea of this technique is to find the best separating hyperplane between the two classes in a given dataset. The mathematical formulation of SVM can be divided into two parts: the optimization problem and the decision function.

The decision function takes an input vector x and returns its predicted class label based on whether the output of the hyperplane is positive or negative. The details of the optimization process are discussed in (Chang and Lin, 2004 [6]; Cristianini and Shawe-Taylor, 2000 [8]; Gunn, 1998) [17].

Thereafter, SVM finds the best separating hyperplane by solving an optimization problem that maximizes the margin between the two classes, subject to constraints that ensure all data points are correctly classified with a margin of at least $1-\xi_i$. The decision function then predicts the class label of new data points based on the output of the hyperplane.

The *svm* function from the **e1071** library available on CRAN has been employed (Karatzoglou et al., 2004)[19]
svm_ mod <- svm(as.factor(Y) ∼ ., data=training_ data, , cost = 10, gamma = 1/length(data), probability = TRUE)

2.5 Random Forests (RF)

Random Forest is an ensemble of learning algorithm developed by Breiman (2001) [5]. It is a type of ensemble learning method that combines multiple decision trees for making predictions. The algorithm is called "random" because it uses random subsets of the features and random samples of the data to build the individual decision trees. The data is split into training and testing sets. The training set is used to build the model, and the testing set is used to evaluate its performance. At each node of a decision tree, the algorithm selects a random subset of the features to consider when making a split. This helps to reduce overfitting and increase the diversity of the individual decision trees.

The following R script runs the randomForest function from the **randomForest** package (Liaw and Wiener(2007).)[18]

RF_ mod <- randomForest(as.factor(Y) ∼., data = Training_ data, mtry=ncol(data)-1,ntree=1000)

3 Applications in finance

A significant and growing part of economists is moving towards using the tools offered by ML to conduct innovative empirical analyses. The reason is twofold. ML has allowed economists to use new databases (multidimensional, images, texts) which until then remained unusable with traditional methods; it also opened the way for exploring new problems important to the discipline, notably problems where the prediction of an event is the main research question.

Thus, ML can be understood as a methodological but also conceptual advance in the discipline. It broadens the deductive approach in economics, It now also proposes to explore fields of research where we let the data speak in order to predict certain processes. In this sense, ML poses today as the best way to listen carefully to what the data has to tell us. ML therefore adds to the economist's toolbox not only to exploit new data and incorporate new methods, but also, ultimately, to address and solve new problems.

3.1 Construction of higher and new measures

Researchers can use ML to construct superior and new measures, studies of this archetype use ML to extract information from unconventional high dimensional data such as text, images or videos and construct a numerical measure of 'an economic variable. Using higher metrics reduces attenuation bias, leading to more accurate estimates. New measures allow new analyzes with previously unmeasurable economic aspects.

3.2 Algorithmic trading

Algorithmic trading refers to the use of algorithms to make better trading decisions. Usually, traders build mathematical models that monitor economic news and trading activities in real time to detect any factors that could force security prices up or down. The model comes with a set of predetermined instructions on various parameters such as timing, price, quantity and other factors to make trades without the active participation of the trader.

Unlike human traders, algorithmic trading can analyse large volumes of data simultaneously and make thousands of trades every day. Machine learning enables rapid trading decisions, giving human traders an advantage over the market average. Additionally, algorithmic trading does not make trading decisions based on emotions, which is a common limitation among human traders whose judgement may be affected by emotions or personal aspirations. The trading method is mainly used by hedge fund managers and financial institutions to automate trading activities.

3.3 Sentiment Measures

In finance, our interest lies primarily in sentiment aggregated to markets such as the stock market, which is the most common target of ML-based sentiment measures. The majority of relevant studies use measures of sentiment towards stocks to study their effect on future stock returns. There are many studies that construct a measure of investor sentiment from social media e.g. Antweiler and Frank(2004)[1] use naïve Bayes and SVM methods to classify user posts on the Yahoo Finance forum as positive or negative where they aggregate their classifications to construct a measure of stock market sentiment. In addition to text analyses, Obaid and Pukthuanthong (2022)[27] apply machine learning to news photos to derive a sentiment measure for stocks and find that it can replace text-based measures. Other studies use analyst reports or annual reports to measure sentiment.

