A DEA and Random Forest Regression Approach to Studying Bank Efficiency and Corporate Governance

Keyur Thaker, Vincent Charles, Abhay Pant, Tatiana Gherman

Abstract

We employ Data Envelopment Analysis to estimate the new technical, new cost, and new profit efficiency of Indian banks over the period 2008-2018. Then, we use Random Forest Regression to examine the impact of corporate governance (Board Size, Board Independence, Duality, Gender Diversity, and Board Meetings), bank characteristics (Return on Assets, Size, and Equity to Total Assets), and other characteristics (Ownership and Years) on bank efficiency. Among others, we found that board characteristics play a significant role particularly in new profit efficiency; therefore, policymakers and regulators should consider Board Size, Board Independence, Board Meetings, and Duality while framing guidelines for enhancing bank new profit efficiency. We also found that Board Independence plays a vital role in bank new cost efficiency, while Gender Diversity contributes to both new technical and new cost efficiency. This study makes methodological contributions by employing Machine Learning based Random Forest Regression in tandem with Data Envelopment Analysis under a two-phase model to examine corporate governance and bank efficiency, which is a pioneering attempt.

Keywords: Machine learning, data science, data envelopment analysis, performance, bank efficiency, corporate governance.

1. Introduction

The global financial crisis of 2007-2009 was a crucial event that called attention upon the governance of the commercial banks, which has often been considered among the main causes of the crisis (De Haan & Vlahu, 2016). The Basel Committee on Banking Supervision (Basel Committee, 2015) called for effective governance practices to maintain public trust and confidence in the banking systems. In India, the Reserve Bank of India (RBI) appointed a committee, led by P. J. Nayak, to review the governance of boards of banks. The committee's May 2014 report (Nayak et al., 2014) formed the basis for the set of rules for Bank Governance in India. The Nayak Committee suggested a need for moving away from multiple regulations of banks, and a more substantial role of the board in the bank governance. The committee recommended the constitution of a stronger bank board and management, with considerable autonomy in the banking sector.

Bank efficiency is an important indicator of bank performance and managerial efficiency and provides useful information for policy decisions, including for the ones related to bank governance. In time, the Indian banking sector has experienced several transformations, such as enhanced competition from new private banks, deregulation, technological advancements, large-scale financial inclusion, and increased regulatory capital requirements, which have shaped the sector in a particular way. Corporate governance represents a set of rules or laws by which the management of organisations are controlled or directed (Kamarudin et al., 2020). Corporate governance in banks is a micro dimension of governance that focuses on governance within banks (firm-level governance). It is different from country-level governance, which is a macro dimension of governance that deals with governance within the country (Kamarudin et al., 2020). Kraay et al. (2010) identified voice and accountability, political stability and absence of violence, government effectiveness, regulatory capital, rule of law and control of corruption as important dimensions of country-level governance. While some of the studies in the recent past studied the relationship between country-level governance and bank efficiency and productivity (Hussain et al., 2020; Kamarudin et al., 2016; Kamarudin et al., 2020), the main focus of this study is on the impact of corporate governance on bank efficiency in India. The study uses board structure for corporate governance as board structure is a very important dimension of corporate governance (Pant Radhakrishnan, 2019) in banks that significantly influences banks' performance. It was further noted that most of the studies on bank governance have been conducted in the context of developed countries. Despite theoretical and empirical linkages, in India, to the best of the authors' knowledge, Narwal and Pathneja's (2016) study is the only known study on corporate governance and bank productivity. Further, the present study makes methodological contributions by employing a hybrid machine learning based random forest (RF) - data envelopment analysis (DEA) methodology to examine the impact of corporate governance on bank efficiency.

The Indian banking sector is interesting to study from many points of view. First, it is considered as one of the best regulated banking sectors, partly due to which the Indian banks survived the global financial crisis without major issues. Second, as mentioned previously, the Indian banking sector has undergone changes, such as banking reforms, heightened competition, increased regulatory capital, innovation, financial inclusion, rapid technology adoption, the revamping of bank governance, growth, and asset quality crisis. The advent of new private banks from the 1990s has challenged the dominance of the public sector banks and underlined the need for professionalising management and bank governance. Third, India has emerged as the fastest growing economy in the world, and its banking system is set to become the fourth largest by the end of 2020 (KPMG International, 2017). Fourth, the Government owns over 50% of the stake in public sector banks, holding the majority voting rights and the right to interfere in the functioning of the board. Apart from the RBI, being Government-owned, public sector banks have to comply with multiple regulators and supervisors, such as the Central Vigilance Commissioner, Right To Information, Controller & Auditor General and the Ministry of Finance, which severely hampers the board's autonomy and effectiveness.

To address the gap identified, we examine bank governance in India in terms of board characteristics and bank efficiency. We employ DEA in the first phase, and to this aim we consider three measures of bank efficiency, namely new technical (NTE), new cost (NCE), and new profit (NPE) efficiency, defined in line with Tone (2002), for a more complete and comprehensive assessment of banks (when compared to traditional technical - TE, cost - CE, and profit - PE efficiency, respectively). These measures reflect the banks' ability to efficiently convert inputs into outputs, minimise the cost, and maximise the revenues, which Berger and Mester (1997) considered as the essential characteristics of the economic efficiency concept.

In the second phase, the relationship between corporate governance (Board Size, Board Independence, Duality, Gender Diversity, and Board Meetings), bank characteristics (Return on Assets, Size, and Equity to Total Assets), and other characteristics (Ownership and Years) and bank efficiency is examined. More specifically, the relationship among Board Size, Board Independence, Duality, Gender Diversity, and Board Meetings and the three measures of bank efficiency is estimated. We employ conventional regression, OLS Bootstrap and Fractional Logit Bootstrap, and a methodologically superior and novel Random Forest (RF) regression to explain the relationship.

Overall, our study makes several contributions: First, as previously mentioned, we make methodological contributions by employing a machine learning based RF regression technique in tandem with DEA to examine the impact of corporate governance, bank characteristics, and other characteristics on bank efficiency, which is a pioneering attempt. RF has advantages over other conventional methods, such as OLS and its variants. RF has better prediction power, with a lower mean square error (MSE) when compared to variants of OLS. Further, the RF algorithm has no distribution data scale requirements, can overcome multicollinearity, is also tolerant of outliers and noise, and is less likely to have over-fitting issues. Moreover, this method is expected to have a lower computation requirement and better prediction power. Second, there are hardly any studies that explore the linkages between corporate governance and bank efficiency in the Indian context and, to the best of our knowledge, this is the first study that simultaneously examines NTE, NCE, and NPE of banks across state- and privately-owned banks in India. Third, prior studies are limited not just in scope, but also in period covered, whereas this study considers all three types of efficiency over a longer time period of 11 years.

The remainder of the paper is organised as follows. An overview of the Indian banking sector, along with a review of the relevant literature, is given in Section 2. The methodology, encompassing DEA and RF, is covered in Section 3. Section 4 is dedicated to data processing, the choice of inputs and outputs, as well as the determinants of bank efficiency. Data analysis is provided in Section 5. The results are discussed in Section 6, while managerial implications are given in Section 7. Finally, Section 8 concludes the paper.

2. Literature Review

2.1. Structure of the Indian Banking System

With an approximate size of INR 141 trillion (equivalent to US $2 trillion, in the year 2017), the Indian banking sector is the fifth largest banking system in the world and set to become the fourth largest by the end of 2020 and the third largest by 2025 (KPMG International, 2017). The RBI, as the central bank of the country, regulates bank and monetary policies. Banks in India are categorised into three ownership groups, namely public or state-owned, private, and foreign. SBI along with six associates are acknowledged as a separate group of Scheduled Banks due to well-defined statutes (1955 SBI Act and the 1959 SBI Subsidiary Banks Act).

All the state-owned banks, including the SBI, form the public sector banks, with over 50% government ownership and control and around 70% of the total credit and deposits from businesses. By 2018, there were 21 public sector banks, 22 private sector banks, and about 44 foreign banks (except for 4 or 5, most foreign banks had limited presence and representative offices). Collectively, all the scheduled commercial banks had over 143,848 functioning offices, 223,318 ATMs, and over 1.6 bn deposit accounts, making the sector the largest banking sector in the world. Private and foreign banks have technology-intensive operations, require less labour, and operate primarily in urban centres. On the other hand, public-sector banks own the major branch network, with a large number of rural branches and priority-sector obligations; and are a major employer.

