Machine Learning for Econometricians:
The Readme Manual

Marcos López de Prado

Marcos López de Prado
is global head of quantitative research and development at Abu Dhabi Investment Authority in Abu Dhabi, UAE, and is a professor of practice at Cornell University in New York, NY.
ml863@cornell.edu

KEY FINDINGS

- The econometric canon is poorly suited for conducting investment research because of the unstructured nature of financial data as well as the complex relationships involved in economic systems.
- Machine learning (ML) techniques offer the numerical power and functional flexibility needed to identify complex patterns in a high-dimensional space.
- This article offers a bridge from classical statistics to ML so that econometricians may adopt ML tools into their existing research processes.

ABSTRACT

One of the most exciting recent developments in financial research is the availability of new administrative, private sector, and micro-level datasets that did not exist a few years ago. The unstructured nature of many of these observations, along with the complexity of the phenomena they measure, means that many of these datasets are beyond the grasp of econometric analysis. Machine learning (ML) techniques offer the numerical power and functional flexibility needed to identify complex patterns in a high-dimensional space. ML is often perceived as a black box, however, in contrast to the transparency of econometric approaches. In this article, the author demonstrates that each analytical step of the econometric process has a homologous step in ML analyses. By clearly stating this correspondence, the author's goal is to facilitate and reconcile the adoption of ML techniques among econometricians.

In a general sense, econometrics encompasses the set of statistical methods applied to economic and financial data, with the purpose of providing empirical support to economic theories. In practice, however, this set of statistical methods has traditionally concentrated on the multivariate linear regression model. There are several reasons why multivariate linear regression models have been popular for the past 100 years: Economic datasets were mostly numerical, short in length, low-dimensional, and with a low signal-to-noise ratio. Limitations in the data often justified the use of relatively constrained specifications.

In recent years, the quantity and granularity of economic data have improved dramatically. The good news is that the sudden explosion of administrative, private sector, and micro-level datasets offers an unparalleled insight into the inner workings of the economy (Einav and Levin 2014; Kolanovic and Krishnamachari 2017). The bad news is that these datasets pose multiple challenges to the econometric toolkit.

To cite just a few challenges: (1) Some of the most interesting datasets are unstructured.[1] They also can be non-numerical and noncategorical, like news articles, voice recordings, or satellite images. (2) These datasets are high dimensional (e.g., credit card transactions). The number of variables involved often greatly exceeds the number of observations, making it very difficult to apply linear algebra solutions. (3) Many of these datasets are extremely sparse. For instance, samples may contain a large proportion of zeros, where basic notions of association such as correlation do not work well. (4) Embedded within these datasets is critical information regarding networks of agents, incentives, and aggregate behavior of groups of people (Easley and Kleinberg 2010).[2]

As a result of these challenges and the complexities of these new datasets, there are two reasons why there is a limit to how much economists can learn from regression models and other linear algebraic or geometric approaches: (1) even with older datasets, traditional techniques are likely too rudimentary to model complex (e.g., nonlinear and interactive) associations among variables; and (2) liquid securities may be too efficient for traditional techniques because whatever inefficiency remains unexploited is too complex for econometric models. Under this second argument, whatever relationship is identified by regression methods on liquid securities must be spurious by construction.

Machine learning (ML) offers a modern set of statistical tools specifically suited to overcome the challenges of new economic and financial data sources and increasingly complex associations in financial markets. Despite this, the use of ML in academic finance remains the exception rather than the rule. Part of the reason may be the false perception that ML operates as a black box, which contrasts the transparency of standard econometric analyses. The goal of this article is to debunk that false perception. The author argues that each analytical step of the econometric process has a direct homologue in ML analyses. By explicitly stating this correspondence, the author wishes to encourage the adoption of ML techniques among applied economics and finance researchers.

The author's message is that financial datasets are increasingly beyond the grasp of econometrics and that ML is a transparent research tool with an important role to play in financial studies. For all of these reasons, finance professionals and academics should familiarize themselves with these techniques, and economics students should enroll in data science courses (in addition to their mandatory econometrics training).

THE ECONOMETRIC CANON

In the words of William Greene, "the concept of multiple regression and the linear regression model in particular constitutes the underlying platform of most econometric modeling, even if the linear model itself is not ultimately used as the empirical specification" (Greene 2012).

Multivariate linear regression modeling is not a modern technology. Its history goes back to at least 1795, when Carl Friedrich Gauss applied ordinary least squares (OLS) to geodesic and astronomic datasets (Stigler 1981). Interestingly, Gauss thought that OLS was so obvious that it did not merit publication. British eugenicist Sir Francis Galton coined the term *regression* in 1886 because he estimated linear

[1] For example, (Financial Information eXchange) FIX messages recorded by exchanges (Easley et al. 2016). These records convey information about the dynamics of order books, with buyers and sellers interacting with each other at random times (they are inhomogeneous). They share more similarities with the transcript of a strategy game than with a typical macroeconomic data tables and panels.

[2] As Leonhard Euler discovered in 1736, geometric objects (like covariance matrices) fail to recognize the topological relationships that characterize networks.

equations to argue that hereditary human physical and moral traits exhibit a regression toward the mean. Near the turn of the 20th century, Karl Pearson coined the term *regression line* in reference to Galton's argument and introduced the method of moments (which was generalized by Lars Hansen 80 years later). Over the following years, Ronald Fisher studied and proved the mathematical properties of regression analysis and popularized maximum likelihood estimation. These ideas gave birth to much of what we study today under the name *econometrics*. Its canon is best exemplified by the content of standard econometrics textbooks (e.g., Tsay 2013; Greene 2012; Wooldridge 2010; Hayashi 2000).

The same quantitative canon used by economists is called *biostatistics* when applied to biological datasets and *chemometrics* when applied to chemical datasets. Even entry-level biostatistics and chemometrics textbooks, however, often include advanced clustering, pattern recognition, or computational methods, which are largely absent, or less emphasized, in popular econometrics textbooks (e.g., compare Greene 2012 with Otto 2016 and Balding, Bishop, and Cannings 2007). Computational methods have become particularly important in these disciplines because they can replace some (possibly unrealistic) assumptions regarding the data-generating process.

A number of leading economists have voiced their frustration with economists' underutilization of data-driven and numerical methods and the overuse of structural models based on arbitrary assumptions. For instance, Nobel laureate Wassily Leontief expressed this concern almost 40 years ago (Leontief 1982):

> Not having been subjected from the outset to the harsh discipline of systematic fact-finding, traditionally imposed on and accepted by their colleagues in the natural and historical sciences, economists developed a nearly irresistible predilection for deductive reasoning. As a matter of fact, many entered the field after specializing in pure or applied mathematics. Page after page of professional economic journals are filled with mathematical formulas leading the reader from sets of more or less plausible but entirely arbitrary assumptions to precisely stated but irrelevant theoretical conclusions.

A simple bibliometric analysis of the economic literature illustrates that Leontief's lament still applies today to a large extent. For example, the Web of Science[3] reports that 13,772 journal articles have been published on subjects in the intersection of *Economics* and *Statistics and Probability*. Among those publications, only 89 articles (0.65%) contained any of the following terms: classifier, clustering, neural network, or machine learning. In contrast, out of the 40,283 articles in the intersection of *Biology* and *Statistics and Probability*, a total of 4,049 (10.05%) contained any of those terms. Out of the 4,994 articles in the intersection of *Chemistry, Analytical* and *Statistics and Probability*, a total of 766 (15.34%) contained any of those terms. Part of this gap is explained by the deficiencies of old economic datasets, which may have precluded the use of ML and other data-intensive techniques. However, this gap should have narrowed years ago in the face of new unstructured, high-dimensional, sparse, non-numeric economic datasets. Institutional resistance continues to prevent the adoption of ML techniques among academic economists, partly because the peer-review and tenure processes today still favor the adherence to the econometric canon. This resistance to modernization is mirrored by fund management firms, since decades ago these firms (sadly) discovered the effectiveness of journal articles for marketing investment funds.

