Wasserbacher, Helmut; Spindler, Martin

Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls

Digital Finance

Provided in Cooperation with:
Springer Nature

Suggested Citation: Wasserbacher, Helmut; Spindler, Martin (2021) : Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls, Digital Finance, ISSN 2524-6186, Springer International Publishing, Cham, Vol. 4, Iss. 1, pp. 63-88, https://doi.org/10.1007/s42521-021-00046-2

Digital Finance (2022) 4:63–88
https://doi.org/10.1007/s42521-021-00046-2

ORIGINAL ARTICLE

Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls

Helmut Wasserbacher[1] · **Martin Spindler**[2]

Received: 27 May 2021 / Accepted: 17 November 2021 / Published online: 16 December 2021

Abstract

This article is an introduction to machine learning for financial forecasting, planning and analysis (FP&A). Machine learning appears well suited to support FP&A with the highly automated extraction of information from large amounts of data. However, because most traditional machine learning techniques focus on forecasting (prediction), we discuss the particular care that must be taken to avoid the pitfalls of using them for planning and resource allocation (causal inference). While the naive application of machine learning usually fails in this context, the recently developed double machine learning framework can address causal questions of interest. We review the current literature on machine learning in FP&A and illustrate in a simulation study how machine learning can be used for both forecasting and planning. We also investigate how forecasting and planning improve as the number of data points increases.

Keywords Financial planning · Machine learning · Forecasting · Causal machine learning · Big data · Double machine learning

JEL Classification Primary G17 · G31 · C53 · C55

The views and opinions expressed in this document are those of the first author, and do not necessarily reflect the official policy or position of Novartis or any of its officers.

✉ Martin Spindler
martin.spindler@uni-hamburg.de

Helmut Wasserbacher
helmut.wasserbacher@novartis.com

[1] Novartis International AG, Novartis Campus, 4002 Basel, Switzerland

[2] Hamburg Business School, University of Hamburg, Moorweidenstr. 18, 20148 Hamburg, Germany

1 Introduction

Accurate financial forecasts and plans for effective and efficient resource allocation are core deliverables of the finance function in modern companies. Particularly, in volatile or fast-evolving market environments, fast and reliable forecasting and planning are crucial (Becker et al. 2016). High-quality forecasting is among the defining characteristics of strong finance functions (Roos et al. 2020). It is therefore hardly surprising that most larger companies have dedicated teams for financial planning and analysis (FP&A) within their finance function.

The increasing availability of big data, coupled with new analysis techniques, provides an opportunity for FP&A to generate more and better insights at a faster pace, generating more value for the company. Machine learning is a set of techniques developed in computer science and statistics that appear particularly well suited to this context. The aim of our paper is to show how machine learning can be used for FP&A and which pitfalls can arise in the process. Machine learning has been applied successfully to a variety of predictive tasks, including fraud detection and financial forecasting. Planning and resource allocation, however, represent tasks of a different nature, because they require understanding the effect of an active intervention in a system, such as the market for a product. For this reason, they are causal problems, which are harder to model with machine learning. A large field within machine learning revolves around pattern recognition. Patterns in data, based on correlations, are learned and then used for predictions. In causal tasks, an understanding of the underlying (causal) mechanisms is important when evaluating the effects of interventions (e.g., the implementation of a new business strategy). The emerging field of causal machine learning uses machine learning algorithms for such questions. For instance, the recently developed double machine learning framework reduces the impact of imperfect model specifications, which are hard to avoid in practice in the context of causal analysis.

We structure this paper as follows. In Sect. 2, we briefly review the role of FP&A. Section 3 provides a short, focused overview of machine learning. In particular, we highlight the pitfall of not distinguishing between forecasting and planning. In Sect. 4, we present the results of our literature review, which finds surprisingly few publications of machine learning applications in FP&A. In Sect. 5, we describe and provide the results from a simulation study. We compare a machine learning technique, the lasso, to a linear regression based on the ordinary least squares (OLS) method. In our analysis, we refer back to the distinction between forecasting and planning from Sect. 3, and show how the results differ between the lasso and OLS for both tasks. Finally, we also quantify the benefit of additional data in this simulation.

2 The role of FP&A

Given the importance of financial forecasting, planning and analysis (FP&A) in modern corporations, most larger companies have dedicated teams for these tasks within their finance function, even though the exact organizational design and naming of the department may vary.[1]

The overarching goal of FP&A is to inform and support decisions of management and the board of directors (Oesterreich et al. 2019). FP&A pursues this goal via different routes, helping determine which projects in the company portfolio create value (and are consequently worth funding), and preparing company-wide forecasts and financial plans to ensure that the company can reach its financial goals in the short and long term (Roos et al. 2019). Investments in research and development (R&D) or the expansion of production capacity are balanced with financial obligations to debt holders or equity investors and tax authorities (Brealey et al. 2020). Financial plans are also an important step in the translation of a company's strategic priorities into concrete operational actions. These actions contribute to focusing the organization and the deployed resources behind common goals.

Analyzing the business environment and business dynamics is an integral part of the work performed by FP&A. The insights generated through such analysis can inform the development of forecasts and plans and help in the assessment of how likely these plans are to succeed. During the the execution phase of plans, these insights allow FP&A to understand why actual results may deviate from the plan and to recommend corrective actions. This need for business acumen is likely to continue, even when advanced forecasting methods like those described in this article are used (Möller et al. 2020).

The time horizons considered for financial forecasts and plans usually range from 1 month to several years (Roos et al. 2019; Fischer 2009). The choice of time horizon depends on company-specific circumstances and objectives; for instance, stock-market listed companies typically put additional weight on quarterly figures. In practice, most companies create forecasts and plans for the next fiscal year (sometimes called a budget), which additionally can serve as a management control mechanism (Strauß and Zecher 2013). Rolling forecasts are another form of plan. These are characterized by regular updates, which are typically performed on a monthly or quarterly basis (Hansen 2011).

FP&A relies in large part on quantitative analysis to generate forecasts and plans. Accounting systems are a major internal data source for FP&A (Garrison et al. 2006; Gray and Alles 2015), covering items related to sales (turnover), expenses, and balance sheet positions, which are especially important for cash flow analysis. Other important internal sources of data include those related to human resources (employee numbers, wage costs), supply chain and production (manufacturing costs

[1] In some companies, the (short-term) plans formally expressed in budgets are prepared by controllers (management accountants) within the accounting department (Garrison et al. 2006), while the strategy department formulates the directional (long-term) plans.

at various levels of granularity), and R&D (product development costs, success rates, timelines).

External data sources include market- or product-specific information, such as the size and development of the market and market shares. The exact nature and granularity of these data depends largely on the product or question under analysis, as well as the investment required to access the relevant data (Gray and Alles 2015). For instance, it is not uncommon in the consumer goods industry to have access to transaction-level data (Taddy 2019), covering one's own and competitor products. However, information at this level of specificity is typically used by the marketing and sales department for product-specific tactics. In contrast, FP&A often uses macroeconomic indicators, including GDP, inflation and currency rates.

The development and spread of comprehensive, company-wide IT systems in recent decades has increased the amount and variety of data readily available to FP&A. Increasing digitalization will further accentuate this development, with big data as the crystallizing term. The "Three V's", a common framework to define big data (Laney 2001), allow us to look at the different dimensions that drive this development. First, the amount of information generated, captured and thus accessible for FP&A activities is growing (volume). Second, the speed of information creation and its accessibility is accelerating (velocity); as a consequence, the speed at which information must be analyzed and acted upon increases, too. This calls for automated, real-time analytics and evidence-based planning (Gandomi and Haider 2015). Third, more and more types of information are being gathered or generated and can be analyzed (variety); for instance, stock-market analysts apply sentiment analysis to extract information relevant to stock prices from text documents.

In addition, other dimensions of big data have been proposed (Gandomi and Haider 2015). In the context of FP&A, the additional dimensions of veracity and complexity appear especially relevant. Thanks to the more widespread use of digital tools, the need for data transparency and scrutiny within many companies is increasing, as well. In turn, the need to ensure data quality and reliability is growing (veracity). Moreover, (big) data are generated through multiple sources, both from inside and outside the company. This requires data cleaning, data matching and, ideally centralized storage, which facilitates accessibility (complexity).

As mentioned above, a key output of FP&A is financial forecasts and plans. For data that are more numerous, available more quickly, and are more diverse and of better quality than in the past, FP&A needs to choose adequate tools, such as those provided by machine learning.