2.2. Corporate Governance and Bank Efficiency

The dynamics of corporate governance can be seen through the lens of the agency theory and the principal agent problem, where the divergence between the interests of owners and managers may lead to agency conflicts (Jensen & Meckling, 1976). Fama and Jensen (1983) pointed out that the separation of management from control leads to efficient monitoring. A large body of the existing literature highlighted the importance of board structures in banks. Some studies highlighted the positive impact of board size on firm performance (Adams & Mehran, 2005; Adeabah et al., 2019), while others found a negative impact of board size on firm performance (Boitan & Nitescu, 2019; Liang et al., 2013). Further, De Andres and Vallelado (2008) found an inverted U-shaped relationship between board size and performance. A larger board is expected to better supervise managers and bring more human capital to advise them. However, a larger board size may lead to free riding, larger coordination effort, excessive power to CEO, which can in turn harm bank efficiency (Battaglia & Gallo, 2015).

Board independence is a critical aspect of board quality, as it may improve board monitoring and advising capability. Fama and Jensen (1983) argued that outside directors are better because they are concerned with their reputation as experts. The existing literature highlights positive, negative, or no impact of board independence on firm performance. Board independence is associated with superior performance (Battaglia & Gallo, 2015; Liu et al., 2015), promoting managerial efficiency (Boitan & Nitescu, 2019), positive excess returns (Rosenstein & Wyatt, 1990), and earnings quality (Cornett et al., 2009). Stewardship theory, on the other hand, argues that insiders or executive directors have better firm-specific knowledge than independent directors and, hence, are better at monitoring than independent directors (Donaldson & Davis, 1994). Lastly, some studies found no impact of board independence on performance (Adams & Mehran, 2005; Francis et al., 2012).

CEO duality, defined as a situation where the same individual holds the positions of both CEO and chairman of the board, has also been studied in relation to its impact on bank performance. In this sense, mixed results were obtained, with some studies having found a positive relationship between CEO duality and bank performance (Pathan, 2009), while other studies reported a negative or insignificant relationship between the two (Liang et al., 2013). The number of board meetings can be an indicator of the level of activity of board members and may lead to better bank and stock performance (Francis et al., 2012). Board meetings are an important mechanism that provides board members with a chance to come together and discuss and exchange ideas for better monitoring and strategy (De Andres & Vallelado, 2008). Finally, with respect to Gender Diversity, Boitan and Nitescu (2019) found that female board members related negatively with bank managerial efficiency; while Adeabah et al. (2019) argued in favour of the positive impact of having a maximum of two female directors, with an inverted U-shaped relation suggesting a threshold effect on bank efficiency.

2.3. DEA Studies

While most studies focused on estimating the TE or CE of Indian banks, relatively fewer studies measured PE, such as the studies by Das et al. (2005) and Ray and Das (2010). In this sense, Das et al. (2005) studied the efficiency of Indian banks using DEA and found that the median efficiency scores of banks improved in the post-reform period. Ray and Das (2010), on the other hand, conducted a study on CE and PE of Indian banks using DEA in the post-reform period and found that state-owned banks were more efficient than the private banks.

TE is relevant in a non-market environment, where the prices of inputs and outputs are not reliable or available. Unlike CE, PE is more informative (Ray & Das, 2010), as it requires the prices of both inputs and outputs. Generally, two approaches, the non-parametric DEA and the parametric stochastic frontier analysis (SFA), have been used to measure bank efficiency in the existing literature. In what concerns DEA, a vast body of literature is available in this sense on bank efficiency, such as the studies by Fethi and Pasiouras (2010), Sufian and Kamarudin (2014), Charles et al. (2016), Gulati and Kumar (2016), Kamarudin et al. (2016), Sathye and Sathye (2017), Charles et al. (2018), Nair and Vinod (2019), Hussain et al. (2020), Kamarudin et al. (2020), Saw et al. (2020), and Tsolas et al. (2020), among others.

Fethi and Pasiouras (2010) reviewed bank performance studies during the period 1998 to 2009 and highlighted that most of the studies on bank performance employed an input-oriented approach for efficiency as banks have greater control on their inputs relative to outputs. Sufian and Kamarudin (2014) employed DEA-based semi-parametric Malmquist index for productivity of banks in Malaysia. The empirical findings of the study indicated that productivity improvements in the Malaysian banking system were a result of technological progress. Charles et al. (2016) performed an efficiency analysis of public banks in Argentina and found the years 1999 and 2002 to be the black years for the public banking system in the country. Gulati and Kumar (2016) employed a DEA-based metafrontier approach to estimate the PE of Indian banks and found domestic banks to be less profit efficient relative to foreign banks. Kamarudin et al. (2016) investigated the impact of country-level governance on the revenue efficiency of Islamic and conventional banks in GCC countries. The findings of the study revealed that the rule of law, government effectiveness, and voice and accountability significantly influenced revenue efficiency in these banks. Sathye and Sathye (2017) investigated the impact of ATM intensity on the banks' TE in India and found ATM intensity to be negatively associated with bank efficiency.

More recently, Charles et al. (2018) proposed a new satisficing DEA model and used it to estimate the efficiency of banks in Peru. Nair and Vinod (2019) examined the determinants of CE, allocative, and scope efficiencies of Indian commercial banks. The results showed size to be a significant factor influencing the efficiency of banks. Hussain et al. (2020) studied the price efficiency of Islamic and conventional banks in 19 countries, over the period 2006 to 2017; the findings of the study indicated that voice and accountability positively impacted revenue efficiency whereas political stability, absence of violence and control of corruption had a negative impact. Kamarudin et al. (2020) investigated the impact of country-level governance on productivity of Islamic and conventional banks in Malaysia; the findings of the study revealed that most of the country-based governance factors significantly determined bank productivity for both sets of banks. Saw et al. (2020) examined CE, PE, and revenue efficiency of Islamic and conventional banks for 18 countries. The empirical results of the study showed no significant difference between Islamic and conventional banks. Lastly, Tsolas et al. (2020) developed a novel two-stage hybrid DEA-ANN based model to study the bank branch performance of a large Greek bank.

While studies on bank efficiency are numerous, to the best of our knowledge, none had a comprehensive focus on efficiency and examined multiple efficiencies, as we do in this study; nor did they investigate the relationship with corporate governance variables, even less so in India. This positions the present study as a significant contribution to the literature.

3. Methodology

3.1. Data Envelopment Analysis

Let $\mathbf{X} = (\mathbf{x_j}) \in \mathbf{R}^{m \times n}$ and $\mathbf{Y} = (\mathbf{y_j}) \in \mathbf{R}^{s \times n}$ represent input and output vectors. The production possibility set for the variable returns-to-scale (VRS) can be defined as: $PPS_{VRS} = \{(\mathbf{x}, \mathbf{y}) | \mathbf{x} \geq \mathbf{X}\lambda, \mathbf{y} \leq \mathbf{Y}\lambda, \mathbf{e}\lambda = \mathbf{1}, \lambda \geq \mathbf{0}\}$, where $\mathbf{e}$ is a row vector with all elements as unity and λ is a column vector with all non-negative elements. $\mathbf{e}\lambda = \mathbf{1}$ takes into consideration the variable returns-to-scale and together with $\lambda \geq \mathbf{0}$, imposes a convexity constraint and indicates the permissible ways in which the observations for various DMUs can be combined. The input-oriented TE can be obtained through the following linear programming, wherein θ is the measure of TE:

$$
\begin{aligned}
&[\mathrm{TE}_{VRS}] \ \min \theta \\
&\text{subject to} \\
&\theta \mathbf{x_0} - \mathbf{X}\lambda \geq \mathbf{0}, \\
&\mathbf{Y}\lambda - \mathbf{y_0} \geq \mathbf{0}, \\
&\mathbf{e}\lambda = \mathbf{1}, \\
&\lambda \geq \mathbf{0},
\end{aligned}
\tag{1}
$$