The explosion of economic and financial data should encourage practitioners to reconsider the appropriateness of fitting simple models. The new available datasets include social media, metadata scraped from websites, satellite images, sensor data,

[3] https://www.webofknowledge.com, as of November 26, 2018.

sentiment extracted from text, and microstructural data generated by exchanges. SINTEF has estimated that 90% of all available data have been collected over the previous two years (SINTEF 2013). The International Data Corporation has estimated that 80% of all available data are unstructured (Zwolenski and Weatherill 2014) and, hence, not amenable to traditional quantitative methods. These detailed sources of information were not available a few years ago, and they finally offer us the possibility to develop economic theories grounded in rich empirical evidence. However, these are also 21st century datasets whose structure cannot be easily uncovered with traditional econometric tools. The next section reviews how other fields have applied ML.

HOW SCIENTISTS USE ML

An ML algorithm learns complex patterns in a high-dimensional space with little human guidance on model specification. That ML models need not be specified by the researcher has led many to erroneously conclude that ML must be a black box. In that view, ML is merely an *oracle*,[4] a prediction machine from which no understanding can be extracted. The black box view of ML is a misconception. It is fueled by popular industrial applications of ML, where the search for better predictions outweighs the need for theoretical understanding. In fact, ML models can be interpreted through a number of procedures, such as partial dependence plot (PDP), individual conditional expectation (ICE), accumulated local effects (ALE), Friedman's H-stat, mean decrease impurity (MDI), mean decrease accuracy (MDA), global surrogate, local interpretable model-agnostic explanation (LIME), and Shapley values, among others. The discussion of these approaches is beyond the scope of the article; please see Molnar (2019) for details.

A review of recent scientific breakthroughs reveals radically different uses of ML in science. The following five use cases stand out:

1. **Existence:** ML has been used to evaluate the plausibility of a theory across all scientific fields, even beyond the empirical sciences. Notably, ML algorithms have been used to make mathematical discoveries. ML algorithms cannot prove a theorem; however, they can point to the existence of an undiscovered theorem, which can then be conjectured and eventually proved. In other words, if something can be predicted, there is hope that a causal or logico-deductive mechanism can be uncovered (Gryak, Haralick, and Kahrobaei 2018).
2. **Importance:** ML algorithms can determine the relative informational content of variables (*features*, in ML parlance) for explanatory and/or predictive purposes (Liu 2004). For example, the MDA method follows these steps: (1) Derive the out-of-sample cross-validated accuracy of an ML algorithm on a particular dataset; (2) repeat step 1 after shuffling the observations of individual features or combinations of features; (3) compute the decay in accuracy between steps 1 and 2. Shuffling the observations of an important feature will cause a significant decay in accuracy. Thus, although MDA does not uncover the underlying mechanism, it discovers the variables that should be part of the theory.
3. **Causation:** ML algorithms can search for the structure of causal relationships, by analyzing statistical properties of purely observational data. Over the past three decades, statisticians have developed numerous computational methods and algorithms for the discovery of causal relations, represented as directed acyclic graphs (see Glymour et al. 2019, Athey 2015).

[4] Here the author uses a common definition of *oracle* in complexity theory: A black box that is able to produce a solution for any instance of a given computational problem.

4. **Reductionist:** ML techniques are essential for the visualization of large, high-dimensional, complex datasets. For example, manifold learning algorithms can cluster a large number of observations into a reduced subset of peer groups, whose differentiating properties can then be analyzed (Schlecht et al. 2008).
5. **Retriever:** ML is used to scan through big data in search of patterns that humans failed to recognize. For instance, every night ML algorithms are fed millions of images in search of supernovae. Once they find one image with a high probability of containing a supernova, expensive telescopes can be pointed to a particular region in the universe, where humans will scrutinize the data (Lochner et al. 2016). A second example is outlier detection. Finding outliers is a prediction problem rather than an explanation problem. An ML algorithm can detect an anomalous observation based on the complex structure it has found in the data, even if that structure is not explained to us (Hodge and Austin 2004).

These five ML use cases are associated with analysis of data before any theory has been conjectured. This contrasts with the econometric tradition of fully specifying models after the theory has been conjectured. Because of this tendency, economics is often criticized for being exceedingly abstract, aprioristic, and detached from practical reality.[5] Rather than replacing economic theories, ML could play the critical role of helping economists form theories based on rich empirical evidence. ML provides the opportunity for economists to apply powerful data science tools to the construction of new theories.

A ROADMAP FROM ECONOMETRICS TO ML

ML is an integral part of modern statistics. ML tools can be best understood as the natural evolution of traditional statistics in the computer age (Efron and Hastie 2016). One way to understand this evolution is to examine how ML addresses each step in the typical econometric workflow. Exhibit 1 lists the correspondence between econometric and ML analytical steps, providing a roadmap for economists who wish to modernize their empirical toolbox. Over the remainder of this section, I discuss each of these steps, underscoring the similarities and differences between econometrics and ML.

Goal Setting

Wooldridge (2010) explains that the goal of econometric studies is to determine causal relationships. Economists can rarely carry out a controlled experiment, where one variable is exogenously changed while all other variables are held fixed. Instead, economists run a thought experiment, where the sensitivity of one variable to changes in another is evaluated while the effect of all other relevant variables is controlled.[6] This is the so-called *ceteris paribus* argument that econometrics uses to deduce

[5] In the words of Nobel laureates Paul Romer and Robert Solow, economic theories are sometimes based on "facts with unknown truth value" (Romer 2016) and "generally phony" assumptions (Solow 2010).

[6] Pearl (2009) argues that this thought experiment requires a number of causal assumptions to complement the observational data used in a regression. Without those assumptions, researchers cannot rule out the existence of excluded confounding variables responsible for the phenomenon. Chen and Pearl (2013) conduct a critical examination of the treatment of causation in econometrics, concluding the most econometrics textbooks mistake correlation with causation.

EXHIBIT 1

Correspondence between Steps in the Econometric and Machine Learning Research Processes

STEP	ECONOMETRICS	CAVEAT	ML SOLUTION
Goal setting	Variance adjudication (in-sample)	Biased when confounders are omitted; poor forecasting performance	Out-of-sample prediction: Upper-boundary, importance, causation, reductionist, retriever
Visualization	Time plots, scatter plots, pie charts, histograms, heat maps	Not well suited for high- dimensional datasets with complex associations	*t*-SNE, networks, geospatial, classification trees, treemaps, etc.
Outlier detection	Winsorizing, trimming, Chauvenet's criterion, Mahalanobis distance, Dixon's Q test, etc.	Heuristic approaches	Anomaly detection methods, RANSAC
Feature extraction	PCA	It misses non-linear interactions and label information; it imposes a change of basis	Kernel-PCA, LDA, biclustering
Regression	Algebraic models	High risk of model mis-specification (wrong functional form, missing interaction effects, wrong assumptions, etc.)	Neural networks, SVR, GA, regression trees, etc.
Classification	Logit, probit	Linear specification; categorical or ordinal regressors require dummy variables	RF, SVC, k-NN, etc.
Feature importance	*p*-values	It assumes that the model is correctly specified; under multicollinearity, results are biased and not robust	MDI, MDA per cluster, Shapley values
Model selection/ Overfitting prevention	Forward selection, backward elimination, bidirectional elimination	Selection bias; in-sample procedures; mis-specification	Specification consistent with causal graph. Regularization, ensemble methods, etc.
Model validation	Adjusted *R*-squared (in-sample), analysis of residuals (in-sample)	It disregards prediction error; it does not weight errors by confidence	Out-of-sample (cross- validated): Explained variance, accuracy, F1, cross-entropy

causal relationships. Consider the standard multivariate linear specification applied in econometric studies,

$$y_t = \alpha + \sum_{i=1}^{I} \beta_i X_{t,i} + \sum_{j=1}^{J} \gamma_j Z_{t,j} + \varepsilon_t \tag{1}$$

where $\{X_{t,i}\}$ are the observations of variables that explain $\{y_t\}$, and $\{Z_{t,i}\}$ are the observations from control variables, whose effect on $\{y_t\}$ we do not wish to attribute to $\{X_{t,i}\}$. In other words, the goal of these analyses is to adjudicate to $\{X_{t,i}\}$ the variance of $\{y_t\}$ in-sample, while controlling for the variance adjudicated to $\{Z_{t,j}\}$. This ceteris paribus argument has many advantages when the specification is correct. When the model is mis-specified, however, the variance adjudication is biased (Clarke 2005; Pearl 2009). It is important to recognize that econometrics' goal of variance adjudication is not necessarily compatible with ML's goal of out-of-sample forecasting (Mullainathan and Spiess 2017). Variance adjudication involves computing the best linear unbiased estimator (BLUE), even though there may be biased estimators with lower mean-squared prediction error.