3 Introduction to machine learning

While there is no uniform definition of machine learning, it can be described as a collection of methods that automatically build predictions from complex data (Taddy 2019). In essence, machine learning deploys a function-fitting strategy aiming to find a useful approximation of the function that underlies the predictive relationship between input and output data (Hastie et al. 2009). In this search for

patterns in data (Bishop et al. 2006), which, to a large extent, is executed autonomously, machine learning draws on statistical tools and algorithmic approaches from computer science. In particular, machine learning aims to cope with the situation of high-dimensional data. High dimensionality occurs when the number of input variables (independent variables, features) used to predict the output (dependent) variable is large compared to the number of observations available. Classical statistical techniques do not work in this setting (Taddy 2019).

The three broad categories of machine learning are supervised learning, unsupervised learning and reinforcement learning. Supervised learning is concerned with predicting the value of an output variable based on the values of a set of input variables. For this, supervised learning relies on a set of input and output variables that are jointly observed for each data point (Hastie et al. 2009). A practical example is to predict the sales of a product using input variables such as time of the year, price level, advertising expenditures and availability of competitor products. In contrast, unsupervised learning consists only of a set of input observations for which the joint distribution is known. However, there is no observed output (response). The goal is to directly infer the properties of these observations (Hastie et al. 2009). Classifying customers into (previously unknown) customer archetypes based on their observed characteristics such as buying behavior, age, gender and socio-economic status is an example of unsupervised learning. In reinforcement learning, the algorithm performs a trial-and-error search to maximize a numeric reward signal, in direct interaction with its environment (Sutton and Barto 2018). By interacting with its environment, the algorithm creates its own data from which it can learn. Games such as checkers, chess and go are classical examples in which reinforcement learning is applied. Sometimes cited as a fourth category, semi-supervised learning falls between supervised and unsupervised learning, combining a small amount of fully labeled data as in supervised learning and a large amount of unlabeled data as in unsupervised learning. The objective is to improve supervised learning in situations in which labeled data are scarce (Zhu and Goldberg 2009). For the purposes of FP&A objectives, which mostly revolve around producing forecasts from a set of inputs and assumptions, the predominant choice is typically supervised methods.

Machine learning methods appear especially suitable for the core FP&A task of forecasting because of their focus on predictive performance. These methods manage to identify generalizable patterns that work well on new data, i.e., data outside of the training sample (Mullainathan and Spiess 2017). Through their ability to identify complex structures that have not been specified in advance, they lend themselves to support a high degree of automation in the generation of forecasts. This flexibility has the additional advantage that many off-the-shelf algorithms perform surprisingly well on a variety of tasks. In addition, a large selection of machine learning algorithms are available and are technically easy to use (Mullainathan and Spiess 2017), making them attractive for practitioners.

Besides forecasting, the second core task of FP&A is to provide recommendations for the design of financial plans and for potential corrective actions when deviations from plans occur. In statistical terminology, this requires causal inference techniques, which are fundamentally different from forecasting. Consider the trivial example of hotel occupancy rates and room prices (Athey 2018). High room

prices coincide with high occupancy rates. Thus, price variations are strongly predictive of hotel occupancy. If the goal is to make a forecast, we do not need to be concerned with understanding why occupancy was high. However, if we want to recommend an action to increase the occupancy rate (an intervention) or imagine in retrospect what the occupancy rate would have been if the room rates had been different ["counterfactual" (Pearl and Mackenzie 2018) or "potential outcome" (Rubin 2005)], FP&A requires a causal understanding of the business dynamics. To conclude with this example, a plan consisting of a room price increase will not lead to higher occupancy. Most likely, prices have increased in the past in reaction to high demand, which was stimulated by other factors (e.g., the holiday season). While this trivial example seems obvious, it illustrates a major pitfall: many companies struggle in practice to identify truly causal measures for the effectiveness of their promotional activities. Blake et al. (2015) discuss this phenomenon in the context of large-scale field experiments conducted at the e-commerce platform eBay.

For interventional and counterfactual analysis, data-driven approaches need to produce reliable estimates for the parameters that govern the relationship between input and output variables. Machine learning algorithms are typically not built for this purpose. Historically, the machine learning community has pursued the goal of maximizing predictive performance as opposed to understanding model parameters (Taddy 2019); however, using a tool built for forecasting and assuming that it also possesses the properties required for causal inference in economic applications can be misleading (Mullainathan and Spiess 2017). Maximizing the predictive power of a model to use it for interventional analysis represents a major trap. Indeed, it may even be necessary to sacrifice predictive accuracy to arrive at a correct understanding of the relationships that are relevant for making decisions about interventions (Athey 2018). The current lack of understanding of cause–effect connections is even cited as a fundamental obstacle for machine learning by some authors (Pearl 2019). Nevertheless, many inference procedures include prediction tasks as an important step (Mullainathan and Spiess 2017). Machine learning is especially suited for this step in high-dimensional settings (Belloni et al. 2014b). The double machine learning framework (Chernozhukov et al. 2017), which we will apply in Sect. 5, allows us to take advantage of the predictive performance of machine learning algorithms when seeking solutions for causal problems.

4 Literature review

We conducted a search of the literature across Google, Google Scholar and finance journals on the use of machine learning in FP&A. The use of quantitative methods in the broad field of finance has been studied intensively for close to 40 years (Ozbayoglu et al. 2020), in part because of the general availability of data in this field, the existence of many areas of implementation and the substantial economic impact of financial decisions. Our search yielded surprisingly few recent publications on the use of machine learning explicitly in FP&A and related fields. The key thrust of machine learning in finance is directed towards various applications ultimately linked to forecasting and trading financial instruments such as stocks, bonds,

currencies and derivatives. Credit scoring and fraud detection are other major areas. Examples of recent surveys include Ozbayoglu et al. (2020) and Henrique et al. (2019).

We see two possible reasons for the apparent scarcity of publications on machine learning in FP&A. First, time-series forecasting has been thoroughly covered and researched for many years (De Gooijer et al. 2006). A large variety of tools for this purpose have been developed, both from an academic and theoretical perspective, as well as from the perspective of practitioners, including easy to use off-the-shelf software (Küsters et al. 2006). From a practical FP&A perspective, these tools, together with the domain knowledge of the experts working in the FP&A function, allow practitioners to arrive at results that—by and large—serve sufficiently well to meet the objective of developing financial plans. Especially, practitioners may therefore perceive machine learning as a "so-so" technology (Acemoglu and Restrepo 2018), which is not (yet) quite worth their (full) attention. Thus, the intrinsic urge to look for new tools, including machine learning, in FP&A is still less pronounced than it is, for instance, in stock-market forecasting, where even a relatively small improvement in forecasting accuracy can yield significant economic payoff. We believe that this will change with the further deployment of digitalization and the consequent increase in data availability as described above. Besides improving the precision of financial forecasts, automated forecasts driven by machine learning can also lead to a substantial reduction in costs and to increased flexibility given that the traditional process is quite labor- and time-intensive.

Second, we hypothesize the following reason for the limited number of publications on machine learning in FP&A. The initial development of artificial intelligence and machine learning methods was driven mostly by academia. Because these methods are highly relevant for industrial applications, companies (in particular in the tech field) have shown strong interest in applying and developing them further. Indeed, some of the large tech companies host their own dedicated research teams. However, the limited availability of skilled professionals represents a hurdle to fast diffusion in all corporate functions of a company. Therefore, the application of machine learning for FP&A is still rare in the finance function, even though a host of machine learning publications by the industry has already appeared in other functional areas.[2] Management consultancies have also discovered the benefit of machine learning for finance and FP&A. However, their publications remain general and directional in nature (see, for instance, Balakrishnan et al. 2020; Roos et al. 2020; Tucker et al. 2017; Chandra et al. 2018).

One company that has made public its use of machine learning in FP&A in scientific papers is Microsoft Corporation. In the past several years, Microsoft appears to have followed an innovative approach with machine learning in FP&A as witnessed by three publications from its employees. One paper (Gajewar et al. 2016) compares the performance of random forests to that of traditional time-series methods such as autoregressive integrated moving average (ARIMA), error trend and seasonality

[2] For instance, www.bosch.com/de/forschung/know-how/publikationen (accessed Feb 23, 2021) contains a collection by Robert Bosch GmbH.

