An Introduction to Machine Learning for Economists

Shailu Singh[1]

Associate Professor, Department of Economics, Hansraj College, University of Delhi

1. Introduction

Machine Learning is an area of rapidly growing importance in empirical economics. It uses data to identify complex relationships between economic variables which contribute to a better understanding of the underlying data generating processes with the potential to significantly enhance the accuracy and efficiency of economic models (Varian, 2014). Currently, machine learning algorithms are being widely used in the areas of economic forecasting, financial modeling (Gogas, P., & Papadimitriou, T., 2021), and risk management among many others. In this article, we conduct a very primitive review targeted mainly for undergraduate students, of the role of machine learning in economics by understanding its scope, approach, and link to the conventional methodology. Pointers to some current applications highlight the potential that young researchers can harness by including machine learning techniques in their toolbox.

2. Scope: Why Machine Learning?

Why have economists become so interested in machine learning? Empirical economics has its own toolbox consisting of conventional statistical and econometric techniques, regression analysis being its workhorse. Is it because machine learning offers new, more efficient ways of doing things? Or is it because it allows us to do things that the conventional techniques were ill equipped to handle? The answer to both the questions is in the affirmative. Machine learning uses more data driven methods to identify relationships between variables compared to the traditional more theory driven methods. Historically, these methods were not so widely

[1] Email id: shailusinghs@gmail.com

used in economic applications as the large amounts of data they required were typically unavailable. Today the kind of data available has undergone a sea change both quantitatively and qualitatively. Vast amounts of data are available which makes these methods not only very attractive, but often the only option as conventional methods is not geared to handle data sets of this size. Also, since the early 2000's the use of multiprocessor graphic cards has sped up computer learning and made these tools easily implementable. However, current advances in machine learning allow the use of methods that work with smaller amounts of data as well and give more efficient results than the conventional techniques. The innovation derives from not just from using large datasets more efficiently but also from allowing the use of some very novel data sources for example text data (Khandayet al., 2020; Chatterjee et al., 2019).Also, scholarship has identified areas for combining conventional and machine learning methods and it is safe to say that the intersection between them is growing substantially.

3. Approach: How is it different?

Empirical economics has relied heavily on regression analysis, both parametric and non-parametric. Parametric models assume that we know the function that describes the relation between the dependent and the independent variable/s while non-parametric regression does not assume anything about the relationship between them. Since data is needed to discover the functional form besides estimation of the non-parametric model, these methods typically require more data.

The most basic parametric model, linear regression postulates the mean of the variable of interest as conditional on one or more explanatory variables. A random sample drawn from the population of interest is used to estimate the conditional means by minimizing a loss function such as the residual sum of squares. The R square is a measure of the goodness of fit. The estimated model performs three key functions: first, quantification of relations between economic variables i.e., estimation of the marginal effects; second, prediction based on the estimates obtained and third, hypothesis testing, while more complex methodologies aim to estimate causal effects.

Machine learning, methods are focused primarily on prediction though researchers are working on extending its use to hypothesis testing and causal analysis (Athey, S., 2015) as well. In short, data driven methods are used to get an efficient prediction of a dependent variable given a set of independent variables. The table below shows how the nomenclature of machine learning models differs from the conventional ones.

Table 1: Comparison of Conventional Analysis and Machine Learning

CONVENTIONAL ANALYSIS	**MACHINE LEARNING**
Model	Algorithm
Predictors	Features
Estimation, Prediction and Inference	Prediction
Prediction considered good if high R bar square	Prediction considered good if gives good out of sample predictions
Estimating the model	Training the model

While the number of machine learning algorithms in use is very large, those used for regression and classification such as Linear Regression, Logistic Regression, Stepwise Regression, Multivariate Adaptive Regression Splines (MARS), Locally Estimated Scatterplot Smoothing (LOESS) and Decision Trees are most popular, especially among economists. Other algorithms used are instance based such as k-Nearest Neighbor (kNN) and Support Vector Machines (SVM); Regularization algorithms such as Ridge Regression, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Nets; Ensemble Algorithms such as Boosting, Bootstrapped Aggregation (Bagging) and Random Forests.StatQuest with Josh Starmer is a very useful YouTube channel (https://www.youtube.com/@statquest) to understand each of these methods clearly.