3.4 Measurements of the business leaders characteristics

The large quantity of image data available free of charge on the Internet allows numerous studies to exploit this information and extract several criteria such the appearance of business leaders, their personality traits, and these own convictions. Indeed, recent progress in ML also allows studies that construct measures of executive emotions. For example, Akansu et al. (2017)[2] apply ML-based face reading to videos of CEOs during press interviews to extract facial emotions and quantify the CEO's mood, emotions such as anger, disgust, fear are measured , sadness, happiness or surprise and study their effect on company performance.

3.5 Measures of the company characteristics

Studies construct measures of company characteristics with ML methods are mainly based on measures of financial characteristics and risk exposures of companies. Buehlmaier and Whited (2018)[6] apply ML to annual reports to construct a measure of financial constraints. So ML can also help study corporate culture. Li et al. (2021)[17] extract aspects of corporate culture from conference call transcripts. Indeed, they study the effect on company performance measures such as operational efficiency and company value. Finally, the capabilities of ML enable the construction of new measures of business connectivity.

Credit risk is a typical economic forecasting problem (see Mestiri, S and Hamdi, M.(2012)[24]). its goal is to detect which potential borrowers will eventually default. Tantri (2021)[15] predicts consumer credit default with strengthened regression trees based on borrower data. Mestiri, S.(2023) [21] use credit card transaction data to predict repayment patterns. Corporate credit risk is another area where ML can provide superior credit risk predictions. Tian et al. (2015)[16] and Hamdi and Mestiri (2014)[14] directly predict corporate bankruptcy from corporate financial statements and market data.

3.6 Forecasting of company results

Analysing the determinants of firm-specific outcomes is an important topic of study in the field of corporate finance that can also be the target of ML-based forecasting. Two studies use ML to predict different financial results. Amini et al. (2021)[3] study corporate capital structure as a typical problem in corporate finance and predict corporate leverage based on standard determinants of capital structure. Mestiri, S. (2023)[21] applied random forests to predict future profits of companies based on their accounting data. Corporate misconduct represents another forecasting problem. Bao et al(2020)[4] type of corporate misconduct we will study is accounting fraud. Xiang et al. (2012)[35] apply ML-based textual analysis to predict startup acquisitions based on company data. Non-parametric models (Mestiri and Farhat 2021)[22] have been investigated in the literature.

4 Limitation and Perspective

ML can process high-dimensional numerical data, that is, data composed of a high number of variables compared to the number of observations. These high-dimensional data arise if there are several economically relevant variables or if non-linearities and interaction effects play an important role. ML methods exploit the information content of this data to make predictions with small errors. Secondly, ML allows the exploitation of unconventional data which is used to extract economically relevant information which can then be a starting point for other economic analyses.

There are also limitations and drawbacks to using Machine learning (ML). First, ML methods tend to have low interpret ability. It is often not directly observable how the algorithm generated its results so ML is not generally suited to problems that require in depth understanding. Second, ML requires large data sets. But unfortunately, large scale data is not always available for many research questions in finance. Finally, the use of ML often incurs high computational costs, e.g. neural networks with complex architectures. However, many researchers are still unclear about how and where to apply ML in finance.

Machine learning relies heavily on data and models and lacks universality and autonomy. In terms of accuracy of forecasting, the performance of Machine learning model is not always better than traditional methods, and the time and computing resources required for Machine learning are much higher than traditional methods. This therefore requires high investments but has great potential.

.

5 Conclusion

Machine learning methods are attracting significant attention in finance. The success of these methods stems from their ability to provide flexible regularized approximations to the theoretically optimal decision rules in data-rich environments. The empirical application of machine learning in finance involves several methodological challenges. In this paper, i cover some of the interesting recent developments that address these challenges. This is an exciting and rapidly growing area of research which needs many interesting methodological developments and applications in the future.

Funding The authors have not disclosed any funding.
Conflict of interest Author states there is no confict of interest with the publication of the present manuscript.

References

[1] Antweiler, W. et Frank, MZ (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance* , 59 , 969-1442

[2] Akansu, A. , Cicon, J. , Ferris, SP and Sun, Y. (2017). Firm Performance in the Face of Fear: How CEO Moods Affect Firm Performance. *Journal of Behavioral Finance* , 18 ,373-389

[3] Amini, S. , Elmore, R. , Öztekin, Ö. and Strauss,J. . (2021) Can machines learn capital structure dynamics?, *Journal of Corporate Finance*, 70,

[4] Bao,Y. , Ke,B. , Li,B. , Yu YJ and Zhang,J. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach . *The Journal of Accounting Research* , 58 , 199 - 235

[5] Breiman, L. (2001). Random Forests. *Machine Learning* 45, 5–32.

