Machine Learning with Applications 11 (2023) 100453

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Predicting firm performance and size using machine learning with a Bayesian perspective

Debdatta Saha [a,*], Timothy M. Young [b], Jessica Thacker [a]

[a] Faculty of Economics, South Asian University, New Delhi, India
[b] The University of Tennessee | Center for Renewable Carbon | Bredesen Center, 2506 Jacob Drive, Knoxville, TN 37996-4563, USA

ARTICLE INFO

Keywords:
Machine learning
Firm financial performance
Performance forecasting
Bayesian additive regression tree (BART)

ABSTRACT

This paper investigates the issue of predicting the financial performance of firms in registered manufacturing in developing countries using machine learning methods along with economic theory to explain the findings. While literature suggests that predictability of top line measures related to sales is lower compared to bottom line measures such as net profits for small informal establishments in developing countries, we find the opposite holds true for firms in registered manufacturing in the food processing industry in India based on the results from machine learning techniques of Bayesian additive regression trees (BART), boosted trees, bootstrap forests, and regression tree algorithms. BART models in validation outperformed the other algorithms in predictability of the dependent variables. Across ten validation studies, BART had an average R^2 ranging from 0.922 to 0.934 and boosted tree models had an average R^2 ranging from 0.873 to 0.905 for predicting sales. A key significant independent variable for predicting sales across all categories and algorithms was real raw material expenses explaining approximately 83% to 88% of the total sums of squares in all validations. This is in line with the realities of the food processing industry, which is intensive in its raw material usage. The dependent variable 'profits' (as measured by real profits before depreciation, interest, taxes and amortization, 'real_pbdita') was more difficult to predict relative to sales. BART models again outperformed the other algorithms in validation with an average R^2 ranging from 0.745 to 0.818. Key significant variables in the models were more diverse where raw material expenses, compensation to employees, and net- or total-fixed assets explained the largest proportions of the total sums of squares. The results from a machine learning approach with a Bayesian perspective can enhance the understanding of the mechanisms that translate sales into profits for registered manufacturing, thereby aiding policy-making for small businesses in the formal sector in developing countries.

1. Introduction

Organizational performance is a multidimensional construct (see Richard, Devinney, Yip, & Johnson, 2009 for details), of which financial performance that forms its component is assumed to indicate attainment of economic goals of the organization (Venkatraman & Ramanujam, 1986). However, there is no consensus on the appropriate measure for financial performance of the firm in the literature. Techniques from artificial intelligence and machine learning have been used recently in finance in profusion to investigate this issue. However, the spirit of enquiry has been that of finding out the appropriate financial ratio or variable that can be best predicted in a particular financial context. What this leaves unaddressed is the following: what is the underlying economic logic that explains the predictive power of a particular financial variable? In our approach to this research question, we integrate the findings from machine learning techniques with economic theory regarding firm size and search costs for product placement. While our investigation is limited to firms in registered manufacturing of processed food in India, we believe this approach of embedding the results from machine learning models with economic logic adds additional insight to the channels that drive these results. From an economic policy perspective, this can act as an important guide.

Our central concern is to investigate which entity captures financial performance the best: top line measures, such as sales or bottom line ones, such as profits, when we analyze standardized financial reports of a comparable set of firms in a particular industry. Be it the usage of financial ratios, such as return on assets for conducting an efficiency analysis of a firm (such as in Ozcan & McCue, 1996), the question remains: should the top line (related to sales or revenue) or the bottom line of net profits/net returns be used?

Multiple studies have used both top line and bottom line measures, such as Chen, Motiwalla, and Riaz Khan (2004), where both revenue and net income are used as outputs in the data envelopment analysis

* Corresponding author.
E-mail addresses: debdatta@sau.ac.in (D. Saha), tmyoung1@utk.edu (T.M. Young), jessica.thacker@students.sau.ac.in (J. Thacker).

https://doi.org/10.1016/j.mlwa.2023.100453
Received 24 February 2022; Received in revised form 21 November 2022; Accepted 25 January 2023
Available online 2 February 2023

(DEA) of firm performance or in the creation of a financial performance index, Miyakawa, Miyauchi, and Perez (2017) use machine learning methods to anticipate firm exit, growth of sales as well as profits, Robert Baum and Wally (2003) study the impact of strategic decision making speed on firm performance, where firm performance is measured by sales as well as employment growth and net profits before tax as a percentage of assets. Generally, this approach gives equal importance to the stages of revenue and profit performance. However, top line measures capture a very different aspect of financial performance in comparison to bottom line. The context matters. For instance, De Mel, McKenzie, and Woodruff (2009) finds that predicting top lines are much more difficult for micro- enterprises in developing countries like Sri Lanka than bottom lines, mostly due to lack of proper books of accounts and usage of recall of enterprise owners to register sales. Recall records are subject to error due to memory lapses, as well as lack of interest on the part of the owner to reveal incomes. That microenterprises mostly do not maintain books following modern accounting standards has been noted in other developing countries, such as Africa (see Agyei-Mensah, 2010; Lesser & Moisé-Leeman, 2009; Maseko & Manyani, 2011). Does this apply to other contexts? What about small firms in registered manufacturing in developing countries, which have to maintain not only books of accounts but also respect processing standards? This is what motivates our paper. Here, we investigate the reliability of working with measures of sales/revenue as opposed to profits in another developing country context (India) but for organized manufacturing in a particular industry, viz. food processing.

What we find is that predicting sales is easier than profits for the class of firms we study. The problem addressed by this study is to expand the current literature by predicting firm performance in India using contemporary machine learning algorithms from a Bayesian perspective. There is a plethora of literature using ML techniques to predict financial market indicators, credit risk prediction, financial engineering for equities and bond trading etc. However, we think our study is unique in by predicting firm performance in India with contemporary methods and it will initiate new research in the applications of ML to economic research. We use the stylized facts of the food processing industry to explain why revenue or sales figures might be more predictable than profits. Our claim is that the standards of converting inputs to saleable output as well as standardized books of accounts, whatever be the size of the firm as long as they are a part of registered manufacturing, ensures high predictability of top line measures (like sales). On the other hand, beyond the stage of final output generation, there are multiple pathways to generate profits (such as business to business sales or industrial sales, brand-based marketing, wholesale marketing etc. (see Sutton, 1991). This heterogeneity in firm strategies in the last mile is likely to show up in the data as higher predictability in sales rather than in profits.

This paper, therefore, contributes to the literature on measurement of firm performance by integrating economic theory with contemporary machine learning methods. Our results, coupled with economic logic, demonstrate that standardized reporting of accounts and diversity of marketing channels in registered manufacturing make top line measures more predictable than bottom. The opposite holds true for the analysis of De Mel et al. (2009), which investigates the same question (without using any machine learning techniques) for informal sector firms in Sri Lanka. Our result highlights that firm structure, rather than size, for which the financial measure is analyzed is of prime importance: accounting and book-keeping in registered manufacturing is very different from that in unregistered manufacturing. Our result holds even with a majority of very small-sized firms, just that these are a part of registered manufacturing.

Relevant Literature

In terms of the method of our investigation, our paper is part of the rapidly expanding field of applications of machine learning techniques in economics. Given that this is not a review of paper of existing techniques, we make a small mention of some current applications of ML for prediction tasks in economics, such as customer behavior, market time series prediction, cryptocurrencies, price prediction and bankruptcy prediction. A thorough review is given in Nosratabadi et al. (2020). Note that Goodell, Kumar, Lim, and Pattnaik (2021) provides a comprehensive review of AI and ML in finance using bibliometric techniques. They mention three important research themes in recent times in this field: "(1) portfolio construction, valuation and investor behavior; (2) financial fraud and distress; and (3) sentiment inference, forecasting and planning". From the current direction of recently published papers, this seems to be largely correct. Notable citations with ML applications in financial performance and predicting bankruptcy or financial distress are: (Barboza, Kimura, & Altman, 2017; Bargagli-Stoffi, Niederreiter, & Riccaboni, 2021; Bohanec, Borštnar, & Robnik-Šikonja, 2017; Bragoli et al., 2021; Chen & Du, 2009; Delen, Kuzey, & Uyar, 2013; Foroghi & Monadjemi, 2011; Lam, 2004; Miyakawa et al., 2017; Moscatelli, Parlapiano, Narizzano, & Viggiano, 2020; Sun & Hui, 2006). Miyakawa et al. (2017) use ML to forecast firm performance for Japanese firms. Bohanec et al. (2017) use ML for predicting B2B sales and Qiu, Srinivasan, and Hu (2014) study the usage of annual reports to forecast firm performance using supervised learning methods.

It is important in the context of our study to note the contribution by März, Klein, Kneib, and Musshoff (2016) where a Bayesian geo-additive quantile regression approach with gradient boosting was used to estimate farmland rental rates. Closely related to our work is Pap, Mako, Illessy, Kis, and Mosavi (2022), which uses genetic algorithms and BART ML to predict the importance of each of the intra- and extra-organizational factors in identifying the firms' performance as well as employee well-being in Europe using the European Company Survey 2019 data, whereas our work is in the context of an emerging economy like India.

Our research is related to the cluster of research papers that work with financial analysis using variables or ratios from annual reports of firms. There are, roughly, three ways of analyzing financial variables/ratios from balance sheets/annual statements drawn from company annual reports. One set of papers applies ML techniques to textual analysis of both standardized format as well as non-standard annual reports (Back, Toivonen, Vanharanta, & Visa, 2001; Kloptchenko, Eklund, Back, Karlsson, Vanharanta, & Visa, 2002; Qiu et al., 2014). Financial ratios or variables are a part of this textual analysis. A second category of papers combines textual analysis with a separate investigation of financial variables of the firm (Lang & Stice-Lawrence, 2015; Li, 2010). A third branch of the literature considers only the financial variables and applies ML techniques on them (Delen et al., 2013; Miyakawa et al., 2017). Our paper is closely linked to this last branch of the literature. However, we differ from the approach of papers in this branch in our investigation method, which we discuss below.

Investigation Technique

Our research question is intimately related to the issue of how firm size is measured, as economic logic indicates firm size should influence the predictive power of sales or profits. Hence, we create two ways in which firm size can be measured: by net fixed assets and by total assets.

Our investigation technique as follows:

Step 1. We create four categories of variables for each measure of firm size, which includes both bottom line and top line measures of firm performance:

- Category 1 (measures firm size by net fixed asset) controls for the cost of indigenous raw materials
- Category 2 (measures firm size by total assets) does not control for the cost of indigenous raw materials
- Category 3 (measures firm size by net fixed asset) introduces various kinds of distribution costs: advertising, marketing and distribution expenses
- Category 4 does the same thing as category 3 but measures firm size by total assets

In this manner, we try to test for the effect of firm size and different distribution cost channels on the predictive power of top line and bottom line measures of a firm's financial performance.

Step 2. We apply four ML techniques, viz. boosted trees, random forests, regression trees and Bayesian Adaptive Regression Tree (BART) to the four categories of variables to identify whether any method predicts either top line or bottom line performance better than the others.

Step 3. We apply economic logic and descriptive statistics to explain the results from the ML analysis.

Contrasting our technique with existing literature

Most papers start with a set of different financial ratios that measure different dimensions of performance, such as liquidity, profitability, solidity, efficiency etc. (such as Kloptchenko et al., 2002) and use ML techniques on all these ratios to identify the ratio that has the highest predictive ability. The difference between top line performance versus bottom line is lost in this analysis. In the process, the economic logic that drives the difference in performance between these two stages of firm performance has no place in this literature.