In contrast, ML algorithms are designed for out-of-sample forecasting.[7] As we saw in an earlier section, this forecasting goal can materialize into five alternative use cases. For instance, under the existence use case, a researcher could feed large amounts of microstructural trade data into an ML algorithm in order to deduce whether flash crashes may be the result of some (yet unknown) mechanism or whether they may be unpredictable black swans. If ML algorithms cannot predict flash crashes with some confidence, then it is unlikely that a theory of flash crashes will emerge. The retriever use case is particularly valuable for outlier detection. Standard regression models are susceptible to the presence of outliers. For example, in regression-based, asset-pricing studies, a few outliers can tilt the fair-value line, making one subset of the investment universe appear to be cheaper than reality. Unless those outliers are detected and removed, large portions of the portfolio may be mispriced. For an application of the reductionist use case, researchers could use ML algorithms to classify stocks according to a large variety of features, not just the economic sector they belong to, leading to a better understanding of how companies are interconnected. Varian (2014) provides a strong argument for using ML methods to complement econometric methods in causal studies. Under Varian's view, even a black box approach to ML can be helpful for developing economic theories. Finally, researchers can apply the importance use case to the study of predictive financial variables, as demonstrated by Easley et al. (2018) on a survey of microstructural models.

Choosing one of these goals in advance and declaring a clear research plan before the analysis begins is paramount. A poorly designed ML study, where the questions and goals are not clearly stated at the onset, is likely to lead to false discoveries. I expand on this point in a later section.

Outlier Detection

The visualization of observations often leads to the discovery of datapoints that do not appear to be generated by the same process as the rest. These observations, called outliers, are particularly problematic because they have a strong influence on the fitted model. Applied economics and finance research traditionally uses a number of ad hoc techniques to correct for the presence of outliers. Winsorizing involves setting a cap and a floor on the observations in a dataset, where any observation in excess of the cap/floor is reset to the value of the cap/floor. Trimming also establishes a cap and floor, with the difference that observations exceeding the cap or floor are removed.[8] A limitation of these methods is that they either rely on heuristics or they make an assumption regarding the distribution of the data.

Many ML algorithms can be used to detect outliers, taking into account the data's complex structure and using that knowledge to derive the probability that a particular observation is anomalous. For example, clustering algorithms will put outliers in their own cluster. Classifiers will learn from labeled examples how to recognize which combinations of values characterize an outlier. Tree-based models are particularly resilient to outliers, an instance of which are isolation forests (Liu, Ting, and Zhou 2008).[9] This property makes trees useful in recognizing outliers because the outliers will not bias the estimation of the probability that a particular observation is an outlier.

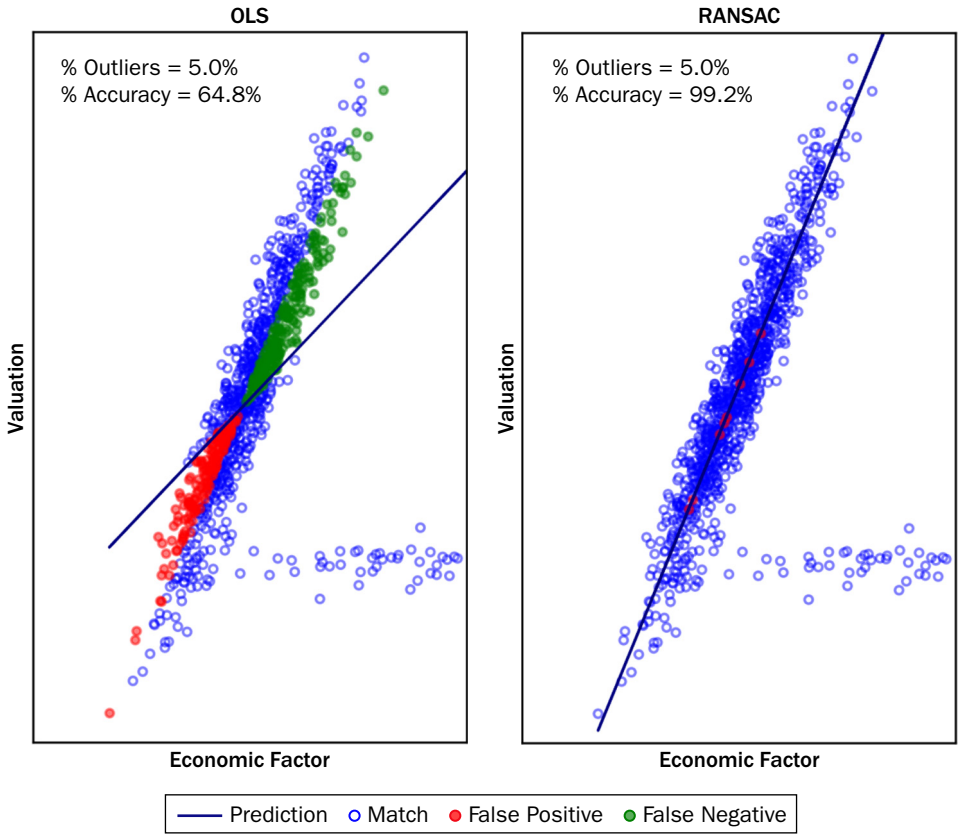
[7] Some econometric models have a forecasting specification, in which case the goal is to adjudicate the future variance. However, those econometric models are not typically fit to maximize out-of-sample forecasting power by means of hyper-parameter tuning via cross-validation or similar methods.

[8] Other methods, like Chauvenet's criterion, Mahalanobis distance, or Dixon's Q test, often rely on the assumption of normality to determine what constitutes an extraordinary event.

[9] For details of Scikit-Learn's implementation of isolation forests, see http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html.

EXHIBIT 2

Estimation in the Presence of Outliers: OLS vs. RANSAC



NOTES: In this cross-sectional regression, outliers account for 5% of the sample cause OLS to misclassify 35.2% of the observations. In contrast, if we apply OLS excluding the outliers detected by RANSAC, the number of misclassified observations is only 0.8% (mostly borderline cases).

In the context of ML regression, a notable example of robust estimation is the random sample consensus (RANSAC) method. This iterative method splits the dataset between inliers and outliers such that the regression may be conducted on the inliers only. RANSAC has been used successfully in computer vision and image processing problems, because it is able to handle a large proportion of outliers in the input (Fischler and Bolles 1981). For example, Exhibit 2 shows how, in the context of cross-sectional studies, a small number of outliers can lead to the misclassification of a large number of securities. The few outliers tilt the regression line, leading to many securities being falsely labeled as cheap (false positives, in red) or falsely labeled as expensive (false negatives, in green). RANSAC is one of several efficient algorithms that can isolate the outliers within a sample, thereby preventing them from unduly biasing model estimates.

Visualization

Until recently, economists had access to a small number of relatively short history datasets. Those datasets were typically visualized using time plots (one variable over time), scatter plots (one variable against another), pie charts (categorical distributions), histograms (numerical distributions), or heat maps (one real variable's distribution against two real variables). Plotting the data is useful to search for problems like

outliers, missing information, or quality issues. The visual patterns that emerge from exploratory data analysis can suggest the formulation of hypotheses, which can then be tested quantitatively. Visualization is also useful for the communication of results.

Traditional visualization tools date back centuries. Since then, datasets have grown in length, complexity, and dimensionality, which demand the use of more modern representation methods. In the particular case of economics and finance, a major source of complexity comes from hierarchical relationships (Simon 1962). Unsupervised learning methods, like the minimum spanning tree algorithm, help represent relationships among variables as networks or tree maps. Supervised methods, like the classification and regression tree (CART) algorithm, visualize how the combination of hierarchies and thresholds can explain outcomes. High-dimensional datasets pose another major challenge because modern economic systems can rarely be represented in two-dimensional or three-dimensional plots. The *t*-distributed stochastic neighbor embedding (*t*-SNE) algorithm is a nonlinear dimensionality reduction technique that embeds a high-dimensional space into two-dimensional or three-dimensional scatter plots such that similar objects are plotted nearby and dissimilar objects are plotted far apart.