(ETS, a variant of exponential smoothing) and seasonal-trend decomposition using loess (STL, another variant of smoothing) for forecasting quarterly revenues by major geographic region and at the global level up to 1 year into the future. Based on their exploratory analysis, the random forest model with a restricted number of features outperformed the traditional time-series methods and the forecasts generated by the domain experts in the Microsoft FP&A department.

A second paper (Barker et al. 2018) describes a machine learning-based solution that forecasts revenue on a quarterly basis, including individual forecasts for 30 products in three different business segments. Specifically, the machine learning forecast used an elastic net, a random forest, a K-nearest-neighbor and a support vector machine. The winner model was then selected via back-testing. The forecasts generated in this way proved to be more accurate than the traditional forecasts generated by FP&A in approximately 70% of the cases. The paper cites the ability to incorporate external information (e.g., temperature as a driver for electricity demand) in regression frameworks as an advantage of these over pure (standard) time-series models. While classical time-series are good at capturing trends and seasonality, they often struggle to incorporate external data. In particular, they generally lack a regularization mechanism, leading to low out-of-sample accuracy for new forecasts, especially in high-dimensional settings. Many machine learning methods include by design mechanisms to avoid overfitting (e.g., regularization for ridge, lasso and elastic net).

Barker et al. (2018) also highlight some requirements that arise from the intent to use the results of machine learning forecasts in a practical manner in a corporate setting. Traditionally, FP&A works with a point estimate, coupled with an estimation of the risks and opportunities around this mid-point. Risks and opportunities typically consist of a list of items or events that will materially impact the business results if they do not turn out as assumed in the mid-point forecast (Conine and McDonald 2017). Judgmental probability estimates provided by subject matter experts are often attached to these items, together with a quantification of the expected impact under the different scenarios.[3] For forecasts generated by traditional statistical or machine learning models, prediction intervals are therefore an important element for FP&A practitioners to quantify the risk in the forecast. However, prediction intervals are not typically part of machine learning models. The solution proposed by Barker et al. (2018) consists of creating intervals from out-of-sample error distributions obtained during back-testing. Other practical requirements in a corporate environment are the need for a mostly automated solution allowing for fast forecast generation as well as the need to ensure high security standards for data storage, processing and access. Financial data such as sales and profits are highly sensitive, and companies are reluctant to release them into public cloud environments. Barker et al. (2018) explain the details of their workflow automation and security controls, which revolve around the Microsoft Azure cloud-computing platform.

[3] Other methods with a very similar intent exist. Examples are the quantification of a best and worst case in addition to the normal or base case, or sensitivity analysis with varying degrees of sophistication.

The third publication (Koenecke and Gajewar 2020) evaluated deep neural networks traditionally used in natural language processing (encoder–decoder LSTMs) and computer vision (dilated convolutional neural networks) to forecast company revenues. The approach incorporated transfer and curriculum learning. For the products and time period under study in this publication, deep neural networks improved predictive accuracy compared to the company's internal baseline, which combined traditional statistical and machine learning methods other than deep neural networks.

In another example of applied machine learning in the area of FP&A, Daimler Mobility used an undisclosed library of machine learning algorithms to generate a monthly forecast set, spanning the next 18 months and updated monthly (Unger and Rodt 2019). In this respect, the approach followed the concept of a rolling forecast. The forecasted set of values comprised key financial performance indicators that were representative of Daimler Mobility's car rental, leasing, financing and fleet management business. According to Unger and Rodt (2019), one of the key advantages of this approach compared to the traditional way of forecasting and budgeting is the speed with which updated forecasts are available, allowing faster adoption of corrective action.

These papers all discuss modern machine learning methods for financial forecasting. In the next section, we will show that these approaches cannot be applied directly to inference problems and how the double machine learning framework overcomes this problem. A first example will illustrate the use of machine learning techniques in FP&A for forecasting. A second example will serve to illustrate the use of double machine learning for planning (inference). Finally, we will explore whether having additional data improves the results for both the forecasting and planning tasks.

5 Simulation example

In this section, we provide the results of a small simulation study. The design of the simulation reflects the setting, types of data and questions that the FP&A department in a large, multinational company could face. We will start with an example in which FP&A is predominantly interested in the accuracy of sales forecasting. We will then carry this example forward into a question related to planning. In this second example, FP&A is interested in assessing the effectiveness of promotional activities in generating sales; in other words, the question of interest relates to causal inference and the answer to this question can inform decisions about resource planning. Finally, we will investigate how the results change if the FP&A department obtains additional data points for their tasks.

5.1 Forecasting

Assume for our stylized simulation the following setting:[4] for a given month n, the FP&A department would like to forecast the sales y_n of a specific product or service. FP&A has collected monthly data over 5 years ($N = 60$) for sales as well as a set of $P = 40$ factors or features that FP&A believes could be predictive of these sales. We represent these factors as $x_{p,n}$ and the corresponding sales with y_n.[5] In practice, there can be a wide range of factors depending on the product or service. Examples include weather conditions and various macroeconomic indicators, but also specific customer shipment patterns or the current competitive market situation. Note that the size of the feature set can easily reach 40 plausible predictors once an initial, smaller feature set is increased due to the inclusion of transformed and newly created features. This step, called feature engineering, can include the creation of lagged variables (e.g., when the effect of the economic situation affects sales several months later) or interaction effects (e.g., when a particular weather situation coincides with a peak shipment date, nullifying or exacerbating the effect of the peak shipment date). A further example is the transformation of categorical variables into several binary values via the so-called one-hot encoding (e.g., when classifying the competitive market situation as "highly competitive", "moderately competitive", "not competitive" and the like).

In addition to developing a set of 40 features, FP&A measures the promotional activity carried out by the company for the product under investigation during the reference timeframe. We denote this promotional activity as d_n. For the purpose of this illustrative simulation, we work with the assumption that the promotional activity can be measured using a single variable. In other words, we do not enter into promotional mix considerations with interaction effects among the different promotional tools. In practice, this single variable could be a summary measure such as the amount of money spent on promotion and advertising; another possibility for a summary measure could be the number of customer calls or minutes of customer interaction. For forecasting and planning activities performed by FP&A at aggregate levels, such as the regional, divisional or group levels, an approach like this, based on a summary measure, is sometimes applied. Extending the analysis to include several marketing variables is possible without any major changes.

Given the nature and intent of promotional activity, it appears natural for FP&A to include d_n in the list of likely predictors for the sales forecasting model. Furthermore, estimating the effect of the promotional activity on sales represents an important question for FP&A, which we will address in the second part of this section, dedicated to planning.

[4] We have intentionally kept the simulation example simple. For instance, we have not added any time-series-specific effects such as a trend component or serially correlated error terms. This allows us to focus on the key elements. For instance, the $N = 60$ data points could represent observations in different countries or sub-markets, which would warrant a cross-sectional approach to the analysis. The conclusions presented in the simulation example will remain largely unchanged.

[5] This is a case of supervised learning, because we have observations for both the input ($x_{p,n}$) and the output (y_n).

To evaluate the accuracy of the sales forecasts, we will follow an out-of-sample evaluation approach. Only the first four years (48 data points) are used to build and train the forecasting models. FP&A then compares the forecasts generated by the models to the actual values from the last year in the available dataset (12 data points). Note that these 12 data points have been intentionally excluded from the model creation phase. While more sophisticated training and evaluation strategies exist (e.g., rolling evaluation windows), the described approach is sufficient for the purpose of this simulation study, because the out-of-sample forecasting performance is evaluated separately for each simulation.

For our simulation, we generate the data as $n = 1, \ldots, N$ independent and identically distributed (i.i.d.) draws from the following model:

$$y_n = \alpha d_n + x_n' \beta_p + \varepsilon_n, \tag{5.1}$$

and

$$d_n = x_n' \gamma_p + \nu_n, \tag{5.2}$$

with $x \sim \mathcal{N}(0, \Sigma)$ where Σ is a $p \times p$ matrix with $\Sigma_{k,j} = c^{|j-k|}$, $\varepsilon \sim \mathcal{N}(0, 2)$ and $\nu \sim \mathcal{N}(0, 2)$.

The second equation captures confounding, i.e., variables that are simultaneously correlated with the outcome variable and the variable of interest. By setting $\alpha = 0$, we assume that the promotional activity undertaken by the company has no effect on sales, i.e., that the promotion efforts are, in reality, a waste of resources. With $c = 0.3$, we include some moderate correlation[6] between features, which can be expected if several features from the same general background (e.g., macroeconomic factors) are included in the model.