Additionally, there is very limited literature for registered manufacturing in India which applies similar ML techniques as we do. Some papers have applied methods such as random forests and tree nets (see Shrivastava, Kumar, Kumar, & Gupta, 2020), artificial neural networks (ANN) and support vector machine (SVM) (see Sehgal, Mishra, Deisting, & Vashisht, 2021) in the context of corporate distress among firms in the Indian context. Corporate distress, though related, is different from measurement of firm performance. In the Indian context, we found only Manogna and Mishra (2021), who identify various financial measures that impact firm performance using Chi-squared automatic interaction detector (CHAID), classification and regression trees (CART), C5.0 and quick, unbiased, efficient statistical tree (QUEST). The study by Pap et al. (2022) used genetic algorithms and BART to predict the importance of each of the intra- and extra-organizational factors in identifying the firms' performance as well as employee well-being in Europe using the European Company Survey 2019 data. **However, this leaves open the question that we address in our research: is it at the stage of sales or at a later stage of profits. This, to the best of our knowledge, has not been answered to date.**

The paper is structured as follows: Section 2 provides the data sources and methodology utilized in our study, Section 3 presents the results with economic intuition behind it and Section 4 concludes.

2. Data sources and methodology

2.1. Dataset and variables studied

We use the CMIE sponsored Prowessdx (for further details, please see https://prowessdx.cmie.com/) for study from 2006–07 to 2018–19 for the registered food manufacturing sector of India. The dependent variable for our analysis is sales deflated by wholesale price index (WPI) of food products (FP) and profits before depreciation, interest, tax and amortization (PBDITA) deflated by WPI of FP. The independent variables for the analysis are classified under four categories where the classification is based on the different measures of firm size[5] namely, net fixed assets and total assets as represented in Table 1. The record length of the public domain dataset was N =337 with 14 independent variables. It is important to note that this type of firm economic data are difficult to collect and organize. Regression tree-based machine learning algorithms do not require normalization and scaling of the data and tend to accommodate poor dataset dimensionality relative to other modeling techniques such as PLS and deep learning. The computational complexity of the machine learning algorithms used in the study with R$^{\rightarrow}$ and SAS$^{\rightarrow}$ software was satisfactory. The code was executed on a conventional 64-bit operating system laptop with an Intel(R) Core (TM) i7-7700HQ CPU @ 2.80 GHz processor with 16.0 GB RAM. Clock times for closure of most algorithms during training and cross validation varied from two to approximately 45 min.

The independent variables common under all these four categories are raw material expenses, energy expenses which comprise power, fuel and water charges and compensation to employees. In the first two categories selling and distribution expenses is studied as aggregate measure, which includes all the advertising, marketing (including rebates and discounts given to consumers, business promotion expenditures, market survey or research expenses) as well as distribution expenses (including all the expenses that the firm incurs in order to deliver its product to the buyer). In category 3 and 4, these selling and distribution expenses are disaggregated into three different sub-categories to study the independent effect of the variables of advertising, marketing and distribution expenses of the firm. Category 1 and category 3 uses net fixed assets (NFA) as a measure of firm size and on the other hand, category 2 and category 4 uses total assets as a measure of firm size. The variables marked in blue are unique to each of these categories to isolate the firm size effect, depending on the usage of indigenous raw materials and disaggregated distribution costs in the four separate categories.

2.2. Machine learning algorithms of the study

2.2.1. Regression trees

Regression trees are part of the methodology of decision trees when the dependent variable is continuous and form the theoretical basis for the advancement of tree-based methods and machine learning. A regression tree is a piecewise constant or piecewise intrinsically linear estimate of a regression function, constructed by recursively partitioning the data and sample space (Loh, 2011). A decision tree partitions the data space of all joint regressor values X into J- disjoint regions $(R_j)_1^J$ (Friedman, 2001). For a given set of joint regressor values X, the tree prediction $\hat{Y} = T_j(X)$ assigns as the response estimate where T is defined as original regression tree, the value assigned to each region containing X:

$$X \in R_j \Rightarrow T_j(X) = \hat{y}_j \tag{1}$$

Given a set of regions, the optimal response values associated with each region are easily obtained, namely the value that minimizes prediction error (risk) in that region:

$$Y_j = arg\, min E_y \left[L(y, y' | X \in R_j) \right] \tag{2}$$

As noted by Friedman (2001), the difficult problem is to find a good set of regions $(R_j)_1^J$. Unlike kernel methods, decision trees attempt to use the data to estimate a good partition instead of a user defined model. Even though regression trees have high explanatory value, the method in which the data space is partitioned results in poor predictability.

2.2.2. Bootstrap forests

'Bootstrap forests' are a tree-based modeling method that builds a collection of individual decision trees and takes the average of their predictions (Breiman, 2001). To build a collection of decision trees, random forest employs a technique known as 'bagging', which combines the results of decision tree models on hundreds of bootstrapped training samples (Breiman, 1996). In order to make these models as independent from each other as possible, each tree is modified in two ways. First, each tree selects a random set of explanatory variables every time a new rule is added. Where p is the number of explanatory variables, either $p/3$ or sqrt(p) are selected, and the best rule is chosen from this subset of variables so that no single split has access to majority of the variables. Subsequently, each tree is built to the point where each leaf contains only one individual. By averaging together this collection of trees, each having random sets of individuals and runs, the prediction model will be more robust and accurate than a model built from a single decision tree (Basuchoudhary, Bang, & Sen, 2017; Breiman, 2001; Matthias & Zou, 2020).

Table 1
Independent variables studied by category grouping (with variable code name in italics).
Source: Authors' own creation.

Category 1 - Indigenous raw material cost as control (size measured by net fixed assets)	Category 2 - No control for indigenous raw material cost (size measured by total assets)
raw material expenses *(real_rawmat_exp)*	raw material expenses *(real_rawmat_exp)*
energy expense (fuel, power, water costs) real_power_fuel_water_charges	energy expense (fuel, power, water costs) *(real_power_fuel_water_charges)*
employee compensation *(real_compensation_to_employees)*	employee compensation *(real_compensation_to_employees)*
selling & distribution costs *(real_selling_distribution_exp)*	selling & distribution costs *(real_selling_distribution_exp)*
net fixed assets *(NFA_FBT)*	*total assets (r_total_assets_FBT)*
indigenous raw material cost (sa_indig_rawmat_pc_total_rawmat)	
Category 3 – Effect of disaggregated distribution costs (size measured by net fixed assets)	**Category 4 – Effect of disaggregated distribution costs (size measured by total assets)**
raw material expenses *(real_rawmat_exp)*	raw material expenses *(real_rawmat_exp)*
energy expense (fuel, power, water costs) *(real_power_fuel_water_charges)*	energy expense (fuel, power, water costs) *(real_power_fuel_water_charges)*
employee compensation *(real_compensation_to_employees)*	employee compensation *(real_compensation_to_employees)*
indigenous raw material cost *(sa_indig_rawmat_pc_total_rawmat)*	indigenous raw material cost *(sa_indig_rawmat_pc_total_rawmat)*
Advertising expenses *(real_advertising)*	Advertising expenses*(real_advertising)*
Marketing expenses *(real_marketing)*	Marketing expenses*(real_marketing)*
Distribution expenses *(real_distribution_exp)*	Distribution expenses*(real_distribution_exp)*
net fixed assets *(NFA_FBT)*	total assets *(r_total_assets_FBT)*

Bootstrap forest is a model that is often used as a benchmark performer for data competitions (Fodor, 2013; Triguero, Río, López, Bacardit, Benítez, & Herrera, 2015). In terms of complexity, the tuning parameters used in bootstrap forest models require minimal customization and can often be kept at their default values (Oshiro, Perez, & Baranauskas, 2012). However, unlike regression trees, the interpretability of bootstrap forest is low due to the possible hundreds of trained trees used in the model, and like most tree-based models, it is prone to overfitting *i.e.,* 'greedy algorithm' (Kuhn & Johnson, 2013).

2.2.3. Boosted trees

'Boosted trees' are similar to bootstrap forests in their use of combined results from multiple trees. Rather than using the bagging approach, however, boosted trees use 'boosting', which builds each tree off the results of the previous (Friedman, 2001; Friedman & Meulman, 2003). Unlike bootstrap forest, each tree in the boosting method is kept simple (the number of splits, or interaction depth, typically reaching only one or two). Boosted trees combine many weak learners into a stronger classifier, where each tree is created iteratively (Friedman, 2001; Friedman & Meulman, 2003). The tree's output ($h(x)$) is given a weight (w) relative to its accuracy. The ensemble output is the weighted sum (Hastie, Tibshirani, & Friedman, 2009): .

$$\hat{y}(x) = \sum_t w_t h_t(x) \tag{3}$$

After each iteration each data sample is given a weight based on its misclassification, *i.e.,* the more often a data sample is misclassified, the more important it becomes. The goal is to minimize an objective function: ,

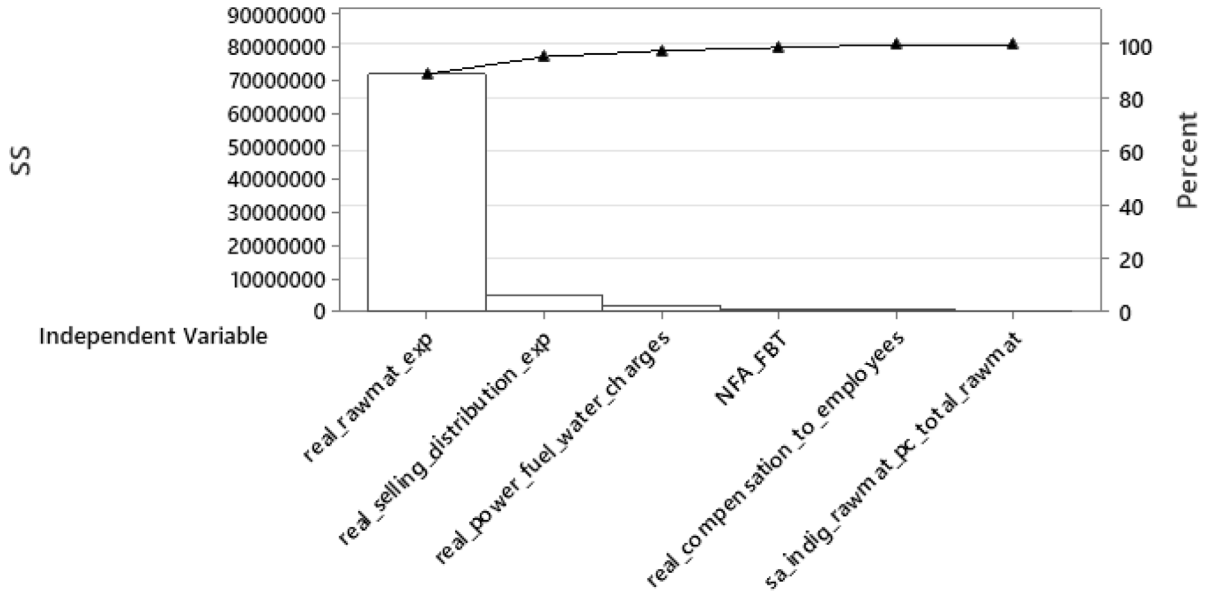
$$O(x) = \sum_i l(\hat{y}_i, y_i) + \sum_t \Omega(f_t) \tag{4}$$

where, $l(\hat{y}_i, y_i)$ is the loss function, *i.e.,* the distance between the truth and the prediction of the ith sample, and $\Omega(f_t)$ is the regularization function, *i.e.,* it penalizes the complexity of the ith tree. The model's final predictions are a weighted sum of individual tree predictions, with each tree improving on the results of the previous tree (Hastie et al., 2009; Mason, Baxter, Bartlett, & Frean, 1999).