Feature Extraction

Feature extraction in ML consists of selecting from the set of features the minimum subset that allows us to achieve our goal. Feature extraction addresses two problems. The first problem is the curse of dimensionality. Suppose we are trying to predict L different labels (i.e., outputs) using N features (i.e., inputs), where each feature can adopt V different values. Even if observations are uniformly distributed across all feature combinations and labels, we would require a number of observations T, where $T > LV^N$, such that we may have one label example per feature combination. For a fixed sample length, T, allowing N to grow means that a large portion of the feature combinations will be fit to a small number of observations.

The second problem addressed by feature extraction is multicollinearity. In an ideal regression model, each of the feature variables is highly correlated to the predicted variable, and the feature variables are uncorrelated with each other. When feature variables are correlated with each other, regression models cannot distinguish between them and the variance of the predicted variable cannot be robustly adjudicated. This correlation among features does not cause problems for forecasting the endogenous variable; however, it makes it difficult to select important features or to test hypotheses.

A traditional econometric solution to both problems is to extract the principal components from the features (inputs) using principal components analysis (PCA), invented in 1901 by Karl Pearson (Pearson 1901). PCA computes the linear combinations of features that are orthogonal to each other and explains most of the variance in the normalized original feature space. For example, Stock and Watson (1998) apply PCA to 224 highly correlated macroeconomic time series and use the leading principal components to predict US industrial production and inflation.

Despite its many properties, PCA has four critical shortcomings. First, it will miss nonlinear interactions between the features. Using the macroeconomic prediction example given earlier, suppose the sample combines observations from a high-volatility regime with observations from a low-volatility regime. Economic fluctuations from the high-volatility regime envelop those from the low-volatility regime, where the distribution of both is separated by a hyperellipsoid. We may want to split the observations from both regimes and study their interactions separately. A PCA analysis cannot do that because the separating hyperellipsoid is not a linear function of the explanatory variables. One solution is to apply the kernel-PCA algorithm, which is a generalization

of PCA using kernel methods (rather than linear algebra). In our example, the kernel-PCA will add one dimension that separates the low-volatility regime observations from the high-volatility regime observations, offering us the possibility of modeling both subsamples separately.

A second limitation of PCA is that it extracts features with no knowledge of the predicted variable (i.e., it is an unsupervised method). That can be problematic in the context of econometrics in particular and supervised learning in general. The PCA transformation is useful only if the predicted variable is highly correlated with the principal components. That is not generally the case. For example, suppose we wish to predict default probabilities in a bond portfolio. We apply PCA to the features, with the result that the marginal distribution of each principal component mixes all outcomes. It would have been useful to extract features that allow us to discriminate best between defaulted and nondefaulted bonds. The linear discriminant analysis (LDA) method finds the linear combination of features that best separates outcomes.

A third caveat of PCA is that extracting principal components involves a change of basis. The resulting extracted features may not be economically intuitive. One ML solution is to apply a bi-clustering algorithm on the correlation matrix of the features, which clusters together features that are mutually redundant. Visually, a clustered correlation matrix is as close to a block diagonal correlation matrix as possible, without a change of basis. The analysis can then be carried out on clustered features rather than principal components.

A fourth caveat of PCA is that the eigenvectors associated with the smaller eigenvalues cannot be robustly estimated. This is particularly true in the context of financial covariance matrices because of the low signal-to-noise ratio. One way to address this problem is to shrink the covariance matrix (Ledoit and Wolf 2004); however, that will remove noise at the expense of removing signal. A second possibility is to regularize the eigenvectors, like the sparse PCA method does. A third possibility is to identify what eigenvalues are associated with noise and shrink only those (Laloux et al. 2000). It is essential that financial researchers apply de-noising and de-toning procedures to financial covariance matrix so as to prevent discoveries that are supported by noise rather than by signal (López de Prado 2019).

Regression

At this stage, the researcher proposes a specification that allows achievement of the stated goal based on the hypothesis suggested by the visualization of the data, taking advantage of the features extracted. In most econometric studies, that means using an algebraic specification like a multivariate linear system of equations with interaction effects. Algebraic functions cannot model complex data patterns like nonlinear relations that exhibit discontinuities (e.g., an activation threshold or regime switches) or topological structures (e.g., hierarchical dependencies with varying degrees of density). In general, it is very difficult to specify a closed-form system of algebraic equations that is able to reflect that complexity.

Let us illustrate this point with an example. Consider the number of possible interaction effects given a number of features. A single equation with N features could contain up to $2^N - N - 1$ multiplicative interaction effects. For 2 features, the only multiplicative interaction effect is x_1x_2; however, for 3 features we have x_1x_2, x_1x_3, x_2x_3, $x_1x_2x_3$. For a mere 10 features, the number of multiplicative interactions rises to 1,013. This does not include other forms of algebraic interaction like x_1/x_2, $x_1\sqrt{x_2}$, $x_1|x_2|$ or nonalgebraic interactions like $x_1\sin[x_2]$, $x_1\log[x_2]$, $\max\{x_1, x_2\}$, etc. An equation with a wide range of interaction effects may exhaust all the degrees of freedom (a saturated model). In practice there is no reason to believe that interaction effects are as simple as x_1x_2, and it is easy for a researcher to omit some of them.

Unlike ML algorithms, econometric models do not learn the structure of the data via a specification search (at least not a large-scale search). The a priori specification choices of researchers can, thus, easily miss a few interaction effects, leaving their econometric model mis-specified. An experiment can clarify what learning the structure of the data means. Suppose that the true process that generates y is $y_t = x_{1,t} + x_{2,t} + x_{1,t}x_{2,t} + \varepsilon_t$, where $\{x_{1,t}\}$, $\{x_{2,t}\}$, and ε_t are independent and identically distributed normal random variables. An economic theory may correctly state that $y_t = f[x_{1,t}, x_{2,t}]$, where $f[\cdot]$ is some real function of $x_{1,t}$ and $x_{2,t}$. If the researcher misses the interaction effect, she will fit the specification $y_t = x_{1,t} + x_{2,t} + \varepsilon_t$. Even though the model incorporates the two theorized variables, the consequences of missing an interaction effect are dramatic. Exhibit 3 compares the predicted y_t with the actual y_t. The out-of-sample correlation between the two is only 0.04. The researcher may falsely conclude that $x_{1,t}$ and $x_{2,t}$ do not explain y_t, discarding a true theory (a type II error).

We can pass the exact same input variables to an ML algorithm and verify whether it recognizes the existence of the interaction effect without our direction. To see how that is possible, consider one of the simplest ML algorithms, the decision tree. This divide-and-conquer algorithm recursively partitions a dataset with complex patterns into subsets with simple patterns, which can then be fit independently with simple linear specifications. For this algorithm, learning means finding the splits that minimize the complexity of the subsets.

Exhibit 4 presents a tree example derived from one run of the decision tree algorithm. In the first split, it isolates 140 observations with high x_2 value, which will be regressed on their own. It splits the remaining 760 observations in terms of their x_1 value, where the 260 with higher x_1 value will be regressed on their own, etc. The algorithm alternates the x_1 splits with x_2 splits because it has recognized that there is an interaction effect involving these two variables.

Exhibit 5 continues the experiment, where this time we fit the same data using a decision tree. Like before, we do not inform the algorithm about the presence of interaction effects. Unlike before, the algorithm has learned about the existence of the $x_{1,t}x_{2,t}$ effect, yielding an out-of-sample correlation between the predicted y_t and the actual y_t of 0.85.

Finally, we repeat this experiment using a random forest (a bootstrap aggregation of decision trees, where candidate features are randomly drawn at each tree level). Exhibit 6 plots the results. The out-of-sample correlation between the predicted y_t and the actual y_t rises to 0.98. The three algorithms (linear regression, decision tree, and random forest) receive the same input data, but they perform very differently in the presence of unknown interaction effects. Appendix A provides the code used to implement this experiment.