We set $\beta = 0$, except for β_{39} and β_{40}, both of which we set equal to 1. Thus, out of the 40 features included in the analysis, only two are actually related to sales. Similarly, we set $\gamma = 0$, except for γ_{39} and γ_{40}, both of which we also set equal to 1. The two features related to sales also determine the amount of promotional activity d_n.[7] ε and ν are random error terms (so-called noise). We report results based on 1000 simulation replications.

It is important to remind ourselves that the FP&A department naturally does not know any details about this data generation process. Only an oracle would know that, in reality, solely 2 of the 40 plausible predictors are linked to sales and that the coefficient in the data-generating process is 0 for the other 38 features. This situation characterizes sparse models. In such models, only a small number of many potential

[6] The value of c represents the correlation between immediate neighbor features (e.g., feature x_p and feature x_{p+1}). Due to the way Σ is constructed, the correlation decays quickly as the distance between features increases (e.g., feature x_p and x_{p+3} have only a correlation of c^3, which is 0.027 for $c = 0.3$).

[7] A simple example can help clarify the intuition behind this setting. Ice cream sales on the beach are probably positively related to weather conditions (feature 1) and the day of the week (feature 2). At the same time, the ice cream salesperson may decide to run some promotional activity when weather conditions are favorable on a weekend. Thus, the same features have an influence both on sales and on promotional activity. We will revert to this illustrative example in the section on planning.

predictors and/or control variables are actually relevant (Belloni et al. 2014a). Identifying them leads to a correct model specification and is the main challenge.[8] Additionally, FP&A does not know that the promotional activity d_n is correlated with the two features that have non-zero coefficients with respect to sales and that the promotional activity has no influence on sales (α, the parameter of interest, is zero). We will come back to this point when we discuss inference.

We now provide results for two forecasting approaches. Both have in common that they rely—in this case correctly—on the typical assumption of a linear relationship between the output variable Y (sales) and the full set of regressors X, which includes 40 presumably predictive features and one variable reflecting promotional activity

$$Y = X'\beta + \varepsilon. \tag{5.3}$$

The first approach is a traditional linear regression based on the ordinary least squares (OLS) method. Formally, OLS optimizes the parameters in such a way as to minimize the mean squared error (MSE)

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\hat{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \{y_i - x_i'\hat{\beta}\}^2, \tag{5.4}$$

where $x_i'\hat{\beta}$ corresponds to the predicted sales value.

The second approach, post-lasso, is a classic machine learning technique. To estimate the coefficients, lasso uses a regularization strategy that is suited to high-dimensional problems in which the number of predictors exceeds or approaches the number of observations, as is the case in our simulation. In the first step, the lasso regression is performed. In the second (i.e., post-lasso) step, the method fits OLS on the coefficients selected in the first step. Formally, lasso optimizes the parameters to minimize MSE subject to a penalty for using parameters

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\hat{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \{y_i - x_i'\hat{\beta}\}^2 + \lambda|\beta|, \tag{5.5}$$

where $x_i'\hat{\beta}$ corresponds again to the predicted sales value.

The key difference between the lasso and OLS is that lasso minimizes a penalized MSE, in which the penalty amount corresponds to the absolute amount of each parameter included in the model, scaled by the tuning- or hyperparameter λ

$$\text{Penalty}_{\text{Lasso}} = \lambda|\beta|. \tag{5.6}$$

[8] By construction, our simulation example is exactly sparse, with parameter values for all non-relevant features exactly equal to zero. For practical applications, a more realistic assumption is approximate sparsity, meaning that all or many features can have non-zero parameter values. Nevertheless, only a limited number of features are needed to approximate the true relationship with sufficient accuracy. We refer interested readers to Belloni et al. (2010). Our simulation could easily be extended to such a setting. Results would remain largely unchanged.

Table 1 Average RMSE of the hold-out-sample from 1000 simulation runs for the forecasting task

RMSE	In sample	Out of sample
OLS	0.738	5.321
Post-lasso	1.991	2.162

A detailed discussion of the theory behind regularization approaches would go beyond the scope of this article. Readers are referred, among many possible sources, to Hastie et al. (2009), Bühlmann and van de Geer (2011) and Taddy (2019). Taddy (2019) sees regularization as "the key to modern statistics" by virtue of its ability to prevent overfitting in high-dimensional settings. Instead, we will recall a few characteristics of the lasso that are particularly relevant to our FP&A example and the corresponding simulation.

The full name of the lasso ("least absolute shrinkage and selection operator") indicates two important characteristics. First, as we can see in the formula for $\text{Penalty}_{\text{Lasso}}$, the absolute size of the coefficients included in the model represents a cost in the minimization of the MSE. Lasso will therefore shrink the coefficients towards zero. This makes the prediction system more stable and avoids overfitting. Second, the lasso-specific penalty in the form of the absolute value of the coefficients has the property that some parameters will be exactly equal to zero. In other words, the lasso will fully exclude some variables from the model and therefore perform automatic variable selection.

As indicated above, the lasso can handle situations in which the number of predictors approaches or even exceeds the number of observations. In our case, the number of predictors (including the measure of promotional activity) is 41 and the number of observations is 48. Although OLS can still be calculated, we will see that its out-of-sample predictive accuracy becomes extremely unreliable. If we were to chose a simulation scenario with 48 or more predictors, OLS could no longer be computed. A second challenge for OLS in settings with many predictors is the increased risk of correlation among the predictors. If predictors are highly correlated among themselves, or if, in an extreme case, there is an exact linear relationship between two predictors (multicollinearity), OLS estimates become unstable. For instance, macroeconomic variables tend to be strongly correlated.

An important ingredient in the lasso is the size of the penalty, which depends on the tuning parameter λ. λ is not determined by the lasso itself, but needs to be selected. Intuitively, λ plays a role in filtering the relevant variables. Several strategies to select λ have been proposed in the literature and are used by practitioners. The most common are cross-validation strategies and information criteria such as Akaike's or Bayes' information criterion. Our simulation study uses the data-dependent penalty level proposed by Belloni and Chernozhukov (2013). We refer interested readers to this source for details.

Compared to the standard lasso approach, which induces bias due to the shrinkage of coefficients, post-lasso has the advantage of a smaller bias, even if the model selected in the first step by lasso fails to include some of the true

predictors. It also converges at a faster rate towards the true parameter values if the model selected by lasso correctly includes all true predictors (in addition to some irrelevant predictors). If lasso selects exactly (only) the true predictors, the post-lasso coefficient estimators are equal to the ones produced by an oracle that is aware of the underlying data-generating process (Belloni et al. 2012, 2014a).

Table 1 summarizes the results of 1000 simulation runs for the forecasting task, comparing OLS to post-lasso.

We report the forecast accuracy in terms of average root-mean-squared error (RMSE) over all simulation runs both on the in-sample and the out-of-sample data set. As outlined above, the in-sample data set consists of 48 data points, which are used to build and train the models. The out-of-sample data set consists of 12 data points, which are intentionally not used in the model construction ("hold-out sample"), allowing the model to be evaluated on new, previously unseen data. The strong focus on forecasting performance on previously unseen data is a hallmark of the machine learning approach.

On the in-sample data, OLS produces a higher predictive accuracy than post-lasso, with an RMSE of 0.738 which is nearly one-third that of the post-lasso RMSE of 1.991. However, the real interest of the FP&A department here is not to model past sales data. Rather, the predictive performance on new data is what matters to FP&A; this is why, the out-of-sample data have been set aside. Here, the OLS RMSE increases substantially to 5.321, more than twice as high as the post-lasso RMSE of 2.162.

We can draw two main conclusions from the simulation. First, the RMSE of standard OLS increases significantly between in-sample and out-of-sample data. With nearly as many features (regressors) as observations in the model, the resulting overfitting is immediately exposed when OLS is evaluated using previously unseen data. Second, the post-lasso RMSE is relatively stable between the in-sample and out-of-sample data. The in-sample performance is thus already indicative of the true predictive power when post-lasso is used on unseen data. Lasso achieves this through the regularization strategy described above, which leads to a very selective inclusion of features and thus parsimonious models. For reference, of the 40 available features in the simulation, post-lasso retains an average of only 1.2 as relevant and shrinks the coefficients of all the others to exactly zero. As a reminder, our simulation includes only two truly relevant features. The out-of-sample RMSE for post-lasso is thus slightly hihger than the perfect RMSE score of 2.0 (equal to the standard error that was selected for the noise parameter ϵ), which would be achieved by an oracle.