Boosting focuses on a learning rate parameter (typically slower than bootstrap forests). A low learning rate allows the model to learn slowly and parse out more subtle interactions within the data, which takes more time and a larger number of trees. A high learning rate allows the model to obtain a more generalized picture of the data using less time and fewer trees. In general, the low learning rate will output a more robust model but takes more computing power. Other parameters that require consideration are the number of trees, the interaction depth of each tree, and the number of individuals each partition in the tree can contain (Ridgeway, 2007). Like bootstrap forest, the results of a boosted tree model can be difficult to interpret, and the model is prone to overfitting (Oshiro et al., 2012; Triguero et al., 2015). Since there are more parameters to consider, boosted tree models are also more difficult to tune than other tree-based models.

2.2.4. Bayesian Additive Regression Trees (BART)

BART was proposed and developed by Chipman, George, and McCulloch (2002, 2010) as an efficient ensemble method for machine learning, combining recent advances in Bayesian modeling with regression tree ideas from machine learning to sensibly search the potential dimensional space of possible models relating the response (Steiner, Zeng, Young, Edwards, Guess, & Chen, 2016). As Mehta (2022) notes, BART is based on the popular Bayes theorem which is used to calculate the posterior probability. Pap et al. (2022) cites its popularity in a multifarious range of applications and notes its usefulness as a nonparametric function estimation using regression trees. As suggested by the name of Bayesian Additive Regression Tree, BART consists of a sum-of-trees model (additive model) and regularization prior (Bayesian method). Mehta (2022) summarizes the key differences of BART with other tree-based machine learning methods, i.e., "the inferences obtained from BART are based on successive iterations of the back fitting



SS	71966554	5192794	1841486	967518	901366	0
Percent	89.0	6.4	2.3	1.2	1.1	0.0
Cum %	89.0	95.4	97.7	98.9	100.0	100.0

Fig. 1. Sums of squares contribution for 'Category 1' independent variables for measuring sales.
Source: Authors' own creation.

algorithm which are effectively an MCMC sample from the induced posterior probability over the sum-of-trees model space. Building a sum of trees and regularization of prior are the two major things that define the BART model". BART has three key aspects: (1) sum-of-trees model; (2) regularization prior and (3) MCMC algorithm (Hill 2010, (Chipman et al., 2010)). Let T_i denote a single tree of B terminal nodes, $M_i = \{\mu_1, \mu_2, \ldots, \mu_B\}$ represent a set of parameter values associated with the B terminal nodes of T_i. Also, instead of fitting a single tree, there are now m trees in the model. Function $g(x;T_i, M_i)$ is defined corresponding to specific (T_i, M_i) which assigns a $\mu \in M$ to a subset of X associated with some terminal node. Now values of response variable Y can be expressed as:

$$Y = \sum_{i=1}^{m} g\left(x; T_i, M_i\right) + \epsilon, \epsilon \sim N\left(0, \sigma^2\right). \tag{5}$$

So, the Bayesian approach using regularization priors is adopted to regularize the model fitting, *i.e.,* a prior is placed on each of three major parameters in Eq. (5), the T_i, its terminal node parameters M_i, and σ. To simplify the prior specification, all T's are assumed to be independent on prior and identically distributed (*i.i.d*); all μ's of M are *i.i.d* given all T's; and σ are independent of all T's and μ's (Chipman et al., 2010).

After placing priors on parameters T_i, M_i, and σ, a posterior distribution $p(y)$ is computed and sampled using Markov Chain Monte Carlo (MCMC), *i.e.,* Gibbs sampling. As summarized in Hill (2011), during MCMC iterations sampling parameters fitting Y in Eq. (5), the size of each of m trees may vary from iteration to iteration. The (T_i, M_i) pair for one tree may be switched to another tree in next time's iteration; the contribution of a particular tree cannot be identified. After a series of MCMC K iterations a sample of K fitting results of Y, $\{\hat{Y}_1, \ldots, \hat{Y}_k\}$ are formed. To estimate or predict Y as

$$\hat{Y} = \frac{1}{K} \sum_{i=1}^{K} \hat{Y}_i. \tag{6}$$

3. Results and discussion

3.1. Descriptive statistics

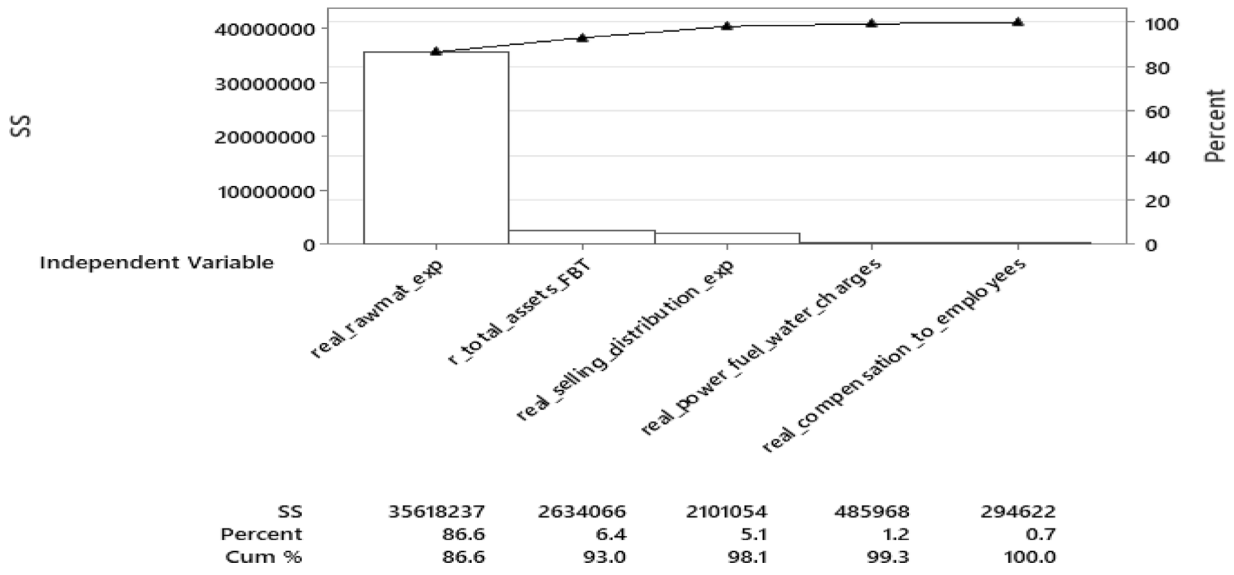
We find that the mean value of sales in our data for all firms is INR 35.11 million, whereas for profits is INR 2.56 million. The standard deviation of sales is INR 112.22 million in contrast with INR 9.98 million for profits. However, the coefficient of variation in both is roughly similar, 3.20 for sales and 3.90 for profits respectively. The overall correlation between sales and profits is not perfect. It stands at 0.683, indicating some degree of diversification of incomes other than from selling produced outputs. Table A.1 in the Appendix shows some of the descriptive statistics emerging from our data when we classify firms by either total assets or net fixed assets. Understandably, the sales (and profits) of firms in the bottom one percent of the distribution of assets (be it total or net fixed assets) considerably lower than that for firms in the highest percentile. More interestingly, the correlation between size and profits is not constant or even with the same sign across the different percentiles of the distribution of asset holdings of firms in our data. A clearer comment on the prediction of sales as opposed to profits from our data comes from our regression tree analysis, BART in particular. The following section discusses the methods we employ to comment on this issue.

3.2. Predicting sales by category and key factors

BART models and boosted tree models performed better in predicting sales across all four categories. Across ten validation studies for the four categories, BART had an average R^2 (R^2 is the coefficient of determination in the validation data sets) ranging from 0.922 to 0.934 for predicting sales (Table 2). Boosted tree models had an average R^2 in validation of 0.873 to 0.905 (Table 2). The key significant independent variable for the categories 1, 2 and 3 was raw material expenses explaining approximately 89%, 87%, and 88% of the total sums of squares across ten validation studies for sales (Figs. 1, 2, and 3). However, when total asset is used for measuring firm size in place

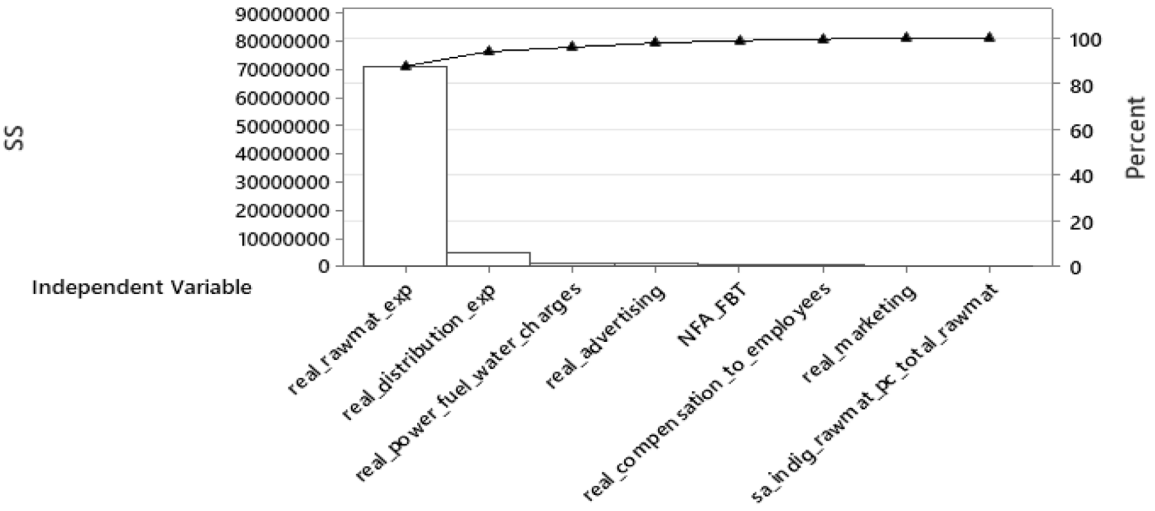
Table 2
R^2 and RMSEP% for sales by algorithm.
Source: Source: Authors' own calculations.