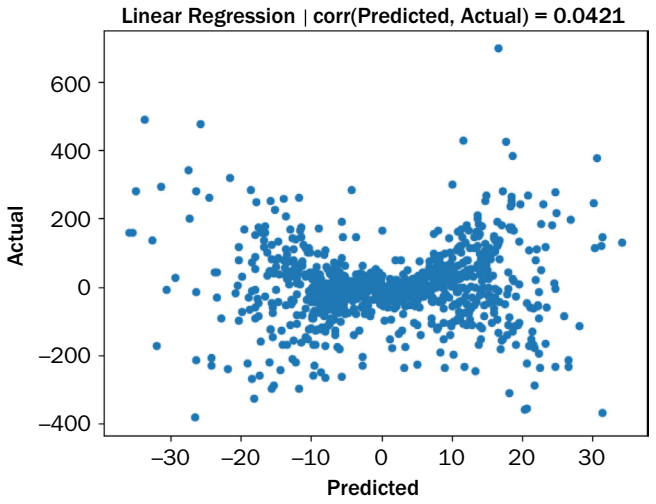
This experiment evidences a key disadvantage of econometric regression models: The researcher must get the specification right or the study will lead to incorrect conclusions. Unfortunately, given how dynamic, complex, and interconnected economic systems are, econometric specifications likely omit important characteristics of the data. Under these circumstances, ML's ability to learn those characteristics makes it perform better than econometric approaches.

Classification

Researchers are sometimes confronted with the problem of predicting a discrete variable that can take a limited number of categorical values. In addition, a continuous variable may be discretized in order to predict a small number of possible outcomes on a larger number of observations per outcome. A classifier is an algorithm that is fit on a training sample, consisting of features associated with outcomes in the form of categories. Once the algorithm has been trained, it can form predictions based

EXHIBIT 3

Out-of-Sample Predictions Using a Linear Regression



on new, previously unseen observations for which the outcome is yet unknown.

Most econometrics textbooks describe two solutions for classification problems: probit (Bliss 1934) and logit (Berkson 1944). The main difference between these models is the shape of the function used to fit the binary outcomes: the inverse cumulative standard normal distribution function in the case of probit and the logarithm of the odds ratio in the case of logit. One disadvantage of logit and probit is their linear specification. A second disadvantage is that the explanatory variables must be real valued. Treatment of categorical (e.g., sectors of the economy) or ordinal (e.g., credit ratings) regressors requires encoding through dummy variables. A third disadvantage is that the model may be biased by imbalanced classes, where one of the outcomes is much more prevalent than

EXHIBIT 4

Splits from a Regression Tree

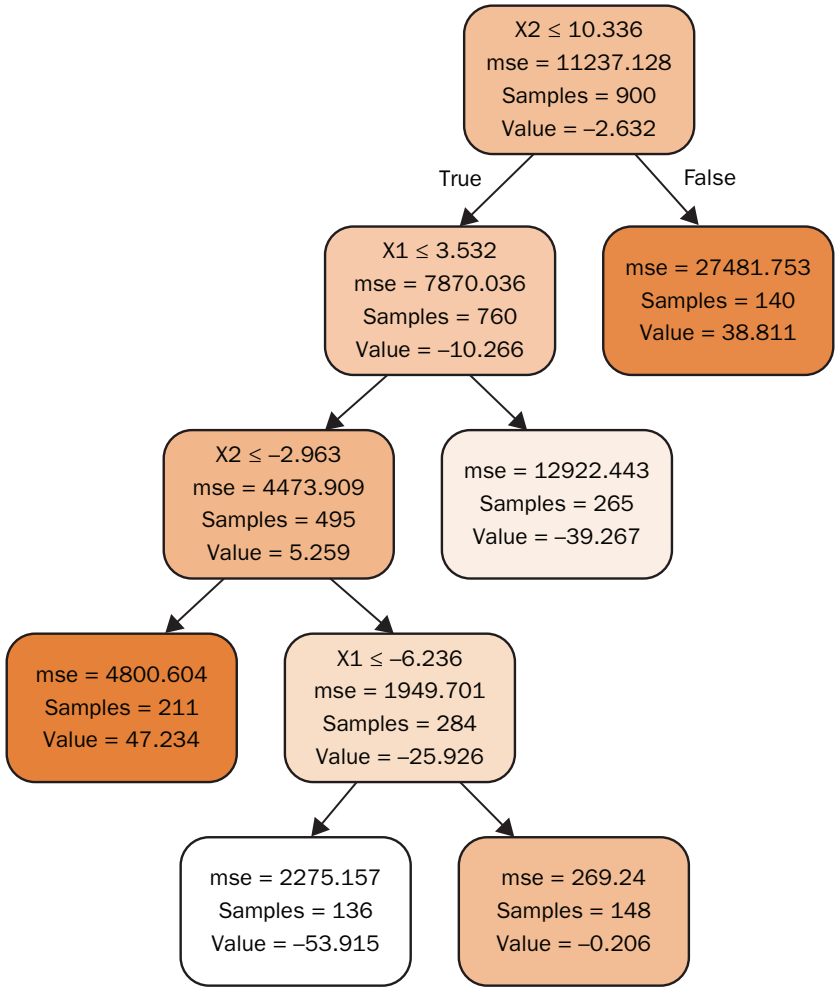


EXHIBIT 5

Out-of-Sample Predictions Using a Decision Tree

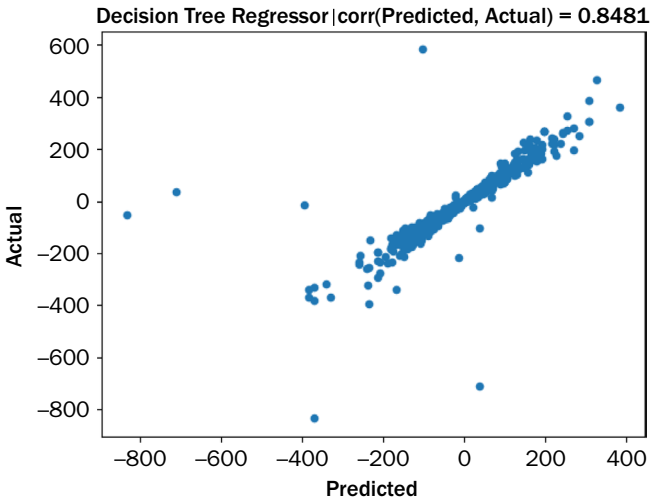
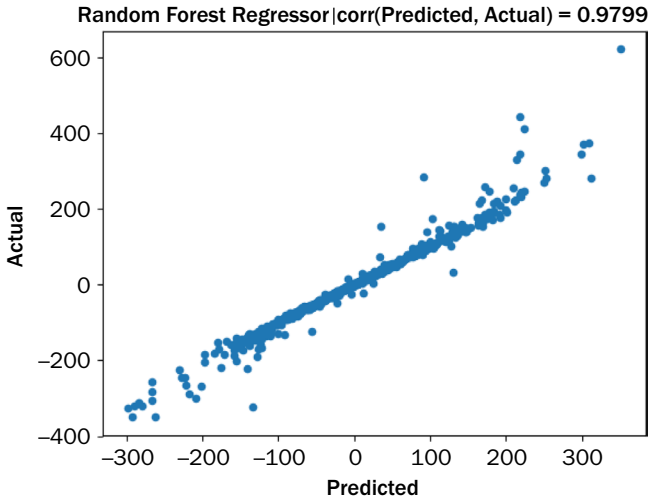


EXHIBIT 6

Out-of-Sample Predictions Using a Random Forest



the other. One solution is to oversample the minority class, as the Synthetic Minority Oversampling Technique (SMOTE) does (see Chawla et al. 2002). Another solution is to undersample the majority class, as the near-miss algorithm does (see Mani and Zhang 2003). Both SMOTE and near-miss rely on K-nearest neighbor (KNN), an ML algorithm.