where "$\mathbf{o}$" represents the DMU of interest, and the TE θ ranges from zero to one. When $\theta = \mathbf{1}$ and the above system has no slacks, then $\mathrm{DMU}_{\mathbf{o}}$ is technically efficient; otherwise, it is deemed to be inefficient. Tone (2002) showed that the measurement of efficiency carried out by means of the traditional methods has shortcomings. To address this issue, he proposed a variant of TE, known as new technical efficiency (NTE), which incorporates the unit input cost $\mathbf{C} = (\mathbf{c_j})$, which can be defined as follows for VRS:

$$
\begin{aligned}
&[\mathrm{NTE}_{VRS}] \ \min \hat{\theta} \\
&\text{subject to} \\
&\hat{\theta} \hat{\mathbf{x}}_{\mathbf{0}} - \hat{\mathbf{X}}\lambda \geq \mathbf{0}, \\
&\mathbf{Y}\lambda - \mathbf{y_0} \geq \mathbf{0}, \\
&\mathbf{e}\lambda = \mathbf{1}, \\
&\lambda \geq \mathbf{0},
\end{aligned}
\tag{2}
$$

where $\hat{\mathbf{X}} = (\hat{\mathbf{x}}_{\mathbf{j}}) \in \mathbf{R}^{m \times n}$, $\hat{\mathbf{x}}_{\mathbf{j}} = (\mathbf{c_j x_j})^{\mathbf{T}}$, $\hat{\theta}$ is the NTE obtained based on the cost-based production possibility set, $PPS^{c}_{VRS} = \{(\hat{\mathbf{x}}, \mathbf{y}) | \hat{\mathbf{x}} \geqq \hat{\mathbf{X}}\lambda, \mathbf{y} \leqq \mathbf{Y}\lambda, \mathbf{e}\lambda = \mathbf{1}, \lambda \geqq \mathbf{0}\}$.

Similarly, the new cost efficiency (NCE), $\hat{\delta} = \frac{\mathbf{e}\hat{\mathbf{x}}^{*}_{\mathbf{0}}}{\mathbf{e}\hat{\mathbf{x}}_{\mathbf{0}}}$, can be obtained from the optimal solution, $\hat{\mathbf{x}}_{\mathbf{0}}$, using the following model:

$$
\begin{aligned}
&[\text{NCE}_{VRS}] \ \min \ \mathbf{e}\hat{\mathbf{x}} \\
&\text{subject to} \\
&\hat{\mathbf{x}} - \hat{\mathbf{X}}\lambda \geqq \mathbf{0}, \\
&\mathbf{Y}\lambda - \mathbf{y_0} \geqq \mathbf{0}, \\
&\mathbf{e}\lambda = \mathbf{1}, \\
&\lambda \geqq \mathbf{0}.
\end{aligned}
\tag{3}
$$

Note that Model (3) has m input constraints which could be further aggregated as a single constraint in view of the objective function form, $\mathbf{e}\hat{\mathbf{x}}$; hence, Model (3) could be re-written as follows:

$$
\begin{aligned}
&[\text{NCE'}_{VRS}] \ \min \ \mathbf{e}\hat{\mathbf{x}} \\
&\text{subject to} \\
&\mathbf{e}\hat{\mathbf{x}} - \mathbf{e}\hat{\mathbf{X}}\lambda \geqq \mathbf{0}, \\
&\mathbf{Y}\lambda - \mathbf{y_0} \geqq \mathbf{0}, \\
&\mathbf{e}\lambda = \mathbf{1}, \\
&\lambda \geqq \mathbf{0}.
\end{aligned}
\tag{4}
$$

Lastly, one can further define the new profit efficiency (NPE), $\hat{\eta} = \frac{\mathbf{e}\hat{\mathbf{y}}_{\mathbf{0}} - \mathbf{e}\hat{\mathbf{x}}_{\mathbf{0}}}{\mathbf{e}\hat{\mathbf{y}}^{*}_{\mathbf{0}} - \mathbf{e}\hat{\mathbf{x}}^{*}_{\mathbf{0}}}$, which can be obtained from the optimal solution, $(\hat{\mathbf{x}}_{\mathbf{0}}, \hat{\mathbf{y}}_{\mathbf{0}})$, using the following model:

$$
\begin{aligned}
&[\mathrm{NPE}_{VRS}]\ \max\ \mathbf{e}\hat{\mathbf{y}} \quad \mathbf{e}\hat{\mathbf{x}} \\
&\text{subject to} \\
&\hat{\mathbf{x}_0} \quad \hat{\mathbf{x}} = \mathbf{0}, \\
&\hat{\mathbf{y}} \quad \hat{\mathbf{y}_0} = \mathbf{0}, \\
&\mathbf{e}\lambda = \mathbf{1}, \\
&\lambda = \mathbf{0},
\end{aligned}
\tag{5}
$$

where $\hat{\mathbf{x}} = \hat{\mathbf{X}}\lambda$ and $\hat{\mathbf{y}} = \hat{\mathbf{Y}}\lambda$.

In this section, we have defined the three measures of bank efficiency (NTE, NCE, and NPE) and associated models, which will be used in the first phase to derive the efficiency scores. In the next section, we will proceed to introduce the Random Forest regression approach that will be employed in the second phase to examine the relationship between the three types of bank efficiency and corporate governance, bank characteristics, and other characteristics.

3.2. Random Forest

The concept of Classification and Regression Trees (CART), proposed by Breiman et al. (1984), was instrumental in the development of the RF approach. Random Forest, proposed by Breiman (2001), is a machine learning-driven ensemble method based on the CART decision tree and bootstrapping aggregation method. It combines the Bagging integrated learning theory with the random sub-space method (Ho, 1998), wherein Bagging (Breiman, 1996) uses an ensemble of unpruned CART trees constructed from bootstrap samples of the data. In other words, RF combines a number of trees by taking the same number of bootstrap samples from the original data and building a tree based on each bootstrap sample. As mentioned, in RF, the individual trees are not pruned. The selection of variables for each split is based only on a random subset of predictor variables. Each tree is built on an independently extracted sample and it has the same distribution in the forest. Individual trees form what is known as a forest. From the complete forest, the response variable is predicted as an average (in the case of regression) or majority vote (in the case of classification) of the predictions of all trees (Genuer et al., 2017; Hallett et al., 2014; Scornet et al., 2015).

One can construct a regressor from any training set which is drawn using bootstrap sampling from the entire data set T. The data left out from the random selection process are referred to as Out-of-Bag (OOB) samples; otherwise stated, OOB represents the error rate for left-out observations. To be noted that this method performs a cross-validation in parallel using OOB samples. Let T_N be the training set of independent and identically distributed random vectors $(\mathbf{D}, \mathbf{Z})$ containing N examples (also known as instances, patterns), which can be expressed notationally as $T_N = \{(\mathbf{d_1}, \mathbf{z_1}), (\mathbf{d_2}, \mathbf{z_2}), ..., (\mathbf{d_N}, \mathbf{z_N})\}$, where $\mathbf{D} = (D^1, D^2, ..., D^p) \in \mathbf{R}^p$ represents the features (explanatory variables), p is the number of features, $\mathbf{Z} \in \mathbf{R}$ is the target (dependent variable), and $\mathbf{R}$ represents the real number set. For the total number of features p, an integer $mtry$ (representing the number of candidate features randomly selected to split in each non-leaf node) is fixed. At each node, the best feature at each $mtry$ is picked and the node is split into two child nodes. For regression, it is sought to minimise the variance at each node. Technically, in a regression-based prediction problem, an RF is a regression consisting of a collection of numerical tree predictors $\{Q(\mathbf{d}, \Theta_l), l = 1, 2, ..., N_{tree}\}$, where $\{\Theta_l\}$ are independent identically distributed random vectors drawn from the distribution of random vectors $\mathbf{Z}$ and $\mathbf{D}$, and N_{tree} is the number of trees in the forest. The RF numerical predictor is obtained by $\frac{1}{N_{tree}} \sum_{l=1}^{N_{tree}} Q(\mathbf{d}, \Theta_l)$, with the mean-squared error (MSE) for any numerical predictor as $E_{\mathbf{D},\mathbf{Z}}(\mathbf{Z} - Q(\mathbf{D}))^2$, and the root mean squared error (RMSE) as $\sqrt{E_{\mathbf{D},\mathbf{Z}}(\mathbf{Z} - Q(\mathbf{D}))^2}$. The pseudo $R^2 = 1 - MSE/E_{\mathbf{Z}}(\mathbf{Z} - E_{\mathbf{Z}}(\mathbf{Z}))^2$, is a measure of goodness of fit, which highlights how closely the predicted values of a model match the observed values. The ideal value of R-squared is one which shows that the model can explain all the variability in the dependent variable. MSE, $RMSE$, and R^2 are metrics commonly used for the evaluation of the model. Figure 1 illustrates the flow of the RF regression construction.