Figure 1 shows the distribution of the out-of-sample RMSE for the post-lasso forecast over the 1000 simulation runs. The distribution of the errors follows approximately a normal distribution (overlaid as a red line). From a practical perspective, the risk of generating a highly incorrect lasso forecast is therefore limited. Furthermore, the right tail of the lasso errors ends before the mean of the OLS error. This provides additional reassurance when relying on lasso.

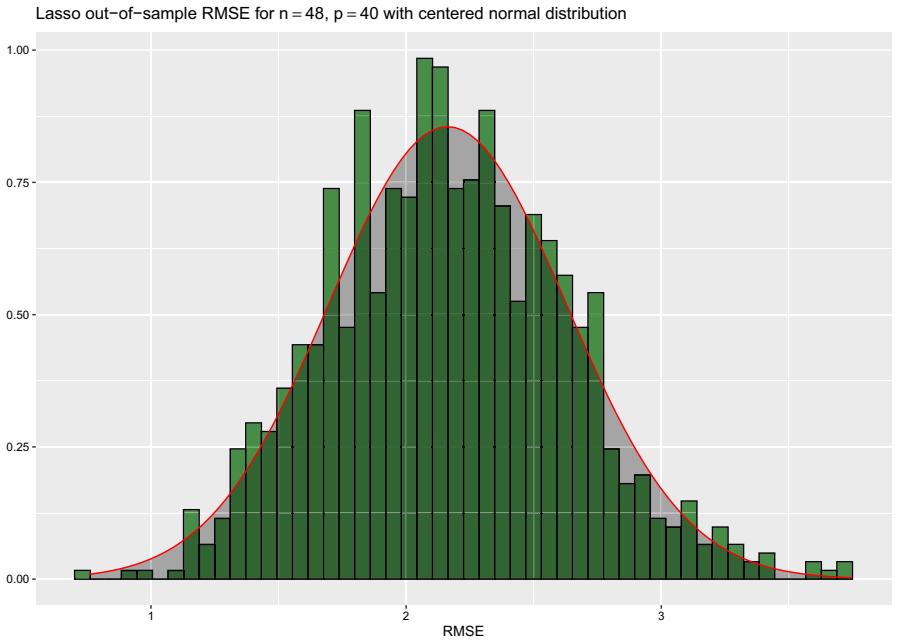

Fig. 1 Distribution of the out-of-sample RMSE for the post-lasso forecast (bars), compared to the normal distribution (red line) (colour figure online)

5.2 Planning

We will now discuss the use of machine learning in financial planning. To come back to our example, the task for the FP&A department consists of evaluating the effectiveness of promotional activity in generating sales; in statistical parlance, the task relates to statistical inference of the effect of a treatment or intervention (i.e., the promotional activity) on an outcome (i.e., sales). This estimate forms the basis for planning and optimizing marketing activities. In our simulation examples, evaluating the effectiveness of promotion equates to estimating the parameter α. As the parameter of interest, α corresponds to the effect of the promotional activity on sales, also called the "lift" in business applications. Let us remind ourselves that in our simulation, only two features are relevant for the sales forecast and that these two features also determine the amount of promotional activity. Thus, we are dealing with confounders, because these two features are correlated with both the treatment and the outcome. Moreover, we have set α to zero, which effectively means that the promotional activity does not have an impact on sales.

In a business environment, this setting could correspond to an ice cream vendor at the beach who spends money on promotion whenever the weather is warm and sunny on the weekends. He ascribes the increased ice cream sales, or at least a part of them, to his promotional efforts, whereas in reality, it is the favorable weather on the weekend that makes people come to the beach and enjoy his ice cream. Similar to the forecasting exercise, the FP&A department is obviously not aware of the

data-generating process governing the simulation and needs to find a way to estimate α.

One approach to estimating the effect of promotion could be to use the parameter estimate for α from the lasso model employed in sales forecasting. However, lasso shrinks parameter estimates because of the penalty loading used in the regularization process and therefore does not generate unbiased estimates of the parameter values, even though it allocates the least possible penalty amount to large signals while retaining the stable behavior of a convex penalty (Taddy 2019). Additionally, lasso estimates predictors sparingly insofar as it sets many parameter estimates to exactly zero. In many cases, the factor measuring promotional activity "may not make it" into the second step of the post-lasso procedure. It is therefore not meaningful to infer from the forecasting model the effectiveness of the promotional activities. We have previously highlighted the warning by Mullainathan and Spiess (2017) and Athey (2018) that using a tool built for forecasting and assuming that its parameters possess the properties required for inference can be misleading.

With the above in mind, one could decide to pursue a hybrid solution with the following approach. Because the lasso has identified the most relevant features for prediction, we carry these forward into the inference model. Additionally, we include in the model the variable of interest (the intervention), which in our example is the variable that represents promotional activity. In a sense, we force this variable of interest into the model. We then estimate the parameter values for all of these features using OLS, which allows us to perform inference on the parameter estimates. In particular, we are able to interpret our parameter of interest α in this model. In our example, α will tell us the effectiveness of the promotional activities. Intuitively, a model that is constructed in this way can be understood as attempting to estimate the effect of promotional activity, while controlling for other factors with proven high predictive power from the forecasting model. For the ice cream vendor at the beach, this corresponds to controlling for the effect of the favorable weather during the weekend and thus deriving an isolated estimate of the effect of promotional activity on sales. This approach can be represented as

$$y_n = \alpha d_n + x'_{p*,n}\beta_{p*} + \varepsilon_n, \tag{5.7}$$

with $p*$ corresponding to the subset of all p features for which $\hat{\beta}_{\text{Lasso}}$ is non-zero.

We will see from the simulation results that this approach, which we will call "naive", grossly fails to discover the true value of the parameter of interest, when modern machine learning methods are used in high-dimensional settings; still, it is widely used by practitioners and applied researchers. In our model, the promotional activity measure is correlated with the features that concomitantly and directly influence sales. In the presence of such confounders, the naive approach will fail.

In short, the naive approach will suffer from omitted variable bias. This is because machine learning methods capture the features correlated with the outcome variable and deliver good predictive performance but often miss variables that are correlated weakly with the outcome but correlated more strongly with the intervention variable. Missing these variables does not harm predictive performance but biases the estimation of the intervention effect, leading to invalid post-selection inference.

For an approach to be valid, it must overcome this problem of imperfect model selection and related omitted variable bias. Double or debiased machine learning, as proposed by Chernozhukov et al. (2017), is one way to do so. The fundamental idea[9] is to reduce, for the estimation of the parameter of interest (i.e., the intervention variable), the sensitivity with respect to errors in selecting and estimating the nuisance parameters (i.e., the other predictors in the model). Technically, this can be achieved by regressing residuals on residuals. The first set of residuals is generated by regressing the outcome variable on the control features, notably using regularizing machine learning methods such as (post-)lasso, random forests, boosted trees or other methods suited for high-dimensional settings. The second set of residuals is generated by regressing the treatment variable on the control features, again using modern machine learning methods. This auxiliary step helps to control for the confounders that might lead to omitted variable bias. Finally, the first set of residuals is regressed on the second set of residuals. The parameter value obtained in this residuals-on-residuals regression represents the effect of the treatment variable on the outcome. This procedure is known as Frisch–Waugh–Lovell partialling out. In our simulation study, machine learning methods are used for partialling out. This approach allows for valid inference compared to the naive approach.

Translated into our stylized simulation, the first regression relates sales to the 40 presumably predictive features; the differences between the predictions $\hat{y}_n$ from this first regression and the actual outcomes (sales) y constitute the first set of residuals r_n^1

$$\hat{y}_n = x_n' \hat{\beta}_p, \tag{5.8}$$

$$r_n^1 = y_n - \hat{y}_n. \tag{5.9}$$

The second regression relates the promotional activity score d to the 40 presumably predictive features; the differences between the predictions from this second regression $\hat{d}_n$ and the actual outcomes (promotional activity) d_n constitute the second set of residuals r_n^2

$$\hat{d}_n = x_n' \hat{\gamma}_p, \tag{5.10}$$

$$r_n^2 = d_n - \hat{d}_n. \tag{5.11}$$

Concretely, we will use a post-lasso approach in the regressions to derive both sets of residuals, but in principle, any other machine learning method could be used, such as random forests or support vector machines. Finally, we regress the residuals from the first regression onto the residuals from the second regression to obtain an estimate for the parameter of interest, α, which represents the impact of promotional activities on sales

[9] Double machine learning also uses cross-fitting, an efficient way of data splitting. Interested readers are referred to Chernozhukov et al. (2017).