Compare an algorithm that forecasted a price change of 1 for a realized price change of 3 with another algorithm that forecasted a price change of –1 for a realized price change of 1. Both algorithms made an error of 2, but only the second lost money. In finance, predicting the sign is often more important than predicting the size. Failing to predict the size is an opportunity loss, but failing to predict the sign is an actual loss. In addition, it is common in finance to find that the sign and size of an outcome depend on different features. The sign of an outcome is often more important than its size, making classification a critical task for researchers. In this sense, an econometrics focus on regression techniques (at the expense of classification techniques) is misplaced (for a discussion of bet sizing, see López de Prado 2018a; 2019).

The ML toolkit offers econometricians numerous alternatives that perform proper statistical classification tasks. Of particular interest are the so-called probabilistic classifiers that output the distribution of probability across all classes for each new observation. Hence the researcher receives a forecast and the confidence associated with that forecast. Popular examples of probabilistic classifiers include naïve Bayes, decision trees, random forests, k-NN, adaBoost, and neural networks. Note that most classifying algorithms can be used for regression problems, just as in the previous section where we utilized decision trees and random forests in a regression problem. The different use depends on the objective function used to fit the model, (e.g., the minimization of the sum of squared residuals in the case of a regression problem or maximum information gain in the case of a classification problem).

Feature Importance

A common problem every econometrician faces is the selection of the model's specification. Solving this problem requires making two decisions at once: first, selecting the set of variables involved in a phenomenon, as discussed in the section dedicated to feature extraction, and second, choosing a functional form that binds those variables together (potentially including nonlinearities and feature interactions). There is no reason to believe that multivariate linear regression offers the best answer to both questions. It is likely that some of the complex interactions between variables in traditional datasets may have been missed by traditional econometric methods.

One unique advantage that ML has over econometrics is that ML allows us to decouple the two previously mentioned decisions, following this sequence: (1) We can assess the importance of a large number of features before settling for a particular functional form; (2) once the important features have been isolated, we can focus our attention on identifying the data-generating process most likely to bind them, with the help of a causal graph; and (3) the causal graph informs the structural equation model, with special attention to including all confounders while avoiding all colliders. In this section, I will explain how ML selects what features are important.

In the context of econometric analyses, researchers assess a feature's importance through its *p*-value, an invention that dates back to the 1700s (Brian and Jaisson 2007). This is the probability that, if the true coefficient of a feature is zero, we could have obtained a result equal or more extreme than the one we have observed. A first caveat of the *p*-value is that it relies on the strong assumption that the model is correctly specified. This assumption is not necessarily true, for example, if the researcher omitted a confounder, included a collider, or chose the wrong functional form (Pearl 2009). Hence, a *p*-value could be low even though the true value of the coefficient is zero (a false positive), or the *p*-value could be high even though the value of the coefficient is not zero (a false negative). The misuse of *p*-values is so widespread that the American Statistical Association has discouraged their use going forward as an established statistical significance (Wasserstein, Schirm, and Lazar 2019). This brings into question decades of empirical research in finance.

To address this first caveat, ML studies determine the significance (importance, in ML parlance) of features using computational methods. For example, a popular feature of importance analysis for tree-based algorithms is the MDI method. At each node, the algorithm selects the feature that splits the subset into two less impure subsets, in the sense that labels are less mixed. Once the algorithm has completed the task, we can derive how much of the overall impurity decrease is attributed to each feature. An advantage of MDI over econometric hypothesis testing is that MDI's computational nature avoids the need for distributional assumptions that could be false.

An earlier section listed the five use cases of ML in science. When discussing the importance use case, I described how scientists apply the MDA method to discover the variables that should be part of the theory. A key difference that separates MDA from both *p*-value and MDI is that MDA assesses significance out of sample. It does so by comparing the cross-validated performance of the model with the variables' observations against the cross-validated performance of the model with shuffled observations (shuffling the observations of one variable at a time). A disadvantage of both *p*-value and MDI is that a variable that is significant for explanatory purposes (in-sample) may be irrelevant for forecasting purposes (out-of-sample).

Another ML technique for assessing feature importance is Shapley values. Shapley values treat features like players in a game. The players form coalitions to achieve an outcome. Shapley values' attribution of the outcome to the individual players satisfies several mathematical properties that make it preferable over *p*-values or MDI. Shapley values also offer a way to estimate the strength of interaction effects (López de Prado 2020).

A second caveat of *p*-values is that, for highly correlated explanatory variables (multicollinearity), *p*-values are nonrobust and often biased. The reason is that the system does not have enough information to discriminate among redundant explanatory variables, leading to substitution effects among related *p*-values. A substitution effect occurs when the importance of a variable cannot be assessed by removing that variable because a substitute variable takes the role of the removed variable, thus leading the researcher to believe that the removed variable was unimportant. An ML solution that deals with substitution effects is to cluster together interdependent

features, and derive MDI and MDA per cluster, rather than MDI and MDA per feature. For a detailed discussion of ML alternatives to *p*-values, see López de Prado (2019).

Model Selection

Among models with comparable explanatory power, parsimonious answers are preferred to overly complex ones (the Occam's razor principle). The traditional econometric method for selecting the most parsimonious model is called the stepwise algorithm (Efroymson 1960). The stepwise algorithm is typically implemented in one of three variants: forward selection, backward elimination, and bidirectional elimination. In the forward selection algorithm, the researcher starts with zero explanatory variables out of N candidates. The N candidates are tested individually, and the one candidate that yields the highest improvement (if any) is added. This procedure is repeated until model improvements cease to be statistically significant. In the backward elimination algorithm, the researcher starts with all N candidates and sequentially eliminates the one variable responsible for the lowest improvement. This procedure is repeated until all remaining variables contribute significantly to the model's fit. In the presence of multicollinearity, backward elimination is not advised because the significance of individual regressors within a large model cannot be determined precisely. In the bidirectional elimination algorithm, the researcher alternates a forward selection step with a backward selection step.

The stepwise approach has received two main criticisms: (1) as is often the case in econometric analysis, the dataset used to train the model (train set) also is used to assess the model (test set), hence all decisions are based on in-sample statistics without regard for their effect on out-of-sample model performance; and (2) the procedure does not control for selection bias under multiple testing. The implication of these criticisms is that econometric models selected by a stepwise procedure are typically train-set overfit and test-set overfit. Let us see why.

A model is said to be overfit when it performs well in sample but performs poorly out of sample. The in sample can be divided into a train set and a test set, resulting in two forms of overfitting. Train-set overfitting occurs when a model so closely fits the train set that it misidentifies noise for signal. Test-set overfitting occurs when a researcher assesses the performance of a model on the test set multiple times and picks the best result, hence concealing the existence of worse outcomes. Furthermore, a model is test-set hyperfit when a higher authority (e.g., a journal) picks the best model among a multiplicity of overfit models (e.g., author submissions). For a discussion of test-set overfitting in the context of investment strategies, see Bailey et al. (2014), López de Prado (2018a, 2018b, 2018c, 2018d, 2019), and López de Prado and Lewis (2018).

One way that ML methods prevent train-set overfitting via conservative model selection is called *regularization*. Regularization works by introducing a penalty for complexity such that the model adds complexity only if it is warranted by a significant gain in explanatory power. ML textbooks typically discuss three regularization methods: (1) Tikhonov regularization (or ridge regression), (2) least absolute shrinkage and selection operator (LASSO), and (3) elastic net. Tikhonov regularization minimizes the sum of squared errors with an L_2 penalty on the solution vector, thus giving preference to solutions with a smaller norm. LASSO introduces an L_1 penalty as an inequality constraint, which forces to zero the coefficient of irrelevant regressors. Elastic net combines the L_1 (LASSO) and L_2 (Tikhonov) penalties. As usual with ML methods, the hyper-parameters that control the penalization functions are found through cross-validation, hence minimizing the model's error on the test set (generalization error). Regularization is of particular interest to econometricians because it enforces a sparsity constraint on the model, which makes it simpler and more interpretable.