RF has several advantages (Hallett et al., 2014). First, the implementation of trees is easy to do due to automation. Second, it is data adaptive and practically model assumption free. Third, unlike with traditional regression models, RF removes concerns regarding the correct modelling of associations between predictors and the outcome, and whether or not non-linear effects or higher order interactions for predictors should be incorporated. Fourth, RF is also known for its high prediction ability. Despite these attractive features, however, RF also has some disadvantages, among which we can mention its not-so-straightforward interpretability when compared to an individual tree; this means that it is not very clear which variables are the most important. As Hallett et al. (2014, p. 527) noted, "variable

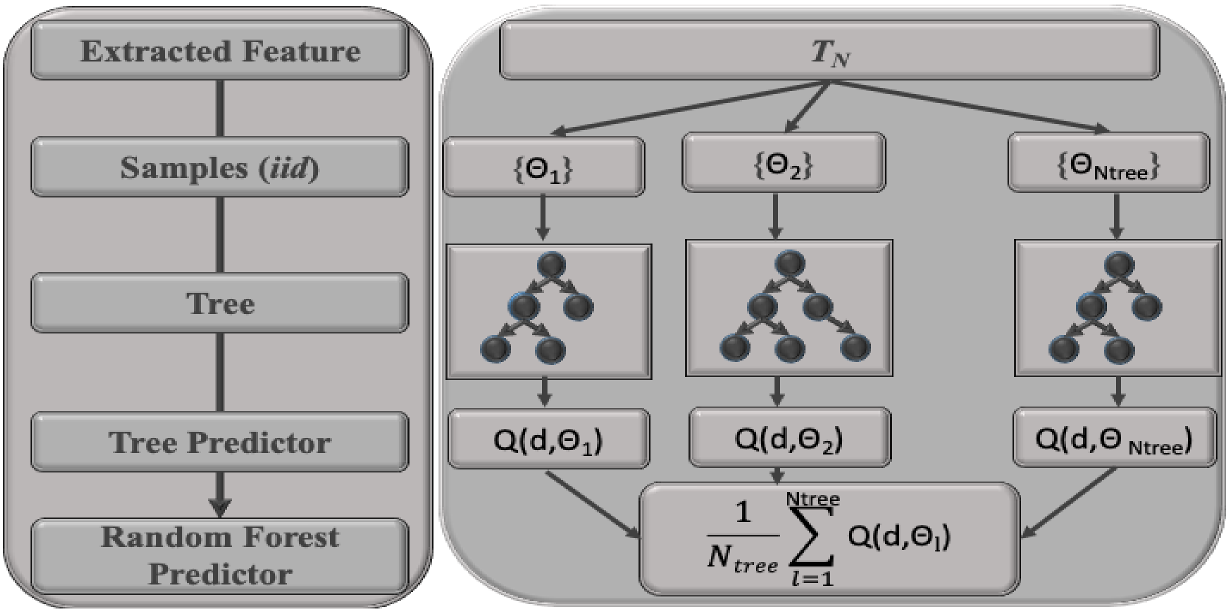


Figure 1: **Flow of random forest regression.**

importance measures provide some insight into an otherwise black box model". In this sense, it should be noted that trees give preference to predictors that have more levels or values (Strobl et al., 2007), and this is because of the way the cut-points are chosen. In this paper, we are interested in RF regression, as despite some shortcomings, the benefits of using RF outweigh the disadvantages. The essence of RF is an improvement to the decision tree algorithm, which combines multiple decision trees. Or otherwise stated, RF is based on the assumption that an ensemble of trees highly increases the prediction accuracy compared to a single tree (Genuer et al., 2017; Hastie et al., 2009), as the ensemble reduces the variance.

4. Data Processing

4.1. Inputs/Outputs

The financial data of the banks considered were collected from statistical tables available on the RBI website; while the corporate governance data, in the absence of reliable and required data in available databases, were hand-picked from the published annual reports of the banks. In view of data availability, we confined our study up to the year 2018. Furthermore, due to the unavailability of governance data, foreign banks were excluded from the study. The public and private sector banks that were merged during the said period were also excluded from the study. Thus, the final sample consisted of 21 public sector banks and 17 private sector banks, over an 11-year long period, which finally yielded a total of 412 observations. As previously mentioned, this study covers a longer period of time when compared to earlier studies. The sample bank list is given in Table A1 in the Appendix.

The literature on bank efficiency recommends mainly two different approaches for input and output variable selection, *i.e.*, intermediation and production (Sealey & Lindley, 1977). While the production approach is appropriate for branch-level studies, a large body of literature considers banks as financial intermediaries between depositors and borrowers. Following Berger and Humphrey (1997) and Fujii et al. (2014), we use the intermediation approach, taking three inputs and three outputs as variables. The inputs are (1) Loanable Funds (Deposits + Borrowings), (2) Labour (measured as number of employees), and (3) Physical Capital (measured as the value of the fixed assets). The prices of inputs are interest paid on funds (deposits and borrowings), employee expenses, and total operating expenses after adjusting for employee expenses. The outputs are (1) Advances, (2) Investments, and (3) Non-Interest Income (which is the income from fee and off-balance sheet activities). With a prolonged trend of tighter interest spreads, banks are increasingly under pressure to generate non-fund-based or fee-based income to tide over low profits and organically build up capital. Hence, non-interest income is considered as an important bank output variable. The prices of outputs are interest income from advances, income from investments, and equal to unity (for Non-Interest Income), respectively. Table 1 presents the summary statistics of these variables.

Table 1: Descriptive Statistics for Variables

Variable	Description	Mean	SD	Minimum	Maximum
Inputs	Labour	26114.95	36396.24	544.00	264041.00
	Physical Capital	1870.02	3329.48	16.41	42918.92
	Loanable Funds	199470.80	295244.10	1101.38	3068485.00
Outputs	Advances	135657.40	203030.40	585.79	1934880.00
	Investments	57985.51	88312.62	361.32	1060987.00
	Non-Interest Income	2542.45	4367.61	8.73	44600.69
Input Prices	Price of Labour	2023.67	3414.14	17.12	33178.68
	Price of Physical Capital	1697.76	2802.49	12.63	26764.77
	Price of Loanable Funds	11156.66	15007.45	52.38	145645.60
Output Prices	Price of Advances	12071.42	16496.76	60.04	141363.20
	Price of Investments	3999.89	6072.70	23.96	70337.62

Note. SD - Standard deviation.

The literature suggests that the choice of the input and output variables is at the researcher's discretion; our choice is in line with Andrieș and Căpraru (2014), Andrieș, Căpraru, and Nistor (2018), Gulati and Kumar (2016), and Ray and Das (2010), among others. Moreover, with 38 DMUs, 3 inputs, and 3 outputs, Cooper et al.'s (2007) rule of thumb, according to which the number of observations should be at least the maximum between the product of the number of input and output variables and three times the sum of the number of input and output variables in order to ensure the discriminatory power among the DMUs, is very well satisfied. For a more comprehensive discussion of the discriminatory power and "curse of dimensionality", as well as the rules of thumb available in the literature and recent developments, the interested reader is referred to the recent study by Charles et al. (2019).

