The University of Texas at Austin
McCombs School of Business

Date Here

CONFRONTING MACHINE LEARNING WITH FINANCIAL RESEARCH

Kristof Lommers, Ouns El Harzli, Jack Kim

Presented by
Minjoo Kim

1. INTRODUCTION

- During the past decade, machine learning has gained substantial interest among scientists but its adoption in financial-economic academia has been rather limited.
- Financial markets present an unique data environment
- Finance's research paradigm differs considerably from other fields in which most machine learning progress has been made.
- In this paper, authors discuss some of the main challenges of machine learning in finance and examine how these could be addressed
- Moreover, authors discuss the various applications of machine learning in the research process such as estimation, empirical discovery, testing, causal inference and prediction.

2. CONFRONTING MACHINE LEARNING WITH FINANCIAL RESEARCH PARADIGM

- There are three main conflicts between finance research and machine learning.

1) First, empirical research in finance has a strong focus on statistical inference, but machine learning has focus on prediction.

2) Second, causal understanding has a central place in financial economics while most machine learning methods do not place much emphasis on causation.

3) Third, financial economics is hypothesis-driven while machine learning tends to be data-driven

-> pre-specified hypothesized relationships vs learn from data and the specific functional form is decided based on the best fit with the underlying data

- As discussed further, machine learning could address the challenges, and can be applied to improve various parts of the empirical research process.

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 1) NON-STATIONARITY

- Machine learning performs better for tasks which are static in nature but financial markets are dynamic throughout time and are constantly subject to change
- However, adjustment can be made for non- stationary and machine learning can be used to provide new ways of tackling non-stationary time series
- Dynamic support vector machines can learn different patterns for regime changes directly from the data (Cao and Gu,2002).
- Shalizi, Jacobs, Klinkner, and Clauset (2011) proposed ensemble methods where the contribution of the different models can vary over time to take non-stationarity into account.
- LSTM -> solves the vanishing gradient problem of RNN by adding gates which enables model to learn both long term/short term time information.

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 2) LIMITED DATA

- Finance is fundamentally not a big data environment compared to other fields where machine learning successes were achieved. (limited time series realization, also limited tradable securities that researchers cannot arbitrarily add)
- Limited data -> increases the risk of overfit and causes other several other problems such as class imbalance, distributional asymmetry, and limited information availability on certain infrequent events

1. Transfer learning

- First, we pre-trained the model on a larger data that shares commonalities with the data of interest. Then, the model is updated with the small data which reinforces the patterns that match and reduces the confidence in the patterns that cannot be seen in the small data of interest.

2. Synthetic data

- Machine learning can learn the underlying structure and properties of the data and generate synthetic data.
- Ex) TimeGAN -> enables the generation of data according to a best-fit neural network approximator

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 3) OVERFIT

- The low signal-to-noise environment in financial markets naturally increase the probability for machine learning to pick up spurious patterns.

1. Cross validation

- The limited data availability in finance limits the usage of prominent techniques against overfitting such as K-fold cross validation -> nested cross validation for limited data/ blocked cross validation for time series

2. Ensemble

- Ensemble methods allow to combine a set of different learning algorithms or the same learning algorithm with different hyperparameters and features ex) Random forest

3. Regularization

- Manage the complexity of the model via penalization of the loss function
- lasso, ridge, and elastic net
- Neural network -> dropout

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 3) OVERFIT

4. Combining the theory with machine learning

- Israel, Kelly and Moskowitz (2020) argue that combining theory with machine learning could mitigate the issue of overfitting and spurious relationships
- Gu, Kelly and Xiu (2019) show improved predictions by embedding theoretical considerations such as no-arbitrage and the linkage between return and risk in equilibrium in an autoencoder model
- Applying aforementioned machine learning methodologies and the philosophy of out-of-sample generalizability could help combat p-hacking, and thus in-sample overfitting, in financial research

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 4)CURSE OF DIMENSIONALITY

- The curse of dimensionality : exponential increase in the required amount of data to sufficiently train a model when the dimensionality increases.
- It is strongly present in finance given the high dimensionality in terms of potential explanatory features (ex: factor zoo) and relatively limited time series data for these variables.
- Dimensionality reduction could be achieved through feature selection or feature extraction.
- Feature selection : selecting the most relevant features
- Feature extraction : transforming features into a lower dimensionality representation
- Unlike traditional feature extraction, Machine learning can handle non-linearity (autoencoder, Kernel PCA).