Other ML methods to prevent train-set overfitting include the following: (a) early stopping exits an optimizer as soon as an increase in generalization error is detected or marginal gains in the model's predictive power do not exceed a given threshold; (b) pruning reduces the size of decision trees by eliminating splits associated with minor information gains; (c) dropout randomly deactivates units from a neural network's hidden layers, forcing the remaining units to become more generally useful (less fine-tuned); (d) kernel-based algorithms, like support vector machines, use the bandwidth parameter to determine the smoothness of the fit; and (e) bootstrap aggregation (bagging) averages the forecasts from many algorithms of the same class, where each algorithm has been fit to a sample randomly drawn with replacement. Bagging addresses train-set overfitting by reducing the variance of the estimation error.[10]

Chen and Pearl (2013) show that the correct specification of an econometric model requires knowledge of the causal graph that represents the interactions between all the variables involved in a phenomenon. Unfortunately, the econometric toolkit does not include methods for identifying causal graphs from a set of causal graphs consistent with the observed probability distribution. The implication is that econometric models are often mis-specified because of the omission of confounding variables, inclusion of colliders, etc. Luckily, the ML toolkit offers econometricians several algorithms for exploratory causal analysis, such as the PC algorithm introduced by Spirtes and Glymour (1991). For a review of causal discovery methods based on graphical models, see Glymour, Zhang, and Spirtes (2019).

Model Validation

A model's goodness of fit estimates the discrepancy between the values expected by the model and the observed values. Dating back to at least 1921, a popular choice in econometrics textbooks has been the coefficient of determination, also known as R^2 (Wright 1921). The R^2 computes in sample (within the train set) the portion of the dependent variable's variance that is explained by the model. The R^2 spuriously rises with the number of independent variables, and the so-called adjusted R^2 corrects for that inflation. Testing the residuals for potential violation of the model's assumptions is as important as achieving a significant R^2. Common tests include Jarque-Bera's normality test, Durbin-Watson's autocorrelation test, and White's heteroscedasticity test.

ML algorithms evaluate the goodness of fit using a wide range of methods. A key distinction between econometric goodness of fit and ML goodness of fit is that the former almost always evaluates performance of a model in sample (in the train set), whereas the latter almost always evaluates the performance of a model out of sample (in the test set), through cross-validation.

For regression problems, common methods used in ML include explained variance, mean absolute error, and median absolute error. For binary classification problems, useful scores include: (1) precision, which computes the proportion of true positives among all predicted positives; (2) recall, which computes the proportion of predicted positives among all true positives; and (3) the F1 score, which is the harmonic average of precision and recall and controls for situations of low precision with high recall (too many false alarms, type I errors) or high precision with low recall (too many misses, type II errors). For multiclass classification problems, useful scores include: (1) accuracy, which computes the proportion of correct predictions, and (2) cross-entropy, which extends the notion of accuracy to incorporate the model's

[10] Under certain conditions, bagging also can reduce the bias of an algorithm because the accuracy of the ensemble can be greater than the average accuracy of the individual algorithms. Another important property of bagging is parallelism, which means that all model instances can be fit simultaneously.

confidence. In doing so, cross-entropy penalizes a model that makes bad predictions with high confidence over a model that makes bad predictions with low confidence.

Suppose an investment model makes many good predictions with low confidence and a few bad predictions with high confidence. Bet sizes are determined as a function of confidence, so this model induces us to take more risk on bad predictions. Traditional goodness-of-fit statistics will assess the model as valuable, even though the model generates net losses. Cross-entropy is particularly useful in recognizing models that look good on paper but are likely to fail in practice.

CONCLUSION

An unresolved contradiction is at the heart of financial economics. First, we are told that markets are extremely efficient. Second, we are told that simple regressions suffice to extract billions of dollars in annual profits. If the first statement is true, then econometric models are not sophisticated enough to recognize complex inefficiencies, and findings are spurious by design. If the second statement is true, then economists must explain why investment factors discovered through econometric methods have performed poorly (Arnott et al. 2019), while funds that do not rely on econometric methods are among the best performing in history. One possible explanation is that markets are too efficient for econometrics but are sufficiently inefficient for more modern statistical approaches.

ML offers the opportunity to gain insight from new datasets that cannot be modeled with econometric methods and old datasets that incorporate complex relationships that are still unexplored. Key strengths of ML methodologies include its focus on out-of-sample predictability over variance adjudication; its usage of computational methods to avoid relying on (potentially unrealistic) assumptions; its ability to learn complex specifications, including nonlinear, hierarchical, and noncontinuous interaction effects in a high-dimensional space; and its feature of importance analysis robust to multicollinearity and decoupled from model selection.

The applications of ML to finance extend far beyond the uses described in this article. They include problems outside the traditional scope of econometric methods, such as portfolio construction, bet sizing, complex optimization, sentiment analysis, automation, detection of false investment strategies, graph-theoretic representation of economic systems, and many others (López de Prado 2016, 2018c). In this article I have focused on use cases where ML can complement the use of econometric methods. For instance, we could model complex text, video, and transactional data with ML methods to predict earnings and then link those predicted earnings to prices via a simple (even linear) econometric model.

Finance is not a plug-and-play subject as it relates to ML. Modeling financial series is harder than driving cars or recognizing faces. The numerical power and functional flexibility of ML ensure that it will always find a pattern in the data, even if that pattern is a fluke rather than the result of a persistent phenomenon. Scientists across all disciplines monitor and evaluate the risk of data mining through established methodologies. An oracle approach to ML, where algorithms are developed to form predictions divorced from all economic theory, is likely to yield false discoveries. ML is not a substitute for economic theory but rather a powerful tool for building modern economic theories.

This article has shown that, for every step in the econometric analysis, there is an analogous step in the ML research process. The mapping presented in this article provides a roadmap for econometricians who wish to expand their quantitative toolkit.

APPENDIX A

EXPERIMENT ON INTERACTION EFFECTS

We have conducted an experiment that examines the performance of three regression methods in the presence of interaction effects: multivariate linear, decision tree, and random forest. Function interEffect performs three tasks. First, it creates a random matrix of size 1000x2 by drawing from a normal distribution. The endogenous variable is computed as $y_t = x_{1,t} + x_{2,t} + x_{1,t}x_{2,t} + \varepsilon_t$, where $\{x_{1,t}\}$, $\{x_{2,t}\}$, and ε_t are independent and identically distributed normal random variables. Second, we split the dataset into training and test sets and perform a k-fold cross-validation. For each train set, we fit $\{y_t\}$ against $\{x_{1,t}\}$ and $\{x_{2,t}\}$ and use that fit on the test-set values of $\{x_{1,t}\}$ and $\{x_{2,t}\}$ to predict the test-set values of $\{y_t\}$. These are out-of-sample predictions in the sense that the test-set values were not used to fit the model. This process is repeated cvSplits times. Third, we compare the predicted values with the actual values out of sample and report the correlation for each method.

SNIPPET 1

Python Code that Implements the Experiment on Interaction Effects

```
import matplotlib.pyplot as mpl
import numpy as np,pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import KFold
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
def interEffect(reg,cvSplits=10):
  kf=KFold(n_splits=cvSplits)
  out=pd.DataFrame()
  1) Create data
  x=np.random.normal(0,10,size=(1000,2))
  y=x[:,0]+x[:,1]+x[:,0]*x[:,1]+np.random.normal(size=(1000,))
  2) Fit in sample, predict out of sample
  for train_index,test_index in kf.split(x):
    x_train,x_test=x[train_index],x[test_index]
    y_train,y_test=y[train_index],y[test_index]
    reg_=reg.fit(x_train,y_train)
    y_pred=reg_.predict(x_test)
    out_=pd.DataFrame({'Predicted':y_pred,'Actual':y_test})
    out=out.append(out_,ignore_index=True)
  3) Report results
  out.plot.scatter(x='Predicted',y='Actual')
  corr=out.corr().iloc[0,1].round(4)
  mpl.title(type(reg).__name__+' | corr(Predicted,Actual)='+str(corr))
  mpl.savefig(type(reg).__name__+'.png')
  return
#- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
if __name__=='__main__':
  interEffect(reg=LinearRegression())
  interEffect(reg=DecisionTreeRegressor())
  interEffect(reg=RandomForestRegressor(n_estimators=1000))
```

REFERENCES

Arnott, R., C. Harvey, V. Kalesnik, and J. Linnainmaa. 2019. "Alice's Adventures in Factorland: Three Blunders That Plague Factor Investing." *The Journal of Portfolio Management* 45 (4): 18–36.

Athey, S. 2015. "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. ACM. https://dl.acm.org/doi/10.1145/2783258.2785466.