4.2. Determinants of bank efficiency

Corporate governance studies (such as, Belkhir, 2009; Laeven & Levine, 2009) have generally used either a single parameter to represent bank governance, such as the Governance Index, or multiple proxies, such as Board Size, Board Independence, Gender Diversity, and Number of Board Meetings (De Andres & Vallelado, 2008; Huang, 2010), just to name a few. For example, recently, Andrieș et al. (2018) employed a Corporate Governance Index, while

Battaglia and Gallo (2015), Boitan and Nitescu (2019), and Liang et al. (2013) considered Board Characteristics, such as Size, Meetings, Duality, Independence, and Gender Diversity. Studies have used different measures of bank performance, such as TE (Adeabah et al., 2019; Boitan & Nitescu, 2019) and CE using DEA (Andrieș et al., 2018), and financial ratio-based variables, such as Tobin's Q, ROA, ROE, and P/E (Battaglia & Gallo, 2015; Liang et al., 2013) as dependent variables. While traditional ratios are easy to measure and represent more popular indicators of performance, they do not effectively capture the different dimensions of the operations as they use limited information. Bank performance measures using DEA are more informative as they are captured via multiple input and output variables.

There are three important economic efficiency concepts, namely TE, CE, and PE (Berger & Mester, 1997), which are based on economic optimisation in reaction to market prices and competition, rather than based solely on the use of technology. Moreover, the measurement of different types of efficiency adds some independent information value (Berger & Mester, 1997), and hence we can expect differences in the relationship between Governance variables and different efficiency measures. Following Klaassen and Eeghen (2015), we also control for bank characteristics, such as size, return on assets, equity to total assets, ownership, and year effects. Banks that are profitable are more allocative efficient (Sufian & Majid, 2009) and cost efficient (Hasan & Marton, 2003; Isik & Hassan, 2002), because they are preferred by both depositors and borrowers. Nair and Vinod (2019) conducted a study on allocative, cost, and scope efficiency of Indian banks and took profitability as an important internal determinant of these efficiencies. Casu et al. (2004), in their study on cost efficiency in Italian banks, used performance (measured as the ratio of net income by equity) as an important determinant of cost inefficiency. In their study of bank efficiency in East Asian countries, Ariff and Can (2009) found ROA to be a positively significant factor affecting the TE of banks. Their findings showed that banks that are more profitable are more technically efficient. Therefore, we use performance (measured through ROA) as an important determinant of banks' NTE, NCE, and NPE.

5. Analytics

5.1. Descriptive Analytics

The key financial data of the Indian banks, both public and private, for the period spanning 2008-2018, are given in the Appendix, in Tables A2 and A3, respectively. It can be noticed that public sector banks dominate the banking sector, with almost three times more average deposits and advances when compared to the private sector banks; while the average capital in terms of shareholders' fund for both types of banks is comparable. Public sector banks grew at a slower rate in terms of total assets and income when compared to private banks. Interestingly, the profit of the private banks grew by 4.38 times, while that of the public sector banks declined, with a loss in 2018, mainly due to the high number of write-offs. In terms of other income, private sector banks seem to have been doing better when compared to public sector banks. While in terms of total assets and income, public sector banks have been twice as big as private sector banks, with the profits of the former being about half of the latter. The average ROA of private banks was 0.9 percent, compared to just 0.35 percent for public banks. The difference in ROE between public banks and private banks is also large despite higher leverage in the former. In essence, the income and asset growth of private and public sector banks are comparable, but the profits of private banks are far better, indicating a need for better management and governance of public sector banks.

The year-wise descriptive statistics of the sampled public and private banks are given in the Appendix, in Tables A4 and A5, respectively. We see heterogeneity across the banks, with the largest bank having over 0.26 million employees and the smallest bank having just 1884, in the year 2018. Similarly, we see the total funds and the other income of the largest bank to be about 300 times and 450 times higher, respectively, when compared to those of the smallest bank, in 2018. Such high heterogeneity obviously poses methodological challenges for conventional approaches such as regression and, hence, RF is aptly suited to overcome these issues.

5.2. Phase I: Efficiency analysis

The DEA results are summarised in Table 1 and the efficiency charts in Figure 2. The detailed results of NTE, NCE, and NPE are available upon request from the authors.

New Technical Efficiency: During the study period, the NTE of all the banks was about 96 percent and it remained stable. The efficiency levels of both the public and private sector banks remained very high and comparable. However, in terms of the number of banks situated on the efficiency frontier, we found that over 50 percent of the private banks were efficient, compared to almost 60 percent of the public sector banks. The high NTE indicates that public and private sector banks have been very effective in using their inputs to generate outputs.

New Cost Efficiency: NCE was marginally lower when compared to NTE, with an average of 95 percent. The NCE of both public and private sector banks remained comparable; nevertheless, over the study period, we found about 33 percent of the private banks on the efficiency frontier, compared to just 22 percent of the public sector banks. The Indian banks were effective in controlling their input cost and there was limited scope for improvement (up to 5 percent). The higher NCE of public sector banks when compared to private banks can be due to public sector banks being older and on an average three times in size in terms of the total income and total fund than those of private sector banks.

New Profit Efficiency: Across the ownership groups, the average NPE of public sector banks was a shade better than that of private sector banks. Incidentally, the NPE of both public and private sector banks declined during the period of study. Over the period, and on an average, 54 percent of the private banks were on the efficiency frontier, compared to 48 percent of the public banks. Private sector banks had better net interest margin and higher current and savings ratio than public sector banks. Non-interest income as a proportion of total funds was also significantly better than that of public sector banks. We also saw a larger variation in the average NPE of all the bank groups, when compared to the average NCE and average NTE.

As per the figures in Table 2, 21 banks were on the efficiency frontier for NTE, 19 banks for NPE, and 10 banks for NCE. Interestingly, the average NPE of all the banks was lower while the number of banks on the efficiency frontier was higher, indicating high heterogeneity in NPE across banks. In terms of ownership groups, we found a far larger proportion of private sector banks on the efficiency frontier when compared to public sector banks, suggesting higher inefficiencies in state-owned banks. Although banks across all groups are comparable in terms of NTE, we found that, overall, private banks count higher on the NPE frontier, pointing towards the importance of corporate governance and the management role in bank profitability.