[6] Buehlmaier, MMM et Whited, TM (2018). Are Financial Constraints Priced? Evidence from Textual Analysis. *Review of Financial Studies,* , vol. 31(7), 2693-2728

[7] Chang, C., Lin, C.J (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2 (27),:1-27.

[8] Cristianini, N., Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. *Cambridge University Press*, Cambridge, U.K., 1-189.

[9] Ethem Alpaydin. (2004) Introduction to Machine Learning (Adaptive Computation and Machine Learning). The MIT Press.

[10] Fisher, R. (1933). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179–188.

[11] Gelfand S. B., Ravishankar C. S., and Delp E. J. (1991). An iterative growing and-pruning algorithm for classification tree design. *IEEE Transaction on PatternAnalysis and Machine Intelligence*, 13 (2):163-174.

[12] Goodfellow, I. , Bengio, Y. and Courville, A. (2016). Apprentissage en profondeur . Presse du MIT.

[13] Gunn, S.R. (1998). Support vector machines for classification and regression. *Technical Report*, University of Southampton. .

[14] Hamdi, M. and Mestiri,S. Bankruptcy prediction for Tunisian firms : An application of semi-parametric logistic regression and neural networks approach *Economics Bulletin.*Vol. 34 No 1, pages 133-143

[15] Tantri, P. (2021). Fintech for the Poor: Financial Intermediation Without Discrimination, *Review of Finance*, Volume 25 , 561 - 593 .

[16] Tian, S. , Yu, Y. et Guo, H. (2015). Variable selection and corporate bankruptcy forecasts, *Journal of Banking and Finance*, 52 , 89 - 100

[17] Li, K. , Liu, X. , Mai, F. and Zhang, T. (2021). The Role of Corporate Culture in Bad Times: Evidence from the COVID-19 Pandemic. *Journal of Financial and Quantitative Analysis*, 56 , 2545 - 2583 .

[18] Liaw, A.and Wiener, M.C. (2007). Classification and Regression by randomForest.

[19] Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A., 2004, "kernlab – An S4 Package for Kernel Methods in R," Journal of Statistical Software (11:9), 2004, pp. 1-20

[20] Sami Mestiri (2021) Simulation de prêt personnel en utilisant R shiny. Available at *HAL Working Papers* .

[21] Mestiri, S. (2023) Using R software to applied econometrics *Working Papers* hal-04343931,HAL

[22] Mestiri. S and Farhat.A (2021) Using Non-parametric Count Model for Credit Scoring. *Journal of Quantitative Econonomics* Vol.19, pages 39-49 .

[23] Mestiri.S (2019) How to use the R software, *MPRA Paper, University Library of Munich, Germany*

[24] Mestiri, S and Hamdi, M.(2012). Credit Risk Prediction: A Comparative Study Between Logistic Regression and Logistic Regression with Random Effects *International Journal of Management Science and Engineering Management* 7 (3), Taylor & Francis, 200-204.

[25] Therneau TM, Atkinson EJ. (1997) An introduction to recursive partitioning using the RPART routines. *Technical report Mayo Foundation.*

[26] Ohlson, J.A. (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy *Journal of Accounting Research* 18 (1), 109–131.

[27] Obaid, K. and Pukthuanthong, K. (2022). "A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news," *Journal of Financial Economics*, Elsevier, vol. 144(1), pages 273-297.

[28] Quinlan, J.R (1986). Induction of decision trees. *Machine Learning* 1, 81-106.

[29] Roy, T., Tshilidzi, M., Chakraverty, S. (2021). Speech emotion recognition using deep learning. *New Paradigms in Computational Modeling and Its Applications*, 177-187.

[30] Shetty, S., Musa, M., Brédart, X. (2022). Bankruptcy Prediction Using Machine Learning Techniques. *Journal of Risk and Financial Management* 15(35), 1-10.

[31] Shin, K.S, and Lee, Y.J. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications* 23, 321–28.

[32] Vapnik, V. (1998). The nature of statistical learning theory. New York: Springer.

[33] Venables, W. N. and Ripley, B. D., 2002, Modern Applied Statistics with S, Springer, New York.

[34] Wilson, R.L. and Sharda, R. (1994). Bankruptcy Prediction Using Neural Networks*Decision Support Systems* 11, 545-557.

[35] Xiang, G. , Zheng, Z. , Wen, M. , Hong, J. , Rose, C. et Liu,C. (2012).A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch. Proceding Sixth International AAAI Conference on Weblogs and Social Media , Irlande (vol. pp. 607 – 610).