Table 2 Results from 1000 simulation runs for the planning/inference task

α	Naive	Partialling-out
Mean estimate	0.1604	0.0081
Std. dev.	0.1668	0.1326
t-statistic	30.422	1.934
p value	0.0000	0.0534
Rejection rate	46.1%	4.8%

The *t*-statistic and *p* value refer to the respective mean estimate

$$r_n^1 = \alpha r_n^2 + \varepsilon_n. \tag{5.12}$$

This approach works well in practice, because the residuals-on-residuals approach makes the estimation of the treatment effect less sensitive to errors in the model specification. Athey (2018) provides an intuitive explanation: "[...] in high dimensions, mistakes in estimating nuisance parameters are likely, but working with residualized variables makes the estimation of the average treatment effect orthogonal to errors in estimating nuisance parameters." This is why the family of approaches that use this principle is also referred to as orthogonal machine learning (Taddy 2019). Interested readers are referred to the literature for an in-depth theoretical discussion, including underlying assumptions and formal proofs, which is beyond the scope of this paper. Key sources include Belloni et al. (2014a) and Chernozhukov et al. (2015, 2018). To implement our simulation, we use the partialling-out approach as defined by Chernozhukov et al. (2016) and report the corresponding results under this label.

Table 2 summarizes the results of the two approaches (i.e., "naive" and "partialling out") from 1000 simulation runs. We report the mean estimate for α, the standard deviation of the estimate and the corresponding *t*-statistic and *p* value for a two-sided test of whether the mean is different from zero. The rejection rate represents the proportion of individual simulation runs in which the ingoing assumption of α=0 has been rejected based on the *t* test (at the customary 5% significance level). In other words, these are the instances in which the model incorrectly suggests an effect (positive or negative) of promotional activity on sales.

The simulation results provide several insights. First, and this is the main point we seek to make, the naive approach grossly fails to discover the true value of α, because it suffers from significant bias. Put simply in the context of our simulation, this bias represents systematic over-estimation of α and thus over-estimation of the effectiveness of promotion. On average, the naive approach estimates a value for α of 0.1604, compared to a true value of zero. The partialling-out approach also yields an average positive value for α of 0.0081, but is much closer to the true value of zero. Relatively speaking, the bias of the naive approach is roughly 20 times higher than that of the partialling-out approach.

A second point is that the standard deviation of the estimates for α are similar for both approaches. Figures 2 (naive approach) and 3 (partialling-out approach) show the distribution of the estimates for α from the 1000 simulation runs compared to a

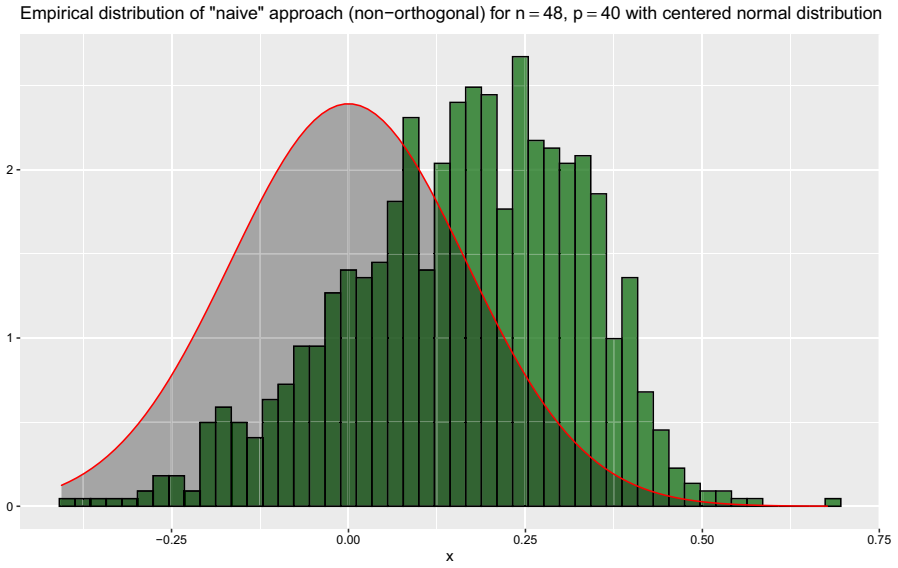

Fig. 2 Distribution of estimator for α from the naive approach

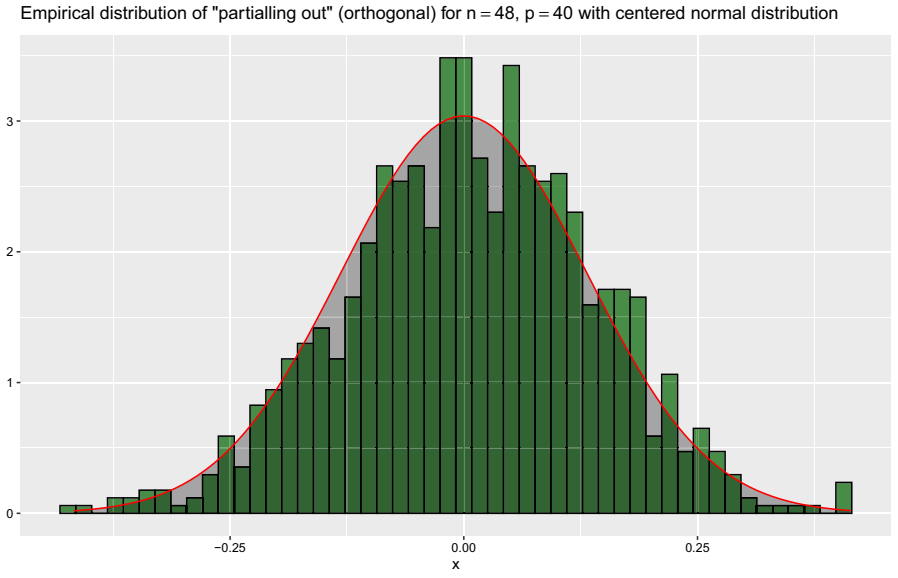

Fig. 3 Distribution of estimator for α from the partialling-out approach

normal distribution curve. Visual inspection suggests that the shapes of both distributions are well approximated by a normal distribution. Of course, the center of the distribution for the naive approach is clearly shifted to the right of zero. This reinforces the point made above that bias is induced by the naive approach.

Table 2 also reports the t-statistic and corresponding p value for a two-sided test of whether the mean estimate of α is zero. Under the naive approach, this hypothesis would be rejected with high confidence (t-statistic of 30), reinforcing the incorrect belief that the promotional efforts positively affect sales. Under the partialling-out approach, the hypothesis of no effect from promotional efforts would not be rejected at the customary 5% threshold level (t-statistic of 1.93). In practice, the FP&A department would of course not benefit from this kind of insight as they would not have access to repeated estimates for α. With the advantage of being able to run multiple simulations, we can use this information to support the point of significant bias in the naive approach. Nevertheless, the rejection rate, reported in the last line of Table 2, provides a good indication of how often FP&A would make an incorrect decision. For each individual run in the simulation, this metric records whether FP&A would (incorrectly) reject the assumption that α is zero at the typical 5% significance level. Under the naive approach, this would happen 46% of the time. Put differently, a bit less than half of the time, FP&A would incorrectly assume that promotional activity does have an effect on sales. With partialling-out, this error drops to slightly below 5%.[10]

In summary, by relying on the naive approach, the FP&A department (or the ice cream vendor) would substantially overestimate the causal effect of the promotional activity on sales. Consequently, this activity would probably be maintained or even increased for this product or service, even though in reality, it does not increase sales. Put differently, the company would draw up plans that allocate resources wastefully on this particular product or market. The impact from falling into this trap could multiply even further across the organization if the results of such an analysis were used as a benchmark for similar products, services or geographic markets. This might happen, for example, if data are not readily available for a particular product (for example, one that is being newly launched) and the decision is made to extrapolate from existing (and potentially wrong) information. Such a situation is even more likely when the existing information appears plausible and suitable[11] and, in addition, is perceived as objective, unbiased (in the sense of free from human/cognitive bias) or even scientific, because it was generated using data-driven methods.