3. ADDRESSING THE CHANLLEGNES OF MACHINE LEARNING IN FINANCE – 4) INTERPRETABILITY

- The potential accuracy improvements afforded by complex, high dimensional, and nonlinear machine learning models often comes at the cost of less interpretable interactions, which is one of the main goal of finance research.
- In recent years, researchers have developed approaches to incorporate interpretability in modern machine learning algorithms by structurally enforcing some of the interpretability constraints.
- Vaughan (2018) proposes explainable neural networks (XNN) by limiting the connections between nodes such that the learned network model is a modified additive index model.
- Researchers can summarize the importance of different input features with post-hoc interpretability (Neural network -> DeepLIFT, random forests ->mean decrease impurity).
- Multiple mechanisms can reveal significant non-linearities and cross-effects that cannot be assessed in traditional finance models.
- In this regard, machine learning can help to get a better understanding of the complex, non-linear and multivariate relationships present in financial markets.

4. APPLICATIONS OF MACHINE LEARNING IN FINANCIAL RESEARCH -1) EMPIRICAL DISCOVERY

- **Machine learning is an ideal candidate for empirical discovery because of its ability to identify complex patterns and break down high-dimensional data into low-dimensional components**
- Instead of using predefined mathematical simplifications to describe the reality of financial markets, one could use machine learning algorithms to discover empirical relationships from the data
- Leippold, Wang, and Zhou (2020) use machine learning for empirical discovery as they generate systematic signals in the Chinese stock market through the use of a machine learning methods.
- Gu (2020) uses feedforward network models to capture the nonlinear interactions between factors in asset price prediction.
- Chen (2020) employs an LSTM-GAN architecture to reveal nonlinear interactions between factors and uncover the importance of latent macroeconomic state variables in asset pricing.

4. APPLICATIONS OF MACHINE LEARNING IN FINANCIAL RESEARCH -2) ESTIMATION

1) Machine learning is particularly useful for processing unstructured data and generating estimates that encapsulate fundamental information contained in the data

- Unstructured data comes in many different sources such as language data (text and voice), visual data (pictures and videos), and geo-spatial data.

2) Machine learning could also be used to estimate financial variables that are traditionally based on qualitative human judgment.

- Ding et al. (2020) show that machine learning methods improve accounting estimates as financial statements sometimes depend on subjective managerial estimates.

3) Machine learning also provides opportunities for improved estimation of structured data because it allows more flexible nonlinear patterns

- Gu, Kelly and Xiu (2020) machine learning can be used for improved measurement of risk premia in asset pricing

4. APPLICATIONS OF MACHINE LEARNING IN FINANCIAL RESEARCH -3) EMPIRICAL TESTING

- Machine learning can be used for the empirical testing of theories providing several advantages compared to conventional statistical models.
- **Firstly, machine learning allows for the measurement and testing of more complex non-linear relationships**
- Stevanovic and Surprenant (2019) and Gu, Kelly and Xiu (2020) observe advantages of machine learning methods relative to conventional econometric methods in asset pricing and macroeconomic predictions as a result of these nonlinear interactions
- **Secondly, machine learning models generally have less assumptions on the relationships and the distributional properties of the data. →This flexibility decreases the tendency for specification errors in financial research which can lead to false rejection of important variables or relationships**
- With post hoc analysis and new methods of significance test that does not rely on strong distributional assumptions, we can do better empirical testing
- Also the machine learning literature provides various tools for more extensive robustness testing.(discussed in overfitting section)

4. APPLICATIONS OF MACHINE LEARNING IN FINANCIAL RESEARCH -4) CASUAL INFERENCE

- There is an inherent difficulty of causality analysis in financial-economic problems because controlled experiments are in most cases not possible
- Machine learning can improve causal inference in two ways

- **1. Firstly, machine learning could be used to improve techniques like regression discontinuity, instrumental variables and difference-indifferences**
- ex) For instance, for the instrumental variables problem, Belloni, Chen, Chernozhukov, and Hansena (2012) have used LASSO regression to determine the optimal set of instruments in the first stage

- **2. Secondly, machine learning offers new ways of addressing causal inference.**

- ex) Tiffin (2019) proposes to leverage the properties of random forests to estimate, in order to isolate the causal factors. Imbens (2016) developed a causal tree.
- Bayesian causal networks -> augment structural equation and represent dependency structure

4. APPLICATIONS OF MACHINE LEARNING IN FINANCIAL RESEARCH -5)PREDICTION

- Given some of the stylized facts of returns such as heavy-tailed distribution, gain-loss asymmetry, and volatility clustering, non-linear tendencies and higher-order interactions in financial relationships should be considered.
- often times, finance research require a large set of predictors.
- For these reasons machine learning tends to outperform in prediction tasks compared to conventional econometric methods
- Gu, Kelly and Xiu (2020) for asset pricing, Stevanovic and Surprenant (2019) for macroeconomics
- This tendency for outperformance of prediction can be especially helpful for tasks, such as asset pricing, where the quality of the predictions are an important determinant to understand the financial mechanisms and test theories

CONCLUSION

- Machine learning algorithms have been mainly developed for particular data environments, and there are some key differences with financial markets in this regard.
- Despite some challenges, machine learning can be unified with financial research when appropriate methodological adjustments are made.

The University of Texas at Austin
McCombs School of Business