The numerous ML algorithms mentioned above are broadly classified under two categories - Supervised and Unsupervised. Supervised algorithms are used to construct predictive models. While the conventional applied economist will think only of linear regression if the dependent variable is continuous and logistic regression if it is binary, machine learning offers many other choices which could reveal aspects of the relationship between two or more variables that the traditional methods may not be able to grasp (Varian, 2014).These

algorithms may be used for classification or numeric prediction. The target or the output feature is the variable that is being predicted using the input features. The algorithm gives clear instructions on what and how it will learn from data. Unsupervised algorithms are used to build descriptive models. This is done by looking for patterns in the data that can then be summarized in new and interesting ways. Unlike Supervised algorithms there is no target variable here; all features are at the same level. These algorithms are used for pattern discovery or clustering/segment analysis.

A key feature of all ML methods is that data is divided into two sets- a **training set** and a **testing set**. The training set is used for estimation or fitting the data. The estimates obtained are then used to see how well they describe the remaining data i.e., the testing set. This is unlike the traditional econometric approach where the entire data is used for estimation purposes and then the coefficient of determination is used to indicate the goodness of fit. This out of sample testing is a characteristic of machine learning methods that makes the prediction more efficient.

Another distinctive feature is the technique of **Cross Validation** (Zhang, Y., & Yang, Y., 2015; Bergmeir,2012). A problem frequently encountered by empirical economists is the choice of the functional form for the model to be estimated. There is a wide range of models that may be employed, parametric and non- parametric models both. Theory often does little to clarify matters in this regard. Students of basic econometrics will be reminded that if the model specified is linear in the parameters linear regression may be applied though the model does not have to be linear in the explanatory variables. In fact, polynomial terms are used to fit more flexible functional forms. For example, a linear relation between Y and X would require just one term in addition to the intercept whereas a U-shaped curve would imply a parabolic relation and hence a quadratic term in addition to the linear term. To go further, if the relation between Y and X involves a point of inflection then a cubic term needs to be added. Thus, the more flexible the relation that one wants to fit, the larger the number of polynomial terms that need to be incorporated in the specified model. The traditional regression techniques allow for the model to be as flexible as one wants, but in doing so degrees of freedom are lost and so the model becomes increasingly imprecise. Too many

predictors than appropriate for estimation, force the use of data reduction techniques which need not be the only choice of the researcher if machine learning models are employed.

Cross validation uses data to choose between alternative models, the objective being to choose the one that gives the best predictions. So, for a particular application, one might think that either of Logistic regression, Support Vector Machine or K- nearest neighbors may be appropriate. The approach is to apply the train-test procedure to each method and compare their prediction performance. Each time a record is kept of how many incorrect predictions are made by each method. A factor contributing to the efficiency of these methods in predicting better is that the division of data used for training and testing is not arbitrary. All the data is divided into a certain number, say 4 blocks. It first uses blocks 1, 2 and 3 for training the data and then uses the block 4 data for testing and makes a note of how well the model predicted the block 4 data. It then repeats the process such that each block is used for testing the data once. It then summarizes the performance of each method and chooses the method that does the best job. Since the data was divided into 4 blocks, this is referred to as fourfold cross validation. If data is divided into k blocks, it is called k fold cross validation. In the extreme case, each block consists of one observation. So, if there are 10 data points and hence 10 blocks, 9 blocks are used for training and one for testing and then summarizing the performance. This is called “leave one out cross validation”. In general, 10-fold cross validation is practiced. Thus cross validation ensures maximum predictive efficiency and is being widely adopted even in conventional models.

4. Conclusion

One of the earliest uses of machine learning in economics was in finance where White (1988) used Neural Networks for predicting daily IBM stock returns. Advancements in the field of computing such as parallel computing made Deep Learning techniques like Recurrent Neural Networks and Convolutive Neural Networks available for economic applications. Currently, algorithms such as Support Vector Machines and Random Forests and techniques such as kernelization, bagging and boosting allow the application of these models to relatively small datasets (Gogas, P., & Papadimitriou, T. (2021) and make it a very potent tool for both

Macro and Micro forecasting. In many applications the results from the application of machine learning methods are more efficient than those obtained from adopting traditional econometric methods (Kreiner, A., & Duca, J., 2020). Increasingly, researchers are finding ways to combine both (Bertoletti et al., 2022) such as the GARCH-SVM (Chen et al., 2010) while integrating machine learning techniques such as cross validation into the traditional too (Bergmeir,,2012).