Table 2: Sector Level Statistics for NTE, NCE, and NPE

Sector	Stats	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
			New	Technical		Efficiency						
All	EB	24	22	20	23	20	21	20	18	17	21	23
	Mean	0.9765	0.9718	0.9575	0.9667	0.9591	0.9622	0.9486	0.9484	0.9645	0.9656	0.9702
	SD	0.0439	0.0589	0.0615	0.0607	0.0736	0.0562	0.0690	0.0621	0.0418	0.0475	0.0466
	CV*	0.0452	0.0610	0.0646	0.0632	0.0772	0.0587	0.0732	0.0658	0.0436	0.0495	0.0483
State	EB	12	12	12	13	12	13	11	10	10	14	13
	Mean	0.9772	0.9820	0.9663	0.9769	0.9764	0.9814	0.9610	0.9549	0.9754	0.9677	0.9704
	SD	0.0397	0.0312	0.0487	0.0402	0.0384	0.0306	0.0514	0.0482	0.0303	0.0541	0.0492
	CV*	0.0411	0.0321	0.0509	0.0416	0.0397	0.0315	0.0541	0.0510	0.0314	0.0565	0.0512
Private	EB	12	10	8	10	8	8	9	8	7	7	10
	Mean	0.9756	0.9593	0.9465	0.9540	0.9376	0.9384	0.9334	0.9404	0.9510	0.9629	0.9699
	SD	0.0498	0.0806	0.0744	0.0788	0.0990	0.0710	0.0852	0.0767	0.0505	0.0392	0.0446
	CV*	0.0518	0.0853	0.0798	0.0838	0.1071	0.0767	0.0926	0.0827	0.0539	0.0413	0.0467
			New	Cost	Efficiency							
All	EB	11	11	9	8	8	9	11	10	11	14	11
	Mean	0.9425	0.9225	0.9071	0.9014	0.8706	0.9139	0.9032	0.9158	0.9285	0.9268	0.9318
	SD	0.0713	0.0879	0.0770	0.0697	0.0964	0.0746	0.0806	0.0736	0.0620	0.0690	0.0595
	CV*	0.0762	0.0959	0.0854	0.0778	0.1114	0.0822	0.0898	0.0809	0.0671	0.0750	0.0642
State	EB	6	6	5	3	2	4	5	4	5	7	5
	Mean	0.9530	0.9406	0.9251	0.9154	0.8737	0.9217	0.9047	0.9140	0.9329	0.9244	0.9287
	SD	0.0512	0.0635	0.0622	0.0487	0.0710	0.0681	0.0681	0.0641	0.0565	0.0723	0.0599
	CV*	0.0542	0.0683	0.0680	0.0538	0.0821	0.0746	0.0761	0.0708	0.0612	0.0790	0.0651
Private	EB	5	5	4	5	6	5	6	6	6	7	6
	Mean	0.9294	0.9001	0.8848	0.8842	0.8668	0.9042	0.9014	0.9180	0.9232	0.9297	0.9356
	SD	0.0904	0.1090	0.0889	0.0878	0.1231	0.0831	0.0960	0.0859	0.0695	0.0669	0.0606
	CV*	0.0987	0.1228	0.1019	0.1007	0.1441	0.0932	0.1081	0.0950	0.0764	0.0730	0.0657
			New	Profit	Efficiency							
All	EB	25	22	18	19	21	19	15	16	19	21	18
	Mean	0.9249	0.8603	0.8041	0.8545	0.8707	0.8616	0.7683	0.6595	0.7694	0.8525	0.8582
	SD	0.1602	0.2246	0.2658	0.2332	0.2324	0.2153	0.2600	0.3515	0.2943	0.2083	0.1875
	CV*	0.1742	0.2627	0.3326	0.2745	0.2686	0.2514	0.3404	0.5362	0.3849	0.2458	0.2198
State	EB	15	11	8	9	11	11	7	8	10	11	10
	Mean	0.9336	0.8547	0.8226	0.8454	0.9089	0.9041	0.7766	0.6442	0.8020	0.8565	0.8539
	SD	8.3299	8.3483	8.3546	8.3500	8.3357	8.3364	8.3649	8.3976	8.3606	8.3475	8.3477
	CV*	9.0152	9.8695	10.2619	9.9798	9.2868	9.3169	10.8828	13.1722	10.5332	9.8481	9.8778
Private	EB	10	11	10	10	10	8	8	8	9	10	8
	Mean	0.9140	0.8672	0.7812	0.8658	0.8235	0.8091	0.7580	0.6785	0.7290	0.8477	0.8635
	SD	0.1949	0.2509	0.3463	0.2762	0.2997	0.2771	0.3206	0.3635	0.3440	0.2325	0.2093
	CV*	0.2163	0.2935	0.4499	0.3237	0.3693	0.3475	0.4292	0.5436	0.4788	0.2783	0.2460

Note. EB - Number of Efficient Banks, SD - Sample Standard Deviation, and CV* - Adjusted Coefficient of Variance.

5.3. Phase II: Determinants of bank efficiency

5.3.1. Random forest regression:

First, all potential predictors were included in the preliminary RF regression model and measures of variable importance, *i.e.*, decrease in mean square errors and decrease in node purities were calculated for each of them. Then, the predictors included in the final model were selected according to measures of variable importance. The analysis using RF regression was conducted for NTE, NCE, and NPE as targets, with RF having been run for each case, by taking corporate governance, bank characteristics, and other characteristics as features (explanatory variables).

In order to find the optimal value of $mtry$, we first set N_{tree} to a large value (N_{tree} = 3000), enough to have RF regression with stable prediction. It should be noted that by default, randomForest() uses $p/3$ variables when building an RF of regression trees. Here, with ten variables, we have $mtry = 3$ for all the three cases of efficiency: NTE, NCE, and NPE; and this is what we report in Table 3 as the initial value. Then, we varied $mtry$ from the smallest value up to the maximum. Thus, we have varied $mtry$ from 1 to 10. For each value of $mtry$, we computed the OOBerror (see Figures 3a, 4a, and 5a). The lowest OOBerror was obtained for $mtry = 2$ for NTE (Figure 3a), $mtry = 3$ for NCE (Figure 4a), and $mtry = 4$ for NPE (Figure 5a).

Table 3: Random Forest Statistics

Type	RF	Data	N_{tree}	$mtry$	MSE	R^2	Cor	Min-Max Accuracy
NTE	Intial	Train	3000	3	0.00322	17.72	-	-
	Tuning	Train	2470	2	0.00319	18.31	0.3917	0.9619
	Tuning	Test	2470	2	0.00199	11.96	0.3663	0.9615
NCE	Intial	Train	3000	3	0.00422	25.47	-	-
	Tuning	Train	2770	3	0.00420	25.87	0.5689	0.9426
	Tuning	Test	2770	3	0.00480	24.75	0.5121	0.9416
NPE	Intial	Train	3000	3	0.04611	28.13	-	-
	Tuning	Train	930	4	0.04627	27.88	0.5293	0.8112
	Tuning	Test	930	4	0.05722	22.08	0.5394	0.8134

Note. N_{tree} - Number of decision trees, $mtry$ - Number of candidate features, and Cor - Correlation.

The second optimal value to find was N_{tree}. For this, OOBerrors were computed for different values of N_{tree}. As it can be observed from Figures 3b, 4b, and 5b, respectively, the error stabilised when the value of the trees was 2470 for NTE, 2770 for NCE, and 930 for NPE, respectively, indicating that the minimum (optimal) number of trees was reached and that the obtained RF reached its optimal prediction error; otherwise stated, the optimal RF regression was reached. Through the iterative learning method, the optimal number of decision trees and the optimal number of splitting features were searched within a certain range, so that each potential regressor could play the maximum role in prediction, accelerate the convergence speed, and improve the regression accuracy. The train set MSEs for the tuned model are 0.00319, 0.00420, and 0.04627, respectively (compared to 0.00322, 0.00422, and 0.04611, respectively, for the initial model with train set), indicating that, overall, RFs yielded an improvement over bagging. The only exception is posed by the case of NPE; nevertheless, the difference is negligible.

The percentage of variation explained is a measure of how well OOB predictions explain the target variance of the training set. As expected, the percentage of variation explained increased after tuning (training data), which indicated that there was an increment in the explanatory power in the predictor variables. The only exception is represented by NPE, for which the R^2 value slightly decreased from 28.13% to 27.88% after tuning (training data);

nonetheless, the difference is minimal and negligible, especially considering the improvement obtained in the number of trees, which decreased significantly from 3000 to 930. Finally, the percentage of variation explained by the tuned model with test data is 24.75% for NCE and 22.08% for NPE, while for NTE it is 11.96%.

MinMax indicates how far the model's prediction is off. For a perfect model, this measure is 1.0. Based on Table 2, we can observe that the min-max accuracy values in the case of NTE and NCE are close to 1, indicating a high prediction power; for NPE, however, the min-max accuracy is 0.8112, which is still a good value, despite being lower than in the case of NTE and NCE.

5.3.2. Traditional Regressions:

We also employ OLS bootstrap and fractional logit bootstrap regression to study the influence of corporate governance, bank characteristics, and other characteristics on NTE, NCE, and NPE of Indian banks (see Table A6 in the Appendix). For both the regressions, Wald χ^2 is significant for all the three types of efficiency. However, OLS bootstrap is more promising when compared to fractional logit bootstrap, as the percentage of variation explained (R^2) by the former is higher than the pseudo R^2 of the latter, for NTE, NCE, and NPE, respectively. This fact can also be observed from the respective $RMSE$ values.

As estimated through the OLS method, duality impacts significantly and positively all three measures of efficiency. Board size and board independence negatively impact on NPE; board meetings negatively affect NCE and NPE; and gender diversity does not have a significant impact on any of the efficiencies. On the other hand, the results of fractional logit bootstrap regression highlight that board size and board independence negatively impact NPE; board meetings negatively impact NCE and NPE; and gender diversity does not have a significant impact on any of the efficiencies. These results are consistent with the ones obtained through the OLS bootstrap. The only difference is represented by duality, which unlike in the case of the OLS method, it is found to impact only one type of efficiency of Indian banks, *i.e.*, NPE.

Moreover, for both OLS bootstrap and fractional logit bootstrap, a larger number (four) of governance variables have a significant impact on NPE. For OLS bootstrap, just two governance variables affect NCE and only one affects NTE, while for fractional logit bootstrap, just one variable impacts NCE and none impacts NTE. In terms of bank characteristics, size and ROA have a positive impact on all the efficiency measures, while equity to total assets only impacts NPE. In terms of other characteristics, ownership impacts only NPE.