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu. 2014. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-Of-Sample Performance." *Notices of the American Mathematical Society* 61 (5): 458–471.

Balding, D., M. Bishop, and C. Cannings. *Handbook of Statistical Genetics*, 3rd ed. Wiley, 2007. https://www.wiley.com/en-us/Handbook+of+Statistical+Genetics%2C+3rd+Edition-p-9780470058305.

Berkson, J. 1944. "Application of the Logistic Function to Bio-Assay." *Journal of the American Statistical Association* 39 (227): 357–365.

Bliss, C. 1934. "The Method of Probits." *Science* 79 (2037): 38–39.

Brian, E., and M. Jaisson. "Physico-Theology and Mathematics (1710–1794)." In *The Descent of Human Sex Ratio at Birth*, pp. 1–25. Springer Science & Business Media, 2007. https://link.springer.com/chapter/10.1007/978-1-4020-6036-6_1.

Chawla, N., K. Bowyer, L. Hall, and W. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16: 321–357.

Chen, B., and J. Pearl. 2013. "Regression and Causation: A Critical Examination of Six Econometrics Textbooks." *Real-World Economics Review* 65: 2–20.

Clarke, K. A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22 (1): 341–352.

Easley, D., and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, 1st ed. Cambridge, UK: Cambridge University Press, 2010.

Easley, D., M. López de Prado, and M. O'Hara. 2016. "Discerning Information from Trade Data." *Journal of Financial Economics* 120 (2): 269–285. https://www.sciencedirect.com/science/article/abs/pii/S0304405X16000246.

Easley, D., M. López de Prado, M. O'Hara, and Z. Zhang. 2018. "Microstructure in the Machine Age." *The Review of Financial Studies* 37 (7): 3316–3363.

Efron, B., and T. Hastie. *Computer Age Statistical Inference*, 1st ed. Cambridge, UK: Cambridge University Press, 2016.

Efroymson, M. "Multiple Regression Analysis." In *Mathematical Methods for Digital Computers*, 1st ed., edited by A. Ralston and H. Wilf. Wiley, 1960. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=430544.

Einav, L., and J. Levin. 2014. "Economics in the Age of Big Data." *Science* 346 (6210). https://www.science.org/doi/10.1126/science.1243089.

Fischler, M., and R. Bolles. 1981. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography." *Communications of the Association for Computing Machinery* 24 (6): 381–395.

Glymour, C., K. Zhang, and P. Spirtes. 2019. "Review of Causal Discovery Methods Based on Graphical Models." *Frontiers in Genetics* 10: 524.

Greene, W. *Econometric Analysis*, 7th ed. Pearson Education, 2012. https://www.amazon.com/Econometric-Analysis-8th-William-Greene/dp/0134461363.

Gryak, J., R. Haralick, and D. Kahrobaei. 2018. "Solving the Conjugacy Decision Problem via Machine Learning." *Experimental Mathematics* 29 (1): 66–78.

Hayashi, F. *Econometrics*, 1st ed. Princeton University Press, 2000. https://press.princeton.edu/books/hardcover/9780691010182/econometrics.

Hodge, V., and J. Austin. 2004. "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review* 22 (2): 85–126.

Kolanovic, M., and R. Krishnamachari. "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing." J.P. Morgan Quantitative and Derivative Strategy, 2017.

Laloux, L., P. Cizeau, J. P. Bouchaud, and M. Potters. 2000. "Random Matrix Theory and Financial Correlations." *International Journal of Theoretical and Applied Finance* 3 (3): 391–397.

Ledoit, O., and M. Wolf. 2004. "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis* 88 (2): 365–411.

Leontief, W. 1982. "Academic Economics." *Science* 217 (4555): 104–107.

Liu, F., K. Ting, and Z. Zhou. 2008. "Isolation Forest." *Proceedings of the Eighth IEEE International Conference on Data Mining* X: 413–422. https://ieeexplore.ieee.org/document/4781136.

Liu, Y. 2004. "A Comparative Study on Feature Selection Methods for Drug Discovery." *Journal of Chemical Information and Modeling* 44 (5): 1823–1828.

Lochner, M., J. McEwen, H. Peiris, O. Lahav, and M. Winter. 2016. "Photometric Supernova Classification with Machine Learning." *The Astrophysical Journal* 225 (2). https://iopscience.iop.org/article/10.3847/0067-0049/225/2/31.

López de Prado, M. 2016. "Building Diversified Portfolios That Outperform Out-of-Sample." *The Journal of Portfolio Management* 42 (4): 59–69.

——. *Advances in Financial Machine Learning*, 1st ed. Wiley, New York, USA, 2018a.

——. 2018b. "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *The Journal of Financial Data Science* 1 (1): 99–110.

——. "Ten Applications of Financial Machine Learning." Working paper, SSRN, 2018c.

——. "Confidence and Power of the Sharpe Ratio under Multiple Testing." Working paper, SSRN, 2018d.

——. *Machine Learning for Asset Managers*, 1st ed. Cambridge, UK: Cambridge Elements, Cambridge University Press, 2019.

——. "Interpretable Machine Learning: Shapley Values." Working paper, SSRN, 2020.

López de Prado, M., and M. Lewis. 2018. "Detection of False Investment Strategies Using Unsupervised Learning Methods." *Quantitative Finance* 19 (9): 1555–1565.

Mani, I., and I. Zhang. 2003. "KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction." *Proceedings of Workshop on Learning from Imbalanced Datasets*. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1603053.

Molnar, C. *Interpretable Machine Learning: A Guide for Making Black-Box Models Explainable*. 2019, https://christophm.github.io/interpretable-ml-book/.

Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Otto, M. *Chemometrics: Statistics and Computer Application in Analytical Chemistry*, 3rd ed. Wiley, 2016. https://www.wiley.com/en-ae/Chemometrics:+Statistics+and+Computer+Application+in+Analytical+Chemistry,+3rd+Edition-p-9783527340972.

Pearl, J. 2009. "Causal Inference in Statistics: An Overview." *Statistics Surveys* 3: 96–146.

Pearson, K. 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine* 2 (11): 559–572.

Romer, P. "The Trouble with Macroeconomics." Working paper, Paul Romer, 2016.

Schlecht, J., M. Kaplan, K. Barnard, T. Karafet, M. Hammer, and N. Merchant. 2008. "Machine-Learning Approaches for Classifying Haplogroup from Y Chromosome STR Data." *PLOS Computational Biology* 4 (6). https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000093.

Simon, H. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106 (6): 467–482.

SINTEF. "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years." *Science Daily*, May 2013, www.sciencedaily.com/releases/2013/05/130522085217.htm.

Solow, R. "Building a Science of Economics for the Real World." Prepared statement of Robert Solow, Professor Emeritus, Massachusetts Institute of Technology, to the House Committee on Science and Technology, Subcommittee on Investigations and Oversight, July 20, 2010.

Spirtes, P., and C. Glymour. 1991. "An Algorithm for Fast Recovery of Sparse Causal Graphs." *Social Science Computer Review* 9 (1): 62–72.

Stigler, S. 1981. "Gauss and the Invention of Least Squares." *Annals of Statistics* 9 (3): 465–474.

Stock, J., and M. Watson. "Diffusion Indexes." Working paper, NBER, 1998.

Tsay, R. *Multivariate Time Series Analysis: With R and Financial Applications*, 1st ed. Wiley, 2013. https://www.wiley.com/en-ae/Multivariate+Time+Series+Analysis%3A+With+R+and+Financial+Applications-p-9781118617908.

Varian, H. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.

Wasserstein, R., A. Schirm, and N. Lazar. 2019. "Moving to a World Beyond P<0.05." *The American Statistician* 73 (1): 1–19.

Wooldridge, J. *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge, MA: MIT Press, 2010.

Wright, S. 1921. "Correlation and Causation." *Journal of Agricultural Research* 20: 557–585.

Zwolenski, M., and L. Weatherill. 2014. "The Digital Universe: Rich Data and the Increasing Value of the Internet of Things. *Journal of Telecommunications and the Digital Economy* 2 (3): 1–9. Available at: https://search.informit.org/doi/10.3316/INFORMIT.678436300116927.