5.3 The value of data

In 2017, "The Economist" (Economist, 2017) asserted in the title of its May 6 edition that data are now the world's most valuable resource. Questions about the value of data as a resource and production factor have generated great interest in academia and policy institutes. One consideration within this vast topic is a (hypothesized) positive feedback loop: more data lead to more data-driven insights, allowing a company to serve its customers better, to attract more customers and, in turn, to collect

[10] Note that one would expect an error rate here of around 5% from a correct model, because the 5% significance level corresponds to a 5% probability of rejecting the null hypothesis when it is, in reality, true.

[11] Blake et al. (2015) highlight in their paper that "[...] the incentives faced by advertising firms, publishers, analytics consulting firms, and even marketing executives within companies, are all aligned with increasing advertising budgets."

Table 3 Average RMSE from 1000 simulation runs for the forecasting task, 48 vs. 60, 72 and 96 training observations

RMSE	48 training obs.		60 training obs.		72 training obs.		96 training obs.	
	In-s.	Out-s.	In-s.	Out-s.	In-s.	Out-s	In-s.	Out-s.
OLS	0.738	5.321	1.101	3.491	1.309	2.983	1.507	2.561
Post-Lasso	1.991	2.162	1.978	2.107	1.988	2.085	1.975	2.015

"In-s." and "Out-s." refer to in-sample and out-of-sample, respectively

even more data. Nevertheless, there seems to be a broad consensus that data are generally governed by decreasing returns to scale, like any other production factor (Varian 2018; Bajari et al. 2019).

In this paper, we will limit ourselves to a short discussion of how the number of observations affects the accuracy achieved by the forecasting and inference methods used by the FP&A department within the frame of our simulation. For empirical results, we refer interested readers to Bajari et al. (2019), which contains a study of the performance of Amazon's retail forecasting system. The study finds performance gains in the time dimension (i.e., from longer data history), but not in the product dimension (i.e., panel data forecasts do not improve with more products within a category). An interesting finding is the overall improvement of forecasts over time (controlling for the length of data history and the number of products), suggesting positive effects from improved technology (e.g., new machine learning models, better hardware or adaptation of organizational practices).

In our simulation, the hypothetical FP&A department uses a training set of 48 observations, 40 predictive features and one variable of interest for inference (i.e., the measure of promotional activity). In many real-life applications relevant to FP&A departments, the number of observations available for analysis is typically limited. More observations may simply not exist; for instance, new products generate sales data starting only from their launch date. Even if data do exist, collecting, accessing and, if necessary, curating them come at a cost; for instance, companies may limit the amount of directly accessible data history due to system constraints, or data generated prior to the introduction of new software may be inaccessible, in full or in part.

Let us now explore simulation results assuming that the FP&A department has invested in expanding the training set of observations to 60, 72 or 96. The number of features (i.e., 40), the variable of interest (promotional activity measure) and the overall simulation set-up remain unchanged.[12] We again run 1000 simulations. What is the return on accuracy of expanding the observation set?[13]

[12] We intentionally do not allow the number of features to grow with the sample size (see for instance Belloni et al. 2010) to isolate the effect of the additional observations clearly. In practice, a significant extension of the number of observations may require including additional control features.

[13] The natural way to think about the expansion is to assume that the department "digs out" additional historical observations. However, from a theoretical standpoint, the department could also wait 1, 2 or 4 years, respectively, and gather the additional data points over time. In this case, the change in forecasting accuracy could be mistaken for a technology or learning effect by an outside observer.

Table 4 Results from 1000 simulation runs for the planning/inference task, 48 vs. 60, 72 and 96 training observations

α	48 training obs.		60 training obs.		72 training obs.		96 training obs.	
	Naive	Part.-out	Naive	Part.-out	Naive	Part.-out	Naive	Part.-out
Mean estimate	0.1604	0.0081	0.1263	0.0055	0.0956	− 0.0008	0.0617	0.0042
Std. dev.	0.1668	0.1326	0.1708	0.1255	0.1694	0.1166	0.1471	0.0952
t-statistic	30.422	1.934	23.385	1.381	17.847	− 0.217	13.257	1.400
p value	0.0000	0.0534	0.0000	0.1675	0.0000	0.8279	0.0000	0.1619
Rejection rate	46.1%	4.8%	45.9%	5.9%	38.6%	6.4%	28.9%	4.8%

The *t*-statistics and *p* values refer to the respective mean estimates. "Part.-out" refers to partialling-out

Table 3 reports the forecasting results based on 60, 72 and 96 observations compared to the previous simulation based on 48 observations. For OLS, the in-sample accuracy drops, as witnessed by the increase in RMSE to 1.507 (for 96 observations) from the initial RMSE of 0.738 with 48 observations. However, the out-of-sample accuracy increases: the corresponding RMSE drops to 2.561 (for 96 observations) from the previous RMSE of 5.321 based on 48 observations. In fact, the additional observations reduce the extent of overfitting seen in the initial setting. With 40 features and (only) 48 observations, OLS was actually close to the point of failing. This point would have been reached if the number of features had been equal to or exceeded the number of observations. Intuitively, OLS moves further away from this point by expanding the set of observations (and keeping the number of features constant).

For post-lasso, the results based on 60, 72 and 96 observations are quite similar to those obtained with 48 observations. Neither the in-sample nor the out-of-sample RMSE change notably. As expected and unlike OLS, post-lasso already deals well with the initial situation in which the number of features is close to the number of observations and benefits only marginally from the increase in observations. Put differently, post-lasso does not require investing in the generation or acquisition of additional data. Our finding is consistent with standard stochastic theory.[14]

In summary, while having more data is generally beneficial, expanding the observation set for forecasting in our simulation study creates a tangible advantage only for OLS. If the FP&A department employs post-lasso, which is the preferable method in this setting, the gain in precision from expanding the observation set is very small and, for many practical applications, would not warrant the effort.

We will now look at inference, which entails estimating the (causal) effect of promotional activities on sales. Table 4 reports the inference results for α based on 60, 72 and 96 training observations compared to the previous simulation based on 48

[14] See, for instance, Bühlmann and van de Geer (2011) or Belloni and Chernozhukov (2013). Post-lasso converges towards the true parameter value at a rate of $n^{-1/4}$, which is slower than the OLS rate of $n^{-1/2}$. The value of additional data is thus generally smaller for post-lasso than for OLS.

observations. Recall that the true value of α is zero. For OLS, as the number of observations increases, the mean estimate for α decreases to 0.0617 (for 96 observations) from the previous estimate of 0.1604 with 48 observations. However, based on a standard t test, this value is still significantly different from zero (t-statistic of 13.257). In comparison, for the partialling-out approach, the mean estimate for α declines from 0.0081 with 48 observations to 0.0042 for 96 observations, with a minimum of − 0.0008 in the simulation run based on 72 observations. In all three additional scenarios, it is not statistically different from zero (t-statistic of 1.381, − 0.217 and 1.400, respectively).

The expanded set of observations reduces the bias of the naive approach. Intuitively, the risk of imperfect model selection described above becomes smaller. Still, the naive approach exhibits significant bias compared to the true value of α. For the partialling-out approach, the additional observations lead to a mean estimate for α that comes even closer to the true value. Depending on the required precision of the estimate, the FP&A department could benefit from the additional set of observations in its analysis. Again, our finding is consistent with standard theory on convergence rates (see, for instance, Bühlmann and van de Geer 2011 or Belloni and Chernozhukov 2013). Whereas post-lasso converges for forecasting towards the true parameter value at a relatively slower rate of $n^{-1/4}$, the double machine learning estimator of the treatment effect converges at the faster rate of $n^{-1/2}$ (i.e., the same rate as OLS).

6 Conclusion

Digitalization, especially when it couples large amounts of data with appropriate tools for analysis, represents an important opportunity for the financial planning and analysis function. In this article, we have provided an introductory overview of machine learning in this context. By reviewing several relevant theoretical aspects of machine learning and discussing the results of a simulation study, we have demonstrated how machine learning may prove useful for FP&A practitioners. We have paid special attention to explain the distinction between forecasting and planning tasks, the first of which involves prediction and the latter of which involves causal inference. We see the confusion of these two concepts as a major pitfall that practitioners should strive to avoid. Specific approaches to causal machine learning have begun to gain traction, as awareness has increased that the naive application of machine learning can fail in applications that go beyond prediction. This applies to all modern machine learning methods in a high-dimensional setting.