Category 1 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.794	3.79	0.976	1.30	0.928	2.23	0.838	2.41
2	0.790	3.06	0.817	1.93	0.915	1.6	0.937	1.5
3	0.938	2.46	0.955	2.10	0.939	1.85	0.979	1.67
4	0.874	1.98	0.867	2.04	0.805	2.46	0.985	1.14
5	0.883	1.25	0.835	1.26	0.695	4.02	0.971	1.14
6	0.945	2.32	0.866	3.64	0.948	2.28	0.837	2.65
7	0.765	3.53	0.8	3.26	0.78	3.42	0.962	1.44
8	0.814	4.23	0.748	4.93	0.861	3.7	0.929	1.08
9	0.964	1.01	0.966	0.98	0.862	1.96	0.932	2.25
10	0.909	3.61	0.903	3.71	0.925	3.27	0.969	1.89
Average:	**0.868**	**2.72**	**0.873**	**2.52**	**0.866**	**2.68**	**0.934**	**1.72**
SD	**0.072**	**1.10**	**0.076**	**1.30**	**0.083**	**0.85**	**0.054**	**0.56**
Category 2 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.890	2.05	0.950	1.90	0.929	2.49	0.886	3.19
2	0.864	1.23	0.870	1.66	0.917	1.79	0.928	1.65
3	0.907	2.24	0.918	2.88	0.941	2.70	0.981	0.98
4	0.868	1.50	0.867	1.14	0.808	2.74	0.997	0.99
5	0.801	3.18	0.768	2.74	0.695	5.17	0.946	1.19
6	0.864	2.72	0.960	2.04	0.948	2.56	0.806	3.17
7	0.786	2.50	0.766	3.60	0.780	4.00	0.988	0.97
8	0.778	3.42	0.869	3.63	0.861	2.51	0.889	2.24
9	0.967	1.15	0.954	1.70	0.862	3.85	0.937	1.89
10	0.811	3.85	0.911	3.64	0.925	3.68	0.927	1.62
Average:	**0.854**	2.39	**0.833**	2.49	**0.867**	3.15	**0.929**	1.79
SD	**0.060**	**0.93**	**0.071**	0.93	**0.083**	1.01	**0.058**	0.85
Category 3 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RASE %	R^2	RASE%	R^2	RASE%	R^2	RASE %
1	0.834	1.82	0.957	1.32	0.738	3.23	0.762	2.15
2	0.856	1.72	0.962	1.28	0.846	3.12	0.997	0.92
3	0.842	1.88	0.965	1.20	0.762	4.03	0.994	1.05
4	0.890	1.66	0.901	2.81	0.510	6.72	0.948	1.71
5	0.728	2.93	0.805	1.99	0.129	9.53	0.914	1.53
6	0.769	2.74	0.952	1.35	0.737	3.29	0.909	1.89
7	0.788	2.92	0.769	2.33	0.815	3.47	0.965	1.28
8	0.685	3.11	0.846	2.66	0.582	6.95	0.834	2.79
9	0.969	1.40	0.935	1.85	0.632	6.01	0.922	1.95
10	0.765	2.96	0.904	2.24	0.684	5.99	0.973	1.16
Average:	**0.813**	**2.31**	**0.899**	**1.90**	**0.644**	**5.23**	**0.922**	**1.64**
SD	**0.083**	**0.67**	**0.070**	**0.60**	**0.208**	**2.15**	**0.074**	**0.58**
Category 4 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP%	R^2	RMSEP%	R^2	RMSEP%	R^2	RMSEP %
1	0.616	2.49	0.958	1.27	0.734	2.72	0.873	3.58
2	0.843	1.86	0.952	1.29	0.822	1.20	0.984	1.62
3	0.673	2.72	0.940	1.85	0.754	3.09	0.952	1.40
4	0.873	0.96	0.838	2.30	0.396	4.73	0.969	1.48
5	0.671	1.84	0.908	2.04	0.100	6.22	0.876	2.31
6	0.689	1.96	0.950	1.35	0.713	2.54	0.882	2.33
7	0.817	1.01	0.775	2.92	0.769	2.21	0.924	1.71
8	0.649	2.82	0.865	2.06	0.539	4.24	0.893	2.20
9	0.906	1.18	0.948	1.37	0.531	4.34	0.929	1.75
10	0.665	1.92	0.915	1.96	0.670	3.63	0.981	1.72
Average:	**0.740**	**1.88**	**0.905**	**1.84**	**0.593**	**3.49**	**0.926**	**2.01**
SD	**0.107**	**0.67**	**0.061**	**0.53**	**0.246**	**1.45**	**0.044**	**0.64**
Definitions: **R^2** is the coefficient of determination measuring the goodness-of-fit of the model **RMSEP %** is the root mean square error of prediction (in percentage)								



SS	35618237	2634066	2101054	485968	294622
Percent	86.6	6.4	5.1	1.2	0.7
Cum %	86.6	93.0	98.1	99.3	100.0

Fig. 2. Sums of squares contribution for 'Category 2' independent variables for sales.
Source: Authors' own creation.



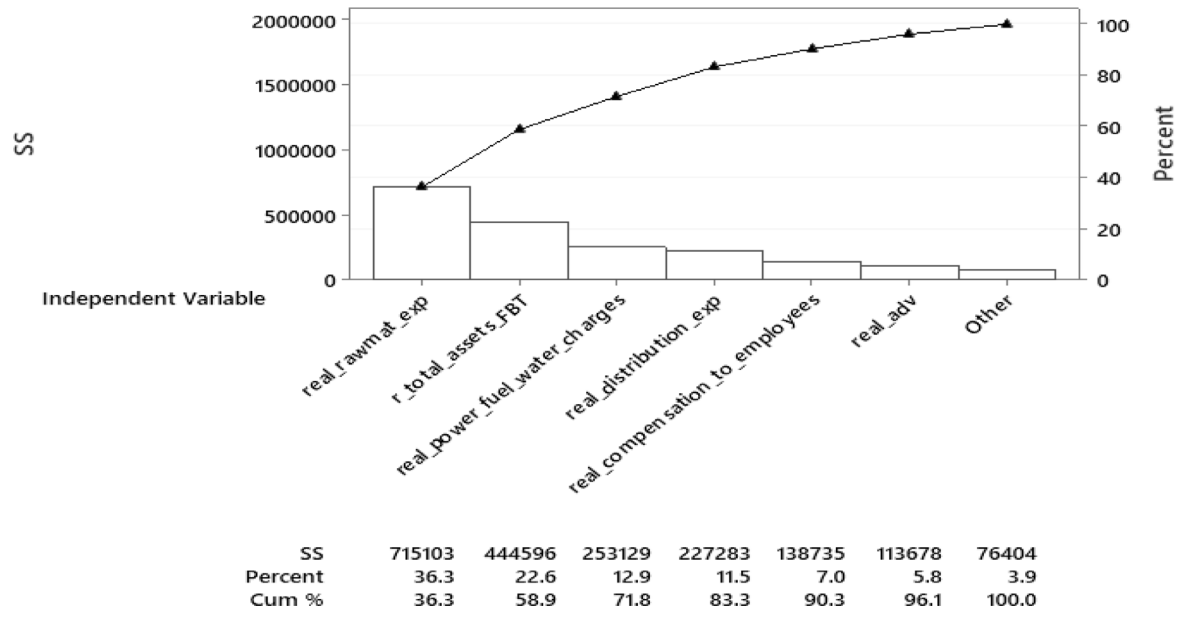
SS	712677	115201	333	1596446	1515434	756184	609630	278929	0
Percent	87.7	6.4		2.0	1.9	0.9	0.8	0.3	0.0
Cum %	87.7	94.1		96.1	98.0	98.9	99.7	100.0	100.0

Fig. 3. Sums of squares contribution for 'Category 3' independent variables for sales.
Source: Authors' own creation.

of net fixed assets then, raw materials, total assets, energy expenses and distribution expenses explains approximately 83% of the total sums of squares across ten validation studies for sales (Fig. 4). An example of the quality of the predictions by the algorithm for sales of the firms for Category 4 for validation study 10 is illustrated in Fig. 5. An example of the quality of the predictions by algorithm for sales of the firms for Category 4 for validation study 10 is illustrated in Fig. 5.

3.3. Predicting profits by category and key factors

The dependent variable profits (as measured by real profits before depreciation, interest, taxes and amortization, '*real_pbdita*') was more difficult to predict relative to sales. This is illustrated by the changes in independent variables across the four categories of independent variables studied. In ten cross validation studies, BART had an average R^2 ranging from 0.745 to 0.818 (Table 3). The other machine learning algorithms in several instances performed better than BART in validation, but across ten validation studies were inferior (Table 3). For Category 1 variables, raw material expenses, net fixed assets and selling and distribution expenses approximately 90% of the total sums of squares across ten validation studies (Fig. 6). For Category 2 variables, total assets, compensation to employees and raw material expenses explained approximately 93% of the total sums of squares across ten validation studies (Fig. 7). For Category 3 variables, net fixed assets, raw material expenses, distribution expenses and compensation to employees explained approximately 93% of the total sums of squares across ten validation studies (Fig. 8). For Category 4 variables, total assets, raw material expenses and compensation to employees explained approximately 96% of the total sums of squares across ten validation studies (Fig. 9). The quality of the predictions by algorithm for real profits for Category 4 for validation study 10 is illustrated in Fig. 10.



SS	715103	444596	253129	227283	138735	113678	76404
Percent	36.3	22.6	12.9	11.5	7.0	5.8	3.9
Cum %	36.3	58.9	71.8	83.3	90.3	96.1	100.0

Fig. 4. Sums of squares contribution for 'Category 4' independent variables for sales.
Source: Authors' own creation.

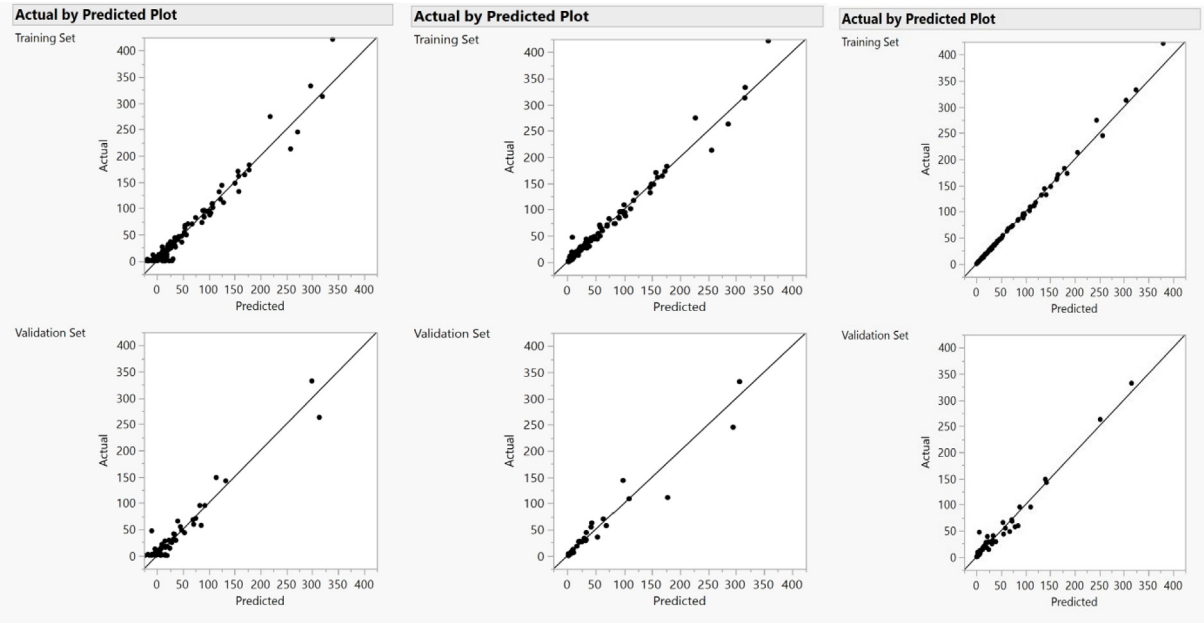


Fig. 5. Comparative predicted versus actual value of sales for validation study 10 for 'bootstrap forest' (left), boosted tree (middle), and BART (right) for 'Category 4' independent variables.
Source: Authors' own creation.