6. Discussion

The Mean Decrease Accuracy, %IncMSE, shows how much our model accuracy decreases if we leave out a particular variable. Similarly, Mean Decrease Gini (%IncNodePurity) is a measure of variable importance based on the Gini impurity index used for calculating the splits in trees; in this case, the higher the value of the Mean Decrease Accuracy or Mean Decrease Gini score, the higher the importance of the variable for the model. %IncMSE is the most robust and informative measure, while %IncNodePurity is a biased measure which should only be used if the extra computation time of calculating %IncMSE is unacceptable. In our case, the %IncMSE results will be used for further analysis. By looking at the results obtained, we can easily observe that in the case of NTE (Figure 3c), the ROA predictor plays an important role in the accuracy of our model, followed by gender diversity, bank size, equity to total assets, duality, year, board independence, ownership, board meetings, and board size; in the case of NCE (Figure 4c), the predictor playing the highest role is bank size, followed by ROA, board independence, board meetings, gender diversity, year, equity to total assets, duality, board size, and ownership; and in the case of NPE (Figure 5c), the ROA predictor plays the greatest role, followed by bank size, bank independence, year, equity to total assets, ownership, gender diversity, duality, board size, and board meetings. We observe that board characteristics could explain NPE better than NTE and NCE. Profit efficiency is more informative (Berger & Mester, 1997) and is derived by using the prices of inputs and outputs as well, where the board would have influence. As the banks operate in a highly regulated environment, what differentiates one bank from another is not the volume of inputs and outputs, but the price; or, otherwise stated, the management of the cost of inputs and prices of outputs. Interestingly, our results differ from Narwal and Pathneja (2016), who found that among all governance variables in their study, none of them explained productivity in banks. Our findings are similar to Liang et al. (2013), who found that the number of board meetings and independent directors have a significant and positive impact on bank performance and asset quality, while board size has a negative effect on bank performance. Our findings further differ from Andrieș et al. (2018), who found a significant and negative impact of the governance structure on bank efficiency, while Pathan and Faff (2013) reported that board size and independent directors decrease bank performance, while gender diversity improves bank performance.

TE is relevant in a non-market environment where the prices of inputs and outputs are not reliable or available. Unlike CE, PE is more informative (Ray & Das, 2010), since as mentioned, it requires the prices of both inputs and outputs. In our study, we observe very little variation across the NTE scores of banks, while there is larger variation in NPE across the different ownership groups and years (see Table 1). We also find that our second phase RF model explains 11.96 percent of the variance in NTE, 24.75 percent of the variance in NCE, and 22.08 percent of the variance in NPE. NPE results are more reasonable when compared to NCE and NTE results. NPE is more informative and offers better insights. Board independence brings diversity, expertise, better monitoring, and advising, which reduces agency cost, improves the decision quality and, hence, bank performance. The number of board meetings is indicative of more discussion, advising, and supervision, and of a positive impact on decision quality (Battaglia & Gallo, 2015); hence, it is a better predictor of bank efficiency.

7. Implications

As previously mentioned, RBI appointed the Nayak Committee to review the governance of boards and give recommendations. The committee submitted its report in May 2014, which later formed the basis for the development of the RBI's set of rules for bank governance; in this sense, our study contributes to the understanding of the role of bank governance in bank efficiency. Nayak et al. (2014) observed that the number of issues (such as business strategy, risk, customer protection, financial results, compliance, and financial inclusion, among others) discussed and deliberated during board meetings are positively related to bank profitability. The committee further pointed out that there is a need to upgrade the quality of the discussion in board meetings and indicated that this is dependent on the skills and the independence of board members. RBI is therefore placing importance on independent directors, non-executive directors, and their role in certain decisions. Moreover, the regulator is keen on knowing the quality of debates in the meetings and the involvement of the independent directors. Our results corroborate the Nayak Committee's findings on the importance of the number of board meetings, board size, and board independence as predictors of bank performance. We also acknowledge the importance of the quality of the discussion in the meetings and the role played by the independent directors; these aspects, however, are beyond the scope of this study due to the unavailability of such qualitative data.

Our study has implications for regulators, bank management, and researchers alike, regarding how the board characteristics relate to the bank managerial efficiency, namely NTE, NCE, and NPE. We found that for NPE, board characteristics play a more significant role; therefore, policymakers and regulators should consider board features such as board size, board independence, board meetings, and duality while framing guidelines for enhancing bank performance. At the same time, board meetings can be taken into account for NCE and duality for NTE and NCE. Furthermore, given the typical characteristics of the features, researchers may want to consider RF regression over the conventional one to overcome the issues around the non-significant features, to improve the quality of the results. In this sense, the RF results show that board gender diversity as a corporate governance characteristic is actually also an important contributor to NTE, although the OLS regression deemed it as not significant. Likewise, equity to total assets is also an important bank characteristic contributing to NTE. On the other hand, RF also indicates that board independence and board diversity are two important corporate governance characteristics that contribute to NCE, although yet again the OLS regression indicates otherwise. This points to the limitations of traditional regression analyses and the usefulness of considering RF to complement such analyses and improve the quality of the insights obtained.

8. Conclusion

In this study, we examined the impact of corporate governance, bank characteristics, and other characteristics on bank efficiency (NTE, NCE, and NPE) of 38 banks in India, across different ownership groups. We took an 11-year long eventful period, spanning the years 2008-2018. We used DEA to estimate bank efficiency in the first phase and RF regression in the second phase. By recognising the limitations of using regression analysis, our study pioneered the use of a methodologically superior and more accurate method, *i.e.*, RF, to find the predictors of bank efficiency.

The first-phase DEA revealed that the average NTE was higher than NCE, while NPE was lower, for all the banks. Banks have been efficient in their intermediation role, but lacked in terms of profitability and low cost. Furthermore, while the average NPE of public sector banks was lower than that of private banks, it showed better improvement over the study period. Interestingly, a larger number of banks were on the efficiency frontier of NPE than of NTE and NCE. Moreover, a higher number of private banks were on the NPE frontier compared to public sector banks. The larger number of banks on the NPE frontier and the lower average NPE across all the banks indicate high heterogeneity in bank efficiency across different ownership and size groups.

The RF regression used in the second phase showed that particularly for NPE, board characteristics play a vital role. Our study found that the predictors of NTE are duality, ROA, bank size, gender diversity, and equity to total assets; the predictors of NCE are board meetings, duality, ROA, bank size, bank independence, and gender diversity; and the predictors of NPE are board size, board independence, board meetings, duality, ROA, equity to total assets, bank size, and ownership. We also found that RF regression performs better than OLS bootstrap and fractional logit bootstrap, providing more convincing results in terms of modelling performance, as well as variable importance.

Overall, our paper makes several contributions: First, this study makes methodological contributions by employing machine learning based RF regression to examine corporate governance and bank efficiency, which is a pioneering attempt; we also argue about its superiority over conventional OLS bootstrap and fractional logit bootstrap. Second, there are hardly any studies in the literature that explored the link between corporate governance and bank efficiency in the Indian context, and our study fills this gap. Third, all three types of efficiency (NTE, NCE, and NPE) are examined over an 11-year period, while most past studies covered one or two types of efficiency, and within a shorter time frame. To the best of our knowledge, this is the only study in the Indian context that covers a longer time period and all three types of efficiency. Overall, our results are more comprehensive and accurate as they better predict a superior measure of bank efficiency, *i.e.*, NPE. At a time when RBI is actively contemplating the revision of corporate governance rules for banks in India, our study has important policy implications.

Acknowledgments

The authors would like to thank the Editor in Chief, the Associate Editor, and two anonymous reviewers for their valuable feedback on the previous version of this manuscript.

References

Adams, R. B., & Mehran, H. (2005). Corporate performance, board structure and its determinants in the banking industry. In *EFA 2005 Moscow Meetings*. Available at SSRN: http://dx.doi.org/10.2139/ssrn.302593

Adeabah, D., Gyeke-Dako, A., & Andoh, C. (2019). Board gender diversity, corporate governance and bank efficiency in Ghana: a two stage data envelope analysis (DEA) approach. *Corporate Governance: The International Journal of Business in Society, 19*(2), 299—320.