Our article has several limitations. It was impossible to cover the vast number of machine learning techniques that exist. Depending on the causal question at hand, a range of econometric approaches (e.g., instrumental variables, synthetic controls or regression discontinuity designs) coupled with machine learning methods may be suitable. We intentionally used a simple data generation process in our simulation; additional elements such as trends or seasonal components or a real-life example could complement our simulation study. Despite these limitations, we believe that our article can be a valuable source of insights into the ways in which FP&A can

benefit from machine learning. With it, we hope to contribute to the adoption of machine learning in this area and help practitioners avoid common mistakes.

Acknowledgements We thank the editor and two anonymous referees for their very helpful comments and suggestions.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation, and work. *The economics of artificial intelligence: an agenda* (pp. 197–236). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.001.0001.

Athey, S. (2018). The impact of machine learning on economics. *The economics of artificial intelligence: an agenda* (pp. 507–547). University of Chicago Press.

Bajari, P., Chernozhukov, V., Hortaçsu, A., & Suzuki, J. (2019). The impact of big data on firm performance: an empirical investigation. *AEA Papers and Proceedings, 109,* 33–37. https://doi.org/10.1257/pandp.20191000.

Balakrishnan, T., Chui, M., Hall, B., & Henke, N. (2020). Global survey: the state of AI in 2020. McKinsey & Company. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020. Accessed 6 Dec 2020.

Barker, J., Gajewar, A., Golyaev, K., Bansal, G., & Conners, M. (2018). Secure and automated enterprise revenue forecasting. In AAAI, pp. 7657–7664.

Becker, S. D., Mahlendorf, M. D., Schäffer, U., & Thaten, M. (2016). Budgeting in times of economic crisis. *Contemporary Accounting Research, 33,* 1489–1517. https://doi.org/10.1111/1911-3846.12222.

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica, 80,* 2369–2429. https://doi.org/10.3982/ECTA9626.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli, 19,* 521–547. https://doi.org/10.3150/11-BEJ410.

Belloni, A., Chernozhukov, V., & Hansen, C. (2010). Inference for high-dimensional sparse econometric models. In Advances in Economics and Econometrics. 10th World Congress of Econometric Society, Aug 2010 III, pp. 245–295. ArXiv, 2011.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives, 28,* 29–50. https://doi.org/10.1257/jep.28.2.29.

Belloni, A., Chernozukov, V., & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies, 81,* 608–650.

Bishop, C. M. (2006). *Pattern recognition and machine learning. Information science and statistics*. Springer (**Softcover published in 2016**).

Blake, T., Nosko, C., & Tadelis, S. (2015). Consumer heterogeneity and paid search effectiveness: a large-scale field experiment. *Econometrica, 83,* 155–174. https://doi.org/10.3982/ECTA12423.

Brealey, R. A., Myers, S. C., & Franklin, A. (2020). *Principles of corporate finance* (13th ed.). McGraw-Hill Education.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications. Springer series in statistics*. Springer.

Chandra, K., Plaschke, F., & Seth, I. (2018). Memo to the CFO: get in front of digital finance - or get left back. McKinsey & Company. https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/memo-to-the-cfo-get-in-front-of-digital-finance-or-get-left-back. Accessed 10 Dec 2020.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review, 107,* 261–65. https://doi.org/10.1257/aer.p20171038.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21,* C1–C68. https://doi.org/10.1111/ectj.12097.

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: an elementary, general approach. *Annual Review of Economics, 7,* 649–688. https://doi.org/10.1146/annurev-economics-012315-015826.

Chernozhukov, V., Hansen, C., & Spindler, M. (2016). High-dimensional metrics in R. arXiv:1603.01700v2.

Conine, T. C., & McDonald, M. (2017). The application of variance analysis in FP&A organizations: survey evidence and recommendations for enhancement. *Journal of Accounting and Finance, 17,* 54–70.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting, 22,* 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001 (**twenty five years of forecasting**).

Economist (2017). The world's most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. Accessed 6 Dec 2020.

Fischer, E. O. (2009). *Finanzwirtschaft für Anfänger. Lehr- und Handbücher zur entscheidungsorientierten Betriebswirtschaft*. Oldenbourg.

Gajewar, A., & Bansal, G. (2016). Revenue forecasting for enterprise products. arXiv:1701.06624.

Gandomi, A., & Haider, M. (2015). Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management, 35,* 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Garrison, R. H., Noreen, E. W., & Brewer, P. C. (2006). *Managerial accounting*. McGraw-Hill/Irwin.

Gray, G. L., & Alles, M. (2015). Data fracking strategy: why management accountants need it. *Management Accounting Quarterly, 16,* 22–33.

Hansen, S. C. (2011). A theoretical analysis of the impact of adopting rolling budgets, activity-based budgeting and beyond budgeting. *European Accounting Review, 20,* 289–319. https://doi.org/10.1080/09638180.2010.496260.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed.). Springer-Verlag.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124,* 226–251. https://doi.org/10.1016/j.eswa.2019.01.012.

Koenecke, A., & Gajewar, A. (2020). Curriculum learning in deep neural networks for financial forecasting. In V. Bitetta, I. Bordino, A. Ferretti, F. Gullo, S. Pascolutti, & G. Ponti (Eds.), *Mining data for financial applications* (pp. 16–31). Springer International Publishing.

Küsters, U., McCullough, B. D., & Bell, M. (2006). Forecasting software: past, present and future. *International Journal of Forecasting, 22,* 599–615. https://doi.org/10.1016/j.ijforecast.2006.03.004 (**twenty five years of forecasting**).

Laney, D. (2001). 3-D data management: controlling data volume, velocity and variety. Application Delivery Strategies by META Group Inc., Gartner. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocity-andVariety.pdf. Accessed 30 July 2020.

Möller, K., Schäffer, U., & Verbeeten, F. (2020). Digitalization in management accounting and control: an editorial. *Journal of Management Control, 31,* 1–8. https://doi.org/10.1007/s00187-020-00300-5.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives, 31,* 87–106. https://doi.org/10.1257/jep.31.2.87.

Oesterreich, T. D., Teuteberg, F., Bensberg, F., & Buscher, G. (2019). The controlling profession in the digital age: understanding the impact of digitisation on the controller's job roles, skills and competences. *International Journal of Accounting Information Systems*. https://doi.org/10.1016/j.accinf.2019.100.

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: a survey. *Applied Soft Computing, 93,* 106384. https://doi.org/10.1016/j.asoc.2020.106384.

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM, 62,* 54–60. https://doi.org/10.1145/3241036.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect* (1st ed.). Basic Books Inc.

Roos, A., Tucker, J., Rodt, M., Stange, S., Ego, P., Boudadi, A., & Sheth, H. (2020). Lessons from best-in-class CFOs. Boston Consulting Group. https://www.bcg.com/publications/2020/lessons-best-in-class-cfos. Accessed 29 July 2020.

Ross, S. A., Westerfield, R. W., & Jordan, B. D. (2019). *Fundamentals of corporate finance* (12th ed.). McGraw-Hill Education.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association, 100,* 322–331. https://doi.org/10.1198/016214504000001880.

Strauß, E., & Zecher, C. (2013). Management control systems: a review. *Journal of Management Control, 23,* 233–268. https://doi.org/10.1007/s00187-012-0158-7.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction. Adaptive computation and machine learning series*. MIT Press.

Taddy, M. (2019). *Business data science: combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw-Hill Education.

Tucker, J., Foldesy, J., Roos, A., & Rodt, M. (2017). How digital CFOs are transforming finance. Boston Consulting Group. https://www.bcg.com/publications/2017/function-excellence-how-digital-cfo-transforming-finance. Accessed 10 Dec 2020.

Unger, G., & Rodt, M. (2019). The art of forward-looking steering: the power of algorithmic forecasting. Boston Consulting Group. https://www.bcg.com/publications/2019/power-of-algorithmic-forecasting. Accessed 30 Nov 2020.

Varian, H. (2018). Artificial intelligence, economics, and industrial organization. *The economics of artificial intelligence: an agenda* (pp. 399–419). University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.001.0001.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 3,* 1–130. https://doi.org/10.2200/S00196ED1V01Y200906AIM006.