3.4. Discussion of results in the light of economic theory

In our dataset, it is not only that BART outperforms other methods in predicting performance, but we also find that sales are not highly correlated with profits and are more predictable compared to the latter. We provide a theoretical explanation below for why this is likely to be the case for firms in formal registered manufacturing in any country.

Consider a firm of size s in formal registered manufacturing, which produces a final product using a standard production technique z. This standardization applies to all firms in the industry comprising firms like this representative firm. The sales of the firm reflects the ability of the firm to convert inputs into saleable output using z. Beyond the stage of sales, the firm has to find avenues to successfully convert its activities into gross profits. We consider the following possibilities now:

(1) *Sales of output exclusively lead to profits:* This case occurs when the source of profits are not diversified beyond actual output. This leads to a tight positive correlation between sales and profits. This might lead to an immediate conclusion that in this case is that predictability of sales and profits are likely to be similar.

However, this is unlikely to be a correct conjecture if the route of converting top lines to bottom lines are also not standardized. In the case of most industries, firms in the manufacturing end of the spectrum have a number of options for converting their saleable output into profits. We consider the following possibilities:

A *Branded sales:* This mechanism ensures that the final product of the firm is marketed under the umbrella of specific brands. For example, in the food processing industry, particularly dairy

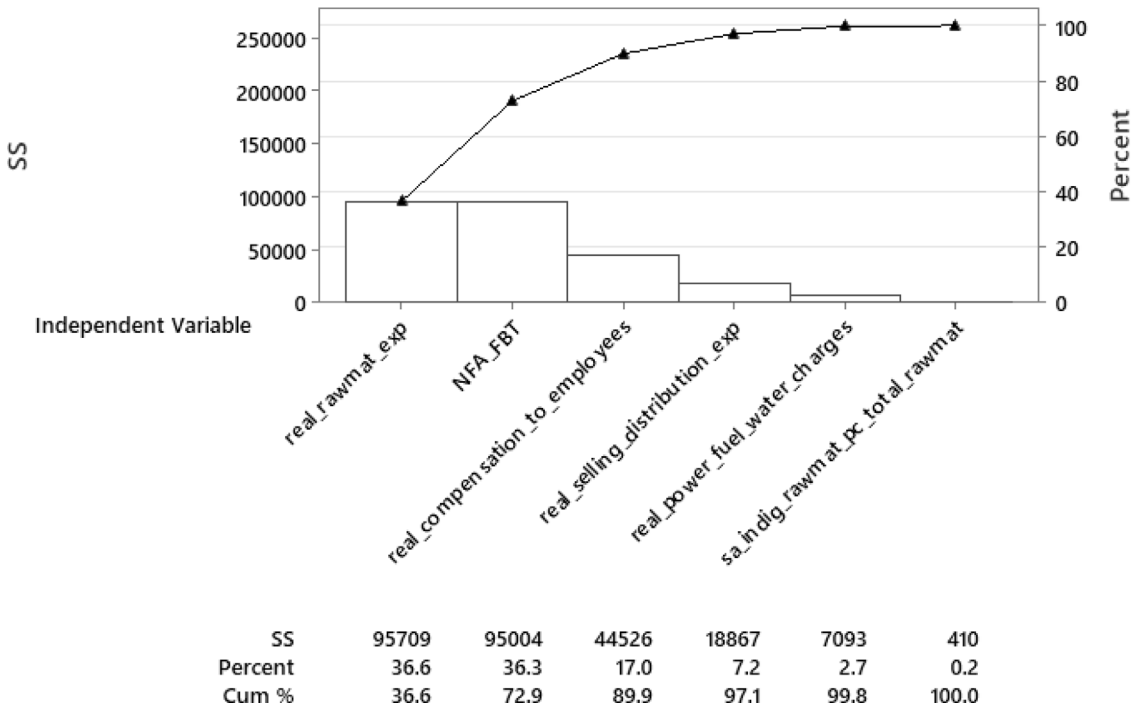
Table 3
$R^2$and RMSEP% for profits (real_pbdita) by algorithm.
Source: Authors' own calculations.

Category 1 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.483	10.35	0.129	9.54	0.046	11.99	0.965	9.37
2	0.574	9.78	0.420	8.36	0.562	9.96	0.915	8.82
3	0.631	8.42	0.543	8.42	0.332	10.92	0.933	8.42
4	0.680	5.65	0.726	4.90	0.705	5.22	0.511	6.20
5	0.477	7.47	0.584	6.93	0.193	13.71	0.882	6.75
6	0.737	7.89	0.762	6.14	0.756	6.24	0.703	6.65
7	0.533	6.67	0.614	5.95	0.688	6.23	0.885	5.37
8	0.600	6.34	0.532	6.47	0.319	7.65	0.713	5.45
9	0.579	8.84	0.507	7.20	0.338	10.41	0.613	5.89
10	0.534	6.12	0.387	8.80	0.308	10.64	0.855	5.90
Average:	**0.583**	**7.75**	**0.520**	**7.27**	**0.430**	**9.30**	**0.798**	**6.88**
SD	**0.083**	**1.59**	**0.181**	**1.47**	**0.246**	**2.81**	**0.153**	**1.46**

Category 2 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.397	12.15	0.143	31.43	0.004	54.90	0.908	5.92
2	0.170	20.75	0.019	44.89	0.278	18.50	0.838	8.31
3	0.532	21.76	0.424	19.42	0.305	21.46	0.963	7.53
4	0.627	14.44	0.509	18.53	0.387	20.99	0.519	14.90
5	0.613	18.17	0.430	19.04	0.490	18.19	0.924	8.05
6	0.728	10.24	0.708	11.35	0.627	12.63	0.795	9.26
7	0.606	18.31	0.724	10.79	0.364	20.74	0.920	8.10
8	0.726	5.72	0.750	6.40	0.401	12.51	0.656	6.67
9	0.684	13.78	0.544	16.53	0.360	20.71	0.734	9.69
10	0.709	10.70	0.629	12.63	0.337	19.94	0.918	6.52
Average:	**0.579**	**14.60**	**0.488**	**18.10**	**0.355**	**22.06**	**0.818**	**8.50**
SD	**0.176**	**5.11**	**0.246**	**10.54**	**0.159**	**12.00**	**0.143**	**2.54**

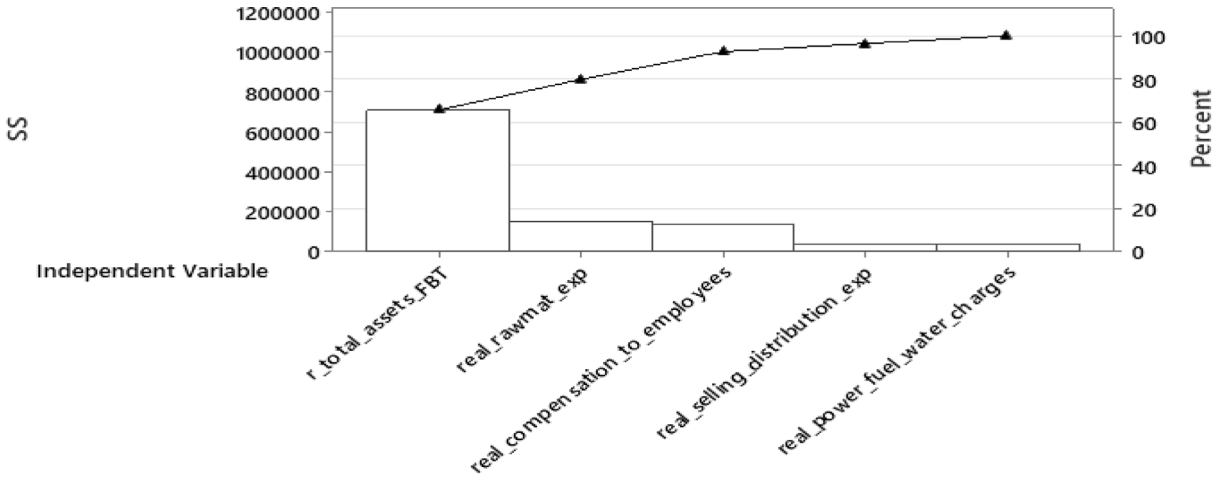
Category 3 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP%	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.540	22.47	0.177	29.91	0.005	54.96	0.870	8.52
2	0.534	23.10	0.481	24.63	0.560	22.27	0.813	8.97
3	0.721	17.89	0.650	19.25	0.333	29.49	0.970	7.86
4	0.667	18.35	0.669	18.81	0.754	17.79	0.618	18.94
5	0.453	28.00	0.682	19.02	0.190	42.43	0.839	9.37
6	0.724	18.83	0.681	19.07	0.754	18.23	0.733	18.94
7	0.555	22.63	0.660	19.45	0.687	19.15	0.876	9.14
8	0.611	20.04	0.661	19.71	0.317	29.68	0.473	30.91
9	0.642	17.88	0.448	24.89	0.336	29.84	0.715	18.29
10	0.556	22.41	0.515	23.17	0.306	30.16	0.892	8.39
Average:	**0.600**	**21.16**	**0.562**	**21.79**	**0.424**	**29.40**	**0.780**	**13.93**
SD	**0.088**	**3.20**	**0.162**	**3.74**	**0.253**	**11.74**	**0.148**	**7.62**

Category 4 Independent Variables								
	Bootstrap Forest		Boosted Tree		Regression Tree		BART	
Validation	R^2	RMSEP%	R^2	RMSEP %	R^2	RMSEP %	R^2	RMSEP %
1	0.392	26.19	0.348	26.35	0.005	54.51	0.916	10.15
2	0.388	26.28	0.280	27.28	0.363	27.36	0.875	11.20
3	0.612	20.04	0.528	21.53	0.268	28.27	0.963	10.40
4	0.600	20.86	0.573	21.19	0.657	19.66	0.300	27.45
5	0.699	18.56	0.488	28.93	0.621	20.01	0.799	15.02
6	0.746	18.08	0.551	21.95	0.579	25.42	0.603	7.49
7	0.550	25.63	0.420	26.72	0.720	16.49	0.869	11.27
8	0.635	18.88	0.529	22.03	0.715	16.73	0.454	23.91
9	0.604	20.91	0.551	22.26	0.489	22.35	0.723	16.62
10	0.734	18.96	0.691	18.44	0.772	15.96	0.945	10.97
Average:	**0.596**	**21.44**	**0.496**	**23.67**	**0.519**	**24.68**	**0.745**	**14.45**
SD	**0.125**	**3.31**	**0.119**	**3.38**	**0.242**	**11.40**	**0.225**	**6.50**



SS	95709	95004	44526	18867	7093	410
Percent	36.6	36.3	17.0	7.2	2.7	0.2
Cum %	36.6	72.9	89.9	97.1	99.8	100.0

Fig. 6. Sums of squares contribution for 'Category 1' independent variables for 'real_pbdita'.
Source: Authors' own creation.