Machine learning has immense potential to advanceeconomic research and the literature is full of innovative applications such as prediction of household solid waste generation (Namoun, A., et al.,2022), time series forecasting (Masini et al., 2023), detecting Covid 19, using text data (Khanday et al., 2020), understanding emotions in text (Chatterjee et al., 2019), personality analysis, predicting personality with social media(Golbeck et al.,2011), selection and productivity of human capital (Chalfin et al., 2016) to cite a few. Table 2 lists some applications with the algorithms used in the studies respectively.

Table 1: Examples of machine learning methods applied in economics-related fields.

Sources	Machine Learning Models	Objectives
Lee et al. (2020)	Support Vector Regression (SVR)	Anomaly Detection
Husejinović (2020)	Naive Bayesian And C4.5 Decision Tree Classifiers	Credit Card Fraud Detection
Zhang (2019)	Improved BP Neural Network	Aquatic Product Export Volume Prediction
Sundar and Satyanarayana (2019)	Multilayer Feed Forward Neural Network	Stock Price Prediction
Hew et al. (2019)	Artificial Neural Network (ANN)	Mobile Social Commerce

Abdillah and Suharjito (2019)	Adaptive Neuro-Fuzzy Inference System (ANFIS)	E-Banking Failure
Sabaitytė et al. (2019)	Decision Tree (DT)	Customer Behavior
Zatevakhina, Dedyukhina, and Klioutchnikov (2019)	Deep Neural Network (ANN)	Recommender Systems
Benlahbib and Nfaoui (2020)	Naïve Bayes and Linear Support Vector Machine (LSVM)	Sentiment Analysis

Source: Extracted from Nosratabadiet al., 2020

Aspiring economists need to understand these methods to fully exploit their research potential, conduct extensive reviews of the applications of these methods in niche areas of economics such as Finance, Energy, Education, Development etc. to identify the gaps and contribute to the discipline.

References

- Athey, S. (2015, August). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 5-6).
- Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685-725.
- Babenko, V., Panchyshyn, A., Zomchak, L., Nehrey, M., Artym-Drohomyretska, Z., & Lahotskyi, T. (2021). Classical machine learning methods in economics research: Macro and micro level example. *WSEAS Transactions on Business and Economics*, *18*, 209-217.
- Basu Choudhary, A., Bang, J. T., & Sen, T. (2017). *Machine-learning techniques in economics: new tools for predicting economic growth*. Springer.
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192-213.
- Bertoletti, A., Berbegal-Mirabent, J., & Agasisti, T. (2022). Higher education systems and regional economic development in Europe: A combined approach using

econometric and machine learning methods. *Socio-Economic Planning Sciences*, *82*, 101231.

- Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124-127.
- Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., & Agrawal, P. (2019). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, *93*, 309-317.
- Chen, S., Härdle, W. K., & Jeong, K. (2010). Forecasting volatility with support vector machine-based GARCH model. *Journal of Forecasting*, *29*(4), 406-433.
- Ghoddusi, H., Creamer, G. G., & Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709-727.
- Gogas, P., & Papadimitriou, T. (2021). Machine learning in economics and finance. *Computational Economics*, *57*, 1-4.
- Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 253-262).
- Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., Rouf, N., & Mohi Ud Din, M. (2020). Machine learning based approaches for detecting COVID-19 using clinical text data. *International Journal of Information Technology*, *12*, 731-739.
- Kreiner, A., & Duca, J. (2020). Can machine learning on economic data better forecast the unemployment rate?. *Applied Economics Letters*, *27*(17), 1434-1437.
- Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76-111.
- Mele, M., & Magazzino, C. (2021). Pollution, economic growth, and COVID-19 deaths in India: a machine learning evidence. *Environmental Science and Pollution Research*, *28*, 2669-2677.
- Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.
- Namoun, A., Hussein, B. R., Tufail, A., Alrehaili, A., Syed, T. A., & BenRhouma, O. (2022). An Ensemble Learning Based Classification Approach for the Prediction of Household Solid Waste Generation. *Sensors*, *22*(9), 3506.
- Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, *8*(10), 1799.
- Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28.
- White, H. (1988, July). Economic prediction using neural networks: The case of IBM daily stock returns. In *ICNN* (Vol. 2, pp. 451-458).
- Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95-112.
- https://www.youtube.com/@statquest