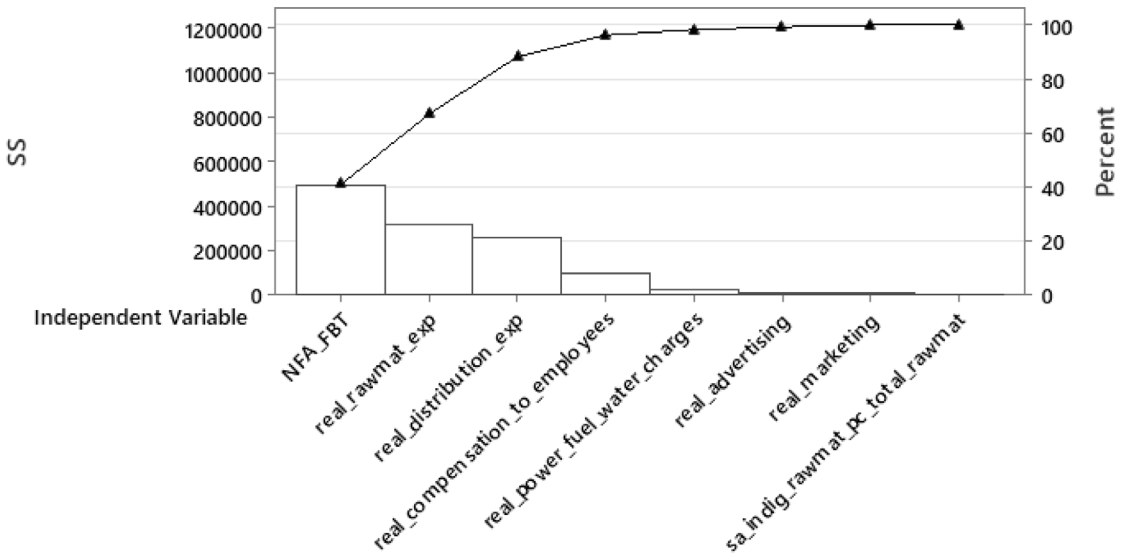
SS	707850	153073	138895	40367	38751
Percent	65.6	14.2	12.9	3.7	3.6
Cum %	65.6	79.8	92.7	96.4	100.0

Fig. 7. Sums of squares contribution for 'Category 2' independent variables for 'real_pbdita'.
Source: Authors' own creation.

products, such as ice-creams or milk and other milk-based products are mostly marketed as branded items. Sutton (1991) models expenditure on advertising as an upfront sunk cost in this industry and uses it to characterize the nature of market concentration and competition in various sub-sectors of the processed food industry. Branding is an expensive and time-consuming process. The current initiatives of some governments, like in India the recent Production Linked Incentive Scheme for Food Processing Industry (PLISFPI) provides assistance to brand-building for export markets is an indicator of the difficulties that small and new entrepreneurs face in creating awareness about branded sales as creating a brand in the international market is challenging (See Sutton's (1991) discussion on difficulty in establishing presence in the global market for even a large company like Unilever).

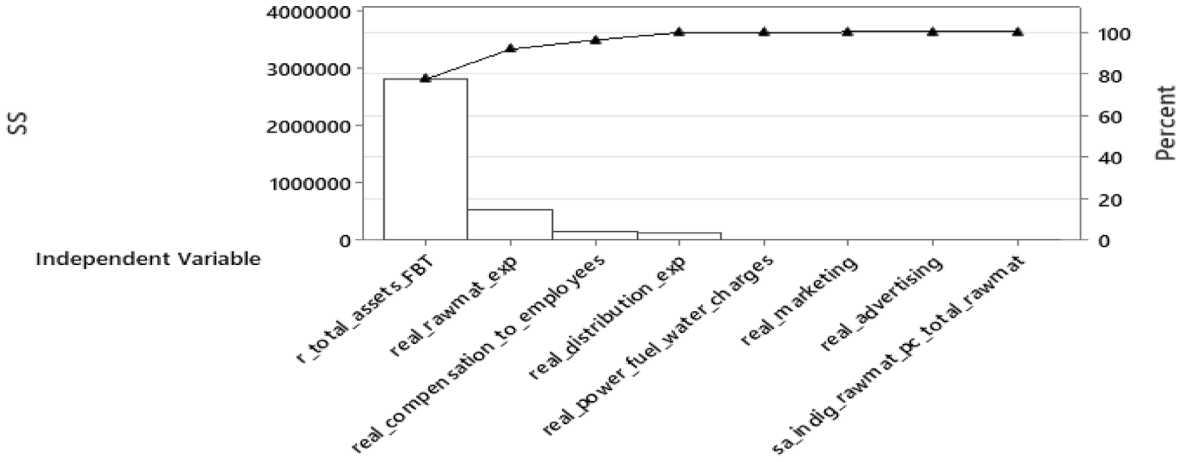
B *Industrial sales*: Firms can bypass the large cost of branded marketing by co- processing for other firms and engage in industrial sales, as noted by Sutton (1991) and Saha (2020). For small and mid-sized firms, this might work as an optimal strategy for survival (Saha, 2020).

C *Wholesale unbranded marketing*: Firms with multiple outputs might offload some of their products in the wholesale unbranded



SS	496618	317553	257467	97283	25051	12090	8360	189
Percent	40.9	26.1	21.2	8.0	2.1	1.0	0.7	0.0
Cum %	40.9	67.0	88.2	96.2	98.3	99.3	100.0	100.0

Fig. 8. Sums of squares contribution for 'Category 3' independent variables for 'real_pbdita'.
Source: Authors' own creation.



SS	2800499	523548	159676	131178	4669	3320	1563	1290
Percent	77.2	14.4	4.4	3.6	0.1	0.1	0.0	0.0
Cum %	77.2	91.7	96.1	99.7	99.8	99.9	100.0	100.0

Fig. 9. Sums of squares contribution for 'Category 4' independent variables for 'real_pbdita'.
Source: Authors' own creation.

segment. Gasoline sales for instance takes place in both the unbranded wholesale market as well as branded sales (Borenstein, 1991), as is the case of pharmaceuticals, which has both the generics segment as well as branded and patented molecules (Bhattacharyya, 2019). Though the margins in wholesale retailing might be minimal, it allows firms to dispose of some by-products, such as broken grains in the case of grain processing (such as in rice-milling), which are not the core revenue generator for the firm (Saha, 2020). Investing in the marketing of peripheral products is an unnecessary expenditure that many firms might desire to avoid and instead focus on the marketing of some star products (see Cooper & Kleinschmidt, 2000).

The implications on profits from these channels, for the same amount of sales, vary due to different price premia attached to these processes. Branded products ensure sellers a higher premium compared to industrial sales, and this in turn attracts a higher premium, in general, compared to wholesale unbranded marketing Saha (2020). As branding is a relatively more expensive exercise, small firms are likely to opt for industrial sales or unbranded marketing as opposed to branded marketing by larger firms. This is likely to lead to a divergence in profits for firms which adopt different marketing strategies for a given level of saleable output produced using standardized technology.

(2) *Profits are generated through financial investments apart from sales of output:* This occurs for firms which diversify their bottom lines beyond the space of produced outputs through financial investments. In

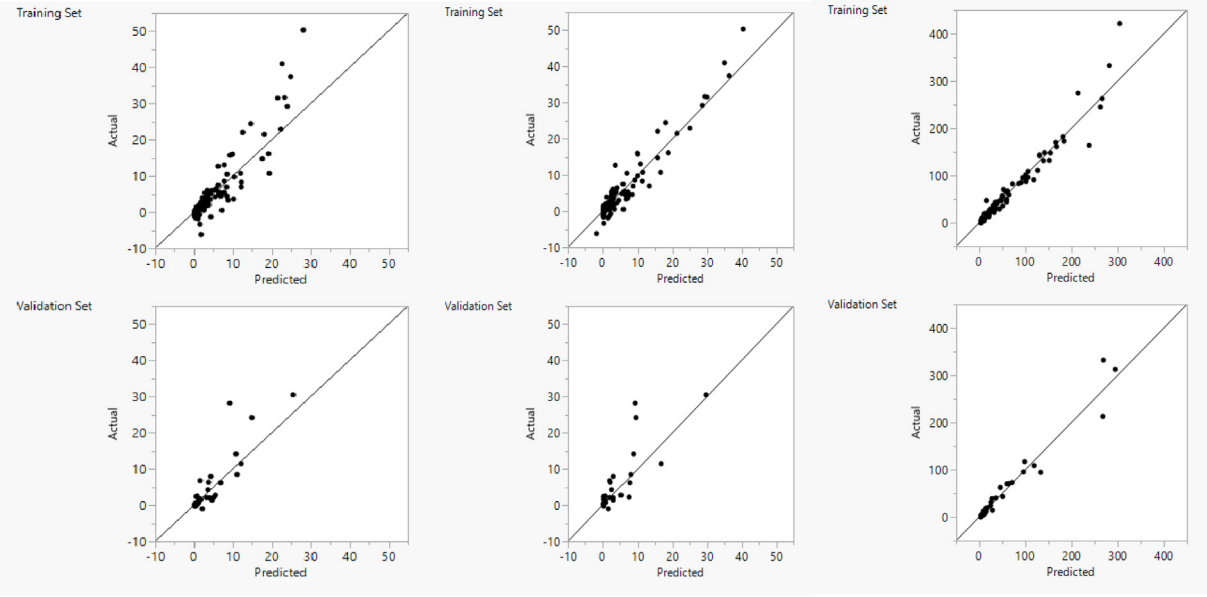


Fig. 10. Comparative predicted versus actual 'real_pbdita' for validation study 10 for 'bootstrap forest' (left), boosted tree (middle), and BART (right) for 'Category 4' independent variables.
Source: Authors' own creation.

this case, the tight correlation between profits and sales is likely to be lower than in the first case. In fact, in our data, the correlation between sales and profits are not high (it stands at 0.683). Even if there is some similarity among firms up to the stage of sales, the final stage leading to profits is divergent across firms potentially explaining why BART underperforms in predicting profits relative to sales.

Unlike the discussion for informal establishments as in De Mel et al. (2009), our discussion above provides a possible explanation regarding higher regularity, and therefore predictability, in sales data rather than profits for registered manufacturing units.

4. Conclusion

Estimating firm performance has a long tradition in economics and finance, due to its important for a number of dimensions: firm survival and firm exit (Balcaen, Manigart, Buyze, & Ooghe, 2012; Bhattacharjee, Higson, Holly, & Kattuman, 2009; Esteve-Pérez & Mañez-Castillejo, 2008) to mention a few), exit of entrepreneurs (see DeTienne, 2010; Wennberg, Wiklund, DeTienne, & Cardon, 2010), performance of teams (DeVaro, 2006), unionization in firms (Machin & Stewart, 1996), effectiveness of industrial policy, with particular reference to taxes and subsidies (Aghion et al., 2015; Rodrik, 2008), impact of public policy on performance of various firm sizes (Dvouletý, Blazkova, & Potluka, 2021) and other miscellaneous factors. It assumes a great deal of significance in developing countries that attempt to jump-start economic growth with industrialization.

There is a paucity of evidence regarding the prediction of financial performance of firms in registered manufacturing in developing countries. Given that these firms are the main indirect tax base for governments in developing countries, accurate prediction acquires an additional imperative other than industrial policy formulation. This study advances the prediction of firm performance using from a machine learning perspective. Note that most of our motivating examples are drawn from the food processing industry, as we intend to investigate the prediction of top line performance as opposed to bottom line performance for this industry for a developing country like India. This industry is important for reasons including and not limited to maintenance of food quality, preservation, food safety assurance (Augustin et al., 2016; Floros et al., 2010; Knorr & Augustin, 2021) providing off-farm employment and generating incomes (Saha, 2020). In India, gross value added by the food processing industry at constant prices 2011–12 for 2019–20 stands at INR 2.24 lakh crore (see https://www.mofpi.gov.in/sites/default/files/1._gvaconstant.pdf). The Government of India has kept a strong focus on this industry as a key driver of industrial growth. Understanding the financial performance of firms in this industry, therefore, has an immediate policy relevance.

Additionally, note that 59 per cent of firms in our sample (5863 firms out of 10,014 firms) have a total asset size less than 4 million INR, which is the 50th percentile in terms of size distribution of all firms. Our investigation shows that it is sales and not profits that is more predictable for this bulk of small firms in registered manufacturing in food processing in India. The dichotomy between the result of small unregistered units in the informal sector and that in the formal sector is highlighted by our analysis: the appropriate yardstick of measuring performance switches sharply from profits to sales as firms change their registration status in developing countries.

Machine learning predicts sales more accurately than profits in the food processing industry for registered manufacturing firms in India. In particular, the Bayesian machine learning algorithm 'BART' outperforms others in predicting performance. Our paper adds to the evolving literature integrating contemporary machine learning methods to deepen our understanding of predicting financial performance of firms.

We should expand the research in this area of firm performance prediction with more ML and DL algorithms if the appropriate datasets can be obtained. We hope to expand this research in the future to work with unstructured data such as scanned image data in text format This would expand the traditional approach by economists that work with data that are highly structured (e.g., cross sectional, time-series or panel).

CRediT authorship contribution statement

Debdatta Saha: Conceptualization, Visualization, Formal analysis, Writing – original draft, Supervision. **Timothy M. Young:** Methodology, Software, Investigation, Formal analysis, Validation, Writing – review & editing. **Jessica Thacker:** Data curation, Resources, Formal analysis Writing– review & editing.

Table A.1
Descriptive statistics of sales and profits (in million INR) for different percentiles for different firm classifications.
Source: Author's own calculations.

Firm Classification in terms of Total Assets			
Percentiles of Total Assets (value in million INR)	Sales (in million INR)	Profits (in million INR)	Correlation between sales and profits (r)
Bottom 1% Value of total assets < 0.005 (#firms=119)	N=15 Mean=1.837 SD=4.109	N=49 Mean=0 .002 SD=0.021	N=15 r=0.095
From 1% upto 5% Value of total assets=[0.005,0.097) (#firms=468)	N=156 Mean=0.095 SD=0.290	N=340 Mean=0.016 SD=0.311	N=156 r=−0.014
From 5% upto 10% Value of total assets=[0.097,0.267) (#firms=586)	N= 360 Mean=0.292 SD=0.561	N= 532 Mean=0.029 SD=0.222	N=360 r=−0.012
From 10% upto 25% Value of total assets=[0.267,1.029) (#firms=1758)	N= 1367 Mean=1.076 SD=1.497	N= 1656 Mean=0.057 SD=0.298	N=1367 r=0.060
From 25% upto 50% Value of total assets=[1.029,4.152) (#firms=2932)	N= 2627 Mean= 4.621 SD= 5.658	N= 2864 Mean=0.218 SD=0.375	N=2627 r=0.267
From 50% upto 75% Value of total assets=[4.152, 17.563) (#firms=2931)	N= 2717 Mean=17.201 SD=17.552	N= 2889 Mean=0.887 SD=1.390	N=2717 r=0.323
From 75% upto 90% Value of total assets=[17.563, 51.358) (#firms=1760)	N=1691 Mean= 40.224 SD=35.695	N= 1754 Mean=2.907 SD=3.867	N=1691 r=0.315
Above 90% Value of total assets above 51.358 (#firms=1172)	N= 1148 Mean=196.419 SD= 277.645	N= 1172 Mean=17.473 SD= 25.874	N=1148 r=0.573
Highest value of asset= 1489.652			

Firm Classification in terms of Net Fixed Assets			
Percentiles	Sales (in million INR)	Profits (in million INR)	Correlation between sales and profits (r)
Bottom 1% Value of assets<0.001 (#firms= 113)	N=47 Mean=0.350 SD= 0.830	N= 97 Mean=0.766 SD=6.910	N=47 r=0.3620
From 1% upto 5% Value of assets=[0.001,0.031) (#firms=447)	N=241 Mean=4.168 SD= 18.806	N=395 Mean=0.044 SD=0.520	N=241 r=−0.0631
From 5% upto 10% Value of assets=[0.031, 0.078) (#firms=561)	N=403 Mean=3.025 SD= 14.389	N= 518 Mean=0.107 SD= 1.095	N=403 r=0.441
From 10% upto 25% Value of assets=[0.078, 0.305) (#firms=1682)	N=1376 Mean= 3.561 SD=7.644	N= 1593 Mean= 0.139 SD=0.528	N=1376 0.335
From 25% upto 50% Value of assets=[0.305,1.239) (#firms= 2,803)	N= 2532 Mean=6.888 SD=11.708	N= 2731 Mean=0.282 SD=0.907	N=2532 r=0.350
From 50% upto 75% Value of assets=[1.239, 5.385) (#firms=2803)	N=2634 Mean=24.507 SD= 55.929	N= 2790 Mean=1.223 SD= 2.419	N=2634 r=0.524
From 75% upto 90% Value of assets=[5.385, 17.775) (#firms=1680)	N=1626 Mean=46.898 SD=66.251	N=1,679 Mean=3.446 SD= 5.848	N=1626 r=0.533
Above 90% Value of assets ≥ 17.77 (#firms=1122)	N=1115 Mean=278.280 SD=278.280	N=1122 Mean=16.511 SD= 26.365	N=1115 r=0.618
Highest value of assets=661.981			

N denotes number of observations available Value of assets is in INR million.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.We have not received any external source of funding for conducting the research published in this research article.

Data availability

Data will be made available on request.

Appendix

See Table A.1.

References

Aghion, P., Cai, J., Dewatripont, M., Du, L., Harrison, A., & Legros, P. (2015). Industrial policy and competition. *American Economic Journal: Macroeconomics*, *7*(4), 1–32. http://dx.doi.org/10.1257/mac.20120103.

Agyei-Mensah, B. K. (2010). Financial management practices of small firms in Ghana: An empirical study. http://dx.doi.org/10.2139/ssrn.1597243, Available at SSRN 1597243.

Augustin, M. A., Riley, M., Stockmann, R., Bennett, L., Kahl, A., Lockett, T., ... Cobiac, L. (2016). Role of food processing in food and nutrition security. *Trends in Food Science & Technology*, *56*, 115–125. http://dx.doi.org/10.1016/j.tifs.2016.08.005.

Back, B., Toivonen, J., Vanharanta, H., & Visa, A. (2001). Comparing numerical data and text information from annual reports using self-organizing maps. *International Journal of Accounting Information Systems*, *2*(4), 249–269. http://dx.doi.org/10.1016/S1467-0895(01)00018-5.

Balcaen, S., Manigart, S., Buyze, J., & Ooghe, H. (2012). Firm exit after distress: differentiating between bankruptcy, voluntary liquidation and M & A. *Small Business Economics*, *39*(4), 949–975. http://dx.doi.org/10.1007/s11187-011-9342-7.

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, *83*, 405–417. http://dx.doi.org/10.1016/j.eswa.2017.04.006.

Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2021). Supervised learning for the prediction of firm dynamics. In *Data science for economics and finance (19–41)*. Springer, Cham, http://dx.doi.org/10.1007/978-3-030-66891-4_2.

Basuchoudhary, A., Bang, J. T., & Sen, T. (2017). *Machine-learning techniques in economics: new tools for predicting economic growth*. New York: Springer, ISBN: 978-3-319-69014-8.

Bhattacharjee, A., Higson, C., Holly, S., & Kattuman, P. (2009). Macroeconomic instability and business exit: Determinants of failures and acquisitions of UK firms. *Economica*, *76*(301), 108–131. http://dx.doi.org/10.1111/j.1468-0335.2007.00662.x.

Bhattacharyya, N. C. (2019). Generic versus branded medicines. *International Journal of Health Research and Medico Legal Practice*, [ISSN: 2454-5139] *5*(1), 1–2, Electronic.

Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2017). Explaining machine learning models in sales predictions. *Expert Systems with Applications*, *71*, 416–428. http://dx.doi.org/10.1016/j.eswa.2016.11.010.

Borenstein, S. (1991). Selling costs and switching costs: Explaining retail gasoline margins. *Rand Journal of Economics*, *22*(3), 354–369. http://dx.doi.org/10.2307/2601052.

Bragoli, D., Ferretti, C., Ganugi, P., Marseguerra, G., Mezzogori, D., & Zammori, F. (2021). Machine-learning models for bankruptcy prediction: do industrial variables matter? *Spatial Economic Analysis*, 1–22. http://dx.doi.org/10.1080/17421772.2021.1977377.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. http://dx.doi.org/10.1007/BF00058655.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Chen, W. S., & Du, Y. K. (2009). Using neural networks and data mining techniques for the financial distress prediction model. *Expert Systems with Applications*, *36*(2), 4075–4086. http://dx.doi.org/10.1016/j.eswa.2008.03.020.

Chen, Y., Motiwalla, L., & Riaz Khan, M. (2004). Using super-efficiency DEA to evaluate financial performance of e-business initiative in the retail industry. *International Journal of Information Technology and Decision Making*, *3*(02), 337–351. http://dx.doi.org/10.1142/S0219622004001045.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning*, *48*(1), 299–320. http://dx.doi.org/10.1023/A:1013916107446.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. http://dx.doi.org/10.1214/09-AOAS285.

Cooper, R. G., & Kleinschmidt, E. J. (2000). 2 new product performance: what distinguishes the star products. *Australian Journal of Management*, *25*(1), 17–46. http://dx.doi.org/10.1177/031289620002500104.

De Mel, S., McKenzie, D. J., & Woodruff, C. (2009). Measuring microenterprise profits: Must we ask how the sausage is made? *Journal of Development Economics*, *88*(1), 19–31. http://dx.doi.org/10.1016/j.jdeveco.2008.01.007.

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: A decisiontree approach. *Expert Systems with Applications*, *40*(10), 3970–3983. http://dx.doi.org/10.1016/j.eswa.2013.01.012.

DeTienne, D. R. (2010). Entrepreneurial exit as a critical component of the entrepreneurial process: Theoretical development. *Journal of Business Venturing*, *25*(2), 203–215. http://dx.doi.org/10.1016/j.jbusvent.2008.05.004.

DeVaro, J. (2006). Teams, autonomy, and the financial performance of firms. *Industrial Relations: A Journal of Economy and Society*, *45*(2), 217–269. http://dx.doi.org/10.1111/j.1468-232X.2006.00425.x.

Dvouletý, O., Blazkova, I., & Potluka, O. (2021). Estimating the effects of public subsidies on the performance of supported enterprises across firm sizes. *Research Evaluation*, *1*(24), http://dx.doi.org/10.1093/reseval/rvab004.

Esteve-Pérez, S., & Mañez-Castillejo, J. A. (2008). The resource-based theory of the firm and firm survival. *Small Business Economics*, *30*(3), 231–249. http://dx.doi.org/10.1007/s11187-006-9011-4.

Floros, J. D., Newsome, R., Fisher, W., Barbosa-Cánovas, G. V., Chen, H., Dunne, C. P., ... Ziegler, G. R. (2010). Feeding the world today and tomorrow: the importance of food science and technology: an IFT scientific review. *Comprehensive Reviews in Food Science and Food Safety*, *9*(5), 572–599. http://dx.doi.org/10.1111/j.1541-4337.2010.00127.x.

Fodor, G. (2013). The ninth annual MLSP competition: first place. In *2013 IEEE international workshop on machine learning for signal processing* (pp. 1–2). http://dx.doi.org/10.1109/MLSP.2013.6661932.

Foroghi, D., & Monadjemi, A. (2011). Applying decision tree to predict bankruptcy. *Vol. 4*, In *2011 IEEE international conference on computer science and automation engineering* (pp. 165–169). IEEE, http://dx.doi.org/10.1109/CSAE.2011.5952826.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *118*, 9–1232. http://dx.doi.org/10.1214/aos/1013203451.

Friedman, J. H., & Meulman, J. J. (2003). Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, *22*(9), 1365–1381. http://dx.doi.org/10.1002/sim.1501.

Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *32*, Article 100577,

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. In *The elements of statistical learning* (pp. 337–387). New York, NY: Springer, http://dx.doi.org/10.1007/978-0-387-84858-7_10.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240.

Kloptchenko, A., Eklund, T., Back, B., Karlsson, J., Vanharanta, H., & Visa, A. (2002). Combining data and text mining techniques for analyzing financial reports. In *AMCIS 2002 Proceedings* (p. 4). http://aisel.aisnet.org/amcis2002/4.

Knorr, D., & Augustin, M. A. (2021). Food processing needs, advantages and misconceptions. *Trends in Food Science & Technology*, *108*, 103–110. http://dx.doi.org/10.1016/j.tifs.2020.11.026.

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. In *Applied predictive modeling* (pp. 61–92). New York, NY: Springer, http://dx.doi.org/10.1007/978-1-4614-6849-3_4.

Lam, M. (2004). Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision Support Systems*, *37*(4), 567–581. http://dx.doi.org/10.1016/S0167-9236(03)00088-5.

Lang, M. H., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting & Economics (JAE)*, *60*(2–3), 110–135. http://dx.doi.org/10.1016/j.jacceco.2015.09.002.

Lesser, C., & Moisé-Leeman, E. (2009). Informal cross-border trade and trade facilitation reform in Sub-Saharan Africa. http://dx.doi.org/10.1787/225770164564.

Li, F. (2010). The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102.

Loh, W. Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. http://dx.doi.org/10.1002/widm.8.

Machin, S., & Stewart, M. (1996). Trade unions and financial performance. *Oxford Economic Papers*, *48*(2), 213–241. http://dx.doi.org/10.1093/oxfordjournals.oep.a028566.

Manogna, R. L., & Mishra, A. K. (2021). Measuring financial performance of Indian manufacturing firms: application of decision tree algorithms. *Measuring Business Excellence*.

März, A., Klein, N., Kneib, T., & Musshoff, O. (2016). Analysing farmland rental rates using Bayesian geoadditive quantile regression. *43*, (4), (pp. 663–698).

Maseko, N., & Manyani, O. (2011). Accounting practices of SMEs in zimbabwe: An investigative study of record keeping for performance measurement (a case study of bindura). *Journal of Accounting and Taxation*, *3*(8), 158–161. http://dx.doi.org/10.5897/JAT11.022.

Mason, L., Baxter, J., Bartlett, P., & Frean, M. (1999). Boosting algorithms as gradient descent in function space. *Vol. 12*, In *Proc. NIPS* (pp. 512–518). https://rchgate.net/publication/2750572_Boosting_Algorithms_as_Gradient_Descent_in_Function_Space.

Matthias, S., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stat Journal*, *20*(1), 3–29. http://dx.doi.org/10.1177/1536867X20909688.

Mehta, S. (2022). *A beginner's guide to bayesian additive regression trees*. Developers Corner, (analyticsindiamag.com).

Miyakawa, D., Miyauchi, Y., & Perez, C. (2017). Forecasting firm performance with machine learning: Evidence from Japanese firm-level data. *RIETI discussion paper series 17-E-068*, https://www.rieti.go.jp/jp/publications/dp/17e068.pdf.

Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G. (2020). Corporate default forecasting with machine learning. *Expert Systems with Applications*, *161*, Article 113567. http://dx.doi.org/10.1016/j.eswa.2020.113567.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. *Mathematics*, *8*(10), 1799.

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition (154–168)*. Berlin, Heidelberg: Springer, http://dx.doi.org/10.1007/978-3-642-31537-4_13.

Ozcan, Y. A., & McCue, M. J. (1996). Development of a financial performance index for hospitals: DEA approach. *Journal of the Operational Research Society*, *47*, 18–26.

Pap, J., Mako, C., Illessy, M., Kis, N., & Mosavi, A. (2022). Modeling organizational performance with machine learning. *Journal of Open Innovation: Technology, Market, and Complexity*, *8*(177), http://dx.doi.org/10.3390/joitmc8040177.

Qiu, X. Y., Srinivasan, P., & Hu, Y. (2014). Supervised learning models to predict firm performance with annual reports: An empirical study. *Journal of the Association for Information Science and Technology*, *65*(2), 400–413. http://dx.doi.org/10.1002/asi.22983.

Richard, P. J., Devinney, T. M., Yip, G. S., & Johnson, G. (2009). Measuring organizational performance: Towards methodological best practice. *Journal of Management*, *35*(3), 718–804. http://dx.doi.org/10.1177/0149206308330560.

Ridgeway, G. (2007). Generalized boosted models: A guide to the gbm package. *Update*, *1*(1), 2007, https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

Robert Baum, J., & Wally, S. (2003). Strategic decision speed and firm performance. *Strategic Management Journal*, *24*(11), 1107–1129.

Rodrik, D. (2008). Normalizing industrial policy. *Commission on growth and development working paper No. 3*, Washington, DC: World Bank, © World Bank. http://hdl.handle.net/10986/28009.

Saha, D. (2020). *Economics of the food processing industry: lessons from Bihar, India*. Springer Nature, ISBN: 978-981-13-8554-4, Electronic.

Sehgal, S., Mishra, R. K., Deisting, F., & Vashisht, R. (2021). On the determinants and prediction of corporate financial distress in India. *Managerial Finance*.

Shrivastava, A., Kumar, N., Kumar, K., & Gupta, S. (2020). Corporate distress prediction using random forest and tree net for India. *Journal of Management and Science*, *1*(1), 1–11. http://dx.doi.org/10.26524/jms.2020.1.

Steiner, S., Zeng, Y., Young, T. M., Edwards, D. J., Guess, F. M., & Chen, C. H. (2016). A study of missing data imputation in predictive modeling of a wood-composite manufacturing process. *Journal of Quality Technology*, *48*(3), 284–296. http://dx.doi.org/10.1080/00224065.2016.11918167.

Sun, J., & Hui, X. F. (2006). An application of decision tree and genetic algorithms for financial ratios' dynamic selection and financial distress prediction. In *2006 international conference on machine learning and cybernetics* (pp. 2413–2418). IEEE, http://dx.doi.org/10.1109/ICMLC.2006.258771.

Sutton, J. (1991). *Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration*. MIT Press, ISBN: 9780262693585.

Triguero, I., Río, S. Del., López, V., Bacardit, J., Benítez, J. M., & Herrera, F. (2015). ROSEFW-RF: the winner algorithm for the ECBDL'14 big data competition: an extremely imbalanced big data bioinformatics problem. *Knowledge-Based Systems*, *87*, 69–79. http://dx.doi.org/10.1016/j.knosys.2015.05.027.

Venkatraman, N., & Ramanujam, V. (1986). Measurement of business performance in strategy research: A comparison of approaches. *Academy of Management Review*, *11*(4), 801–814. http://dx.doi.org/10.5465/amr.1986.4283976.

Wennberg, K., Wiklund, J., DeTienne, D. R., & Cardon, M. S. (2010). Reconceptualizing entrepreneurial exit: Divergent exit routes and their drivers. *Journal of Business Venturing*, *25*(4), 361–375. http://dx.doi.org/10.1016/j.jbusvent.2009.01.001.

Further reading

Athey, S. (2019). 21. The impact of machine learning on economics. In *The economics of artificial intelligence (pp. 507-552)*. University of Chicago Press, http://dx.doi.org/10.7208/9780226613475-023.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *11*, 685–725. http://dx.doi.org/10.1146/annurev-economics-080217-053433.

Dastidar, A. G., & Ahuja, R. (2019). A perspective on the slowdown in private corporate investments in India. In *Opportunities and challenges in development (31–50)*. Singapore: Springer, http://dx.doi.org/10.1007/978-981-13-9981-7_3.

De, P. K., & Nagaraj, P. (2014). Productivity and firm size in India. *Small Business Economics*, *42*(4), 891–907. http://dx.doi.org/10.1007/s11187-013-9504-x.

Faello, J. (2015). Understanding the limitations of financial ratios. *Academy of Accounting and Financial Studies Journal*, [ISSN: 1528-2635] *19*(3), 75, Online.

Gentry, R. J., & Shen, W. (2010). The relationship between accounting and market measures of firm financial performance: How strong is it? *Journal of Managerial Issues*, 514–530, https://www.jstor.org/stable/25822528.

Hammoudi, A., Hoffmann, R., & Surry, Y. (2009). Food safety standards and agri-food supply chains: an introductory overview. *European Review of Agricultural Economics*, *36*(4), 469–478. http://dx.doi.org/10.1093/erae/jbp044.

Hasan, R., & Jandoc, K. R. (2010). The distribution of firm size in India: what can survey data tell us?. *Asian development bank economics working paper series, paper No. 213*, http://dx.doi.org/10.2139/ssrn.1681268.

Henson, S., & Reardon, T. (2005). Private agri-food standards: Implications for food policy and theagri-food system. *Food Policy*, *30*(3), 241–253. http://dx.doi.org/10.1016/j.foodpol.2005.05.002.

Jones, G., & Miskell, P. (2007). Acquisitions and firm growth: Creating unilever's ice cream and tea business. *Business History*, *49*(1), 8–28. http://dx.doi.org/10.1080/00076790601062974.

Majumdar, S. K. (1997). The impact of size and age on firm-level performance: some evidence from India. *Review of Industrial Organization*, *12*(2), 231–241. http://dx.doi.org/10.1023/A:1007766324749.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. http://dx.doi.org/10.1257/jep.31.2.87.

Reddy, K., & Sasidharan, S. (2020). Driving small and medium-sized enterprise participation in GlobalValue chains: Evidence from India. *ADB working paper No. 1118*, https://www.adb.org/publications/driving-sme-participation-global-value-chains-evidence-india.

Sen, S., & Dasgupta, Z. (2018). Financialisation and corporate investments: the Indian case. *Review of Keynesian Economics*, *6*(1), 96–113. http://dx.doi.org/10.4337/roke.2018.01.06.

Shanmugam, K. R., & Bhaduri, S. N. (2002). Size, age and firm growth in the Indian manufacturing sector. *Applied Economics Letters*, *9*(9), 607–613. http://dx.doi.org/10.1080/13504850110112035.