Andrieş, A. M., & Căpraru, B. (2014). The nexus between competition and efficiency: The European banking industries experience. *International Business Review*, *23*(3), 566—579.

Andrieș, A. M., Căpraru, B., & Nistor, S. (2018). Corporate governance and efficiency in banking: evidence from emerging economies. *Applied Economics*, *50*(34-35), 3812—3832.

Ariff, M., & Can, L. (2009). IMF bank-restructuring efficiency outcomes: Evidence from East Asia. *Journal of financial services research*, *35*(2), 167—187.

Basel Committee. (2015). *Corporate Governance Principles for Banking Organizations*. Basel Committee on Banking Supervision, BIS. Available at https://www.bis.org/bcbs/publ/d328.htm.

Battaglia, F., & Gallo, A. (2015). Risk governance and Asian bank performance: An empirical investigation over the financial crisis. *Emerging Markets Review*, *25*, 53—68.

Belkhir, M. (2009). Board of directors' size and performance in the banking industry. *International Journal of Managerial Finance*, *5*(2), 201—221.

Berger, A. N. (2007). International comparisons of banking efficiency. *Financial Markets, Institutions & Instruments*, *16*(3), 119—144.

Berger, A. N., Hancock, D., & Humphrey, D. B. (1993). Bank efficiency derived from the profit function. *Journal of Banking & Finance*, *17*(2-3), 317—347.

Berger, A. N., & Humphrey, D. B. (1997). Efficiency of financial institutions: International survey and directions for future research. *European Journal of Operational Research*, *98*(2), 175—212.

Berger, A. N., & Mester, L. J. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking & Finance*, *21*(7), 895—947.

Boitan, I. A, Nitescu, D.C. (2019). Does corporate governance support efficiency in banking business? Evidence from European systemic banks. *Ekonomicky Casopis*, *67*(5), 525—549.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5—32.

Casu, B., Girardone, C., & Molyneux, P. (2004). Productivity change in European banking: A comparison of parametric and non-parametric approaches. *Journal of Banking & Finance*, *28*(10), 2521—2540.

Casu, B., & Molyneux, P. (2003). A comparative study of efficiency in European banking. *Applied Economics*, *35*(17), 1865—1876.

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929—940.

Charles, V., Peretto, C., & Gherman, T. (2016). Technical Efficiency Analytics for the Public Banks in Argentina. *ICPE Public Enterprise Half-Yearly Journal*, *22*(1-2), 118—139.

Charles, V., Tsolas, I. E., & Gherman, T. (2018). Satisficing data envelopment analysis: a Bayesian approach for peer mining in the banking sector. *Annals of Operations Research*, *269*(1/2), 81—102.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (2nd ed.)*. New York, NY: Springer.

Cornett, M. M., McNutt, J. J., & Tehranian, H. (2009). Corporate governance and earnings management at large US bank holding companies. *Journal of Corporate Finance*, *15*(4), 412—430.

Das, A., Nag, A., & Ray, S. C. (2005). Liberalisation, ownership and efficiency in Indian banking: a nonparametric analysis. *Economic and Political Weekly*, *40*(2), 1190—1197.

De Andres, P., & Vallelado, E. (2008). Corporate governance in banking: The role of the board of directors. *Journal of Banking & Finance*, *32*(12), 2570—2580.

De Haan, J., & Vlahu, R. (2016). Corporate governance of banks: A survey. *Journal of Economic Surveys*, *30*(2), 228—277.

Donaldson, L., & Davis, J. H. (1994). Boards and Company Performance – Research Challenges the Conventional Wisdom. *Corporate Governance: An International Review*, *2*, 151—160.

Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The Journal of Law and Economics*, *26*(2), 301—325.

Fethi, M. D., & Pasiouras, F. (2010). Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey. *European Journal of Operational Research*, *204*(2), 189—198.

Francis, B. B., Hasan, I., & Wu, Q. (2012). Do corporate boards affect firm performance? New evidence from the financial crisis. *Bank of Finland Research Discussion Paper* No. 11/2012. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2041194

Fujii, H., Managi, S., & Matousek, R. (2014). Indian bank efficiency and productivity changes with undesirable outputs: A disaggregated approach. *Journal of Banking & Finance*, *38*, 41—50.

Genuer, R., Poggi, J. M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017, September). Random forests for big data. *Big Data Research*, *9*, 28—46.

Gulati, R., & Kumar, S. (2016). Assessing the impact of the global financial crisis on the profit efficiency of Indian banks. *Economic Modelling*, *58*, 167—181.

Hallett, M., Fan, J., Su, X., Levine, R., & Nunn, M. (2014). Random forest and variable importance rankings for correlated survival data, with applications to tooth loss. *Statistical Modelling*, *14*(6), 523—547.

Hasan, I., & Marton, K. (2003). Development and efficiency of the banking sector in a transitional economy: Hungarian experience. *Journal of Banking & Finance*, *27*(12), 2249—2271.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning. Data mining, inference, and prediction (2nd ed.).* New York, NY: Springer.

Ho, T. K. (1998). Nearest neighbors in random subspaces. Paper presented at the *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Sydney, Australia, 1998.

Hussain, H. I., Kamarudin, F., Mohamad Anwar, N. A., Nassir, A. M., Sufian, F., & Tan, K. M. (2020). Impact of Country's Governance Dimensions on Bank Revenue Efficiency: Overview on Middle East, Southeast Asia, and South Asia Countries. *Transformations in Business & Economics*, *19*(1), 191-228.

Isik, I., & Hassan, M. K. (2002). Technical, scale and allocative efficiencies of Turkish banking industry. *Journal of Banking & Finance*, *26*(4), 719—766.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305—360.

Kamarudin, F., Anwar, N. A. M., Sufian, F., Tan, K. M., & Hussain, H. I. (2020). Does country governance and bank productivity Nexus matters? *Journal of Islamic Marketing.*

Kamarudin, F., Sufian, F. & Nassir, A.M. (2016). Does country governance foster revenue efficiency of Islamic and conventional banks in GCC countries? *EuroMed Journal of Business*, *11*(2), 181—211.

Klaassen, P., & van Eeghen, I. (2015). Analyzing Bank Performance: Linking ROE, ROA and RAROC: U.S. Commercial Banks: 1992-2014. *Journal of Financial Perspectives*, *3*(2). Available at SSRN: https://ssrn.com/abstract=3083577

KPMG International. (2017). *The India-U.S. dynamic – better together.* https://home.kpmg/in/en/home/insights/2017/11/india-us-dynamic-relationship.html

Kraay, A., Kaufmann, D., & Mastruzzi, M. (2010). *The worldwide governance indicators: methodology and analytical issues.* Washington, DC: The World Bank.

Laeven, L., & Levine, R. (2009). Bank governance, regulation and risk taking. *Journal of Financial Economics*, *93*(2), 259—275.

Liang, Q., Xu, P., & Jiraporn, P. (2013). Board characteristics and Chinese bank performance. *Journal of Banking & Finance*, *37*(8), 2953—2968.

Liu, Y., Miletkov, M. K., Wei, Z., & Yang, T. (2015). Board independence and firm performance in China. *Journal of Corporate Finance*, *30*, 223—244.

Nair, A. S., & Vinod, R. (2019). Determinants of the allocative, cost and scope efficiencies of Indian banks. *Applied Economics*, *51*(5), 509—527.

Narwal, K. P., & Pathneja, S. (2016). Effect of bank-specific and governance-specific variables on the productivity and profitability of banks. *International Journal of Productivity and Performance Management*, *65*(8), 1057—1074.

Nayak, P. J., Raman, S., Panse, S., Kar, P., Sengupta, J., Vardhan, H., Sundaresan, S., & Subramanian, K. (2014). *Report of The Committee to Review Governance of Boards of Banks in India.* Mumbai: Reserve Bank of India. https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/BCF090514FR.pdf

Pant, A., & Radhakrishnan, G. (2019). Corporate governance challenges in Indian banks: A public affairs perspective. *Journal of Public Affairs*, *19*(4), e1946.

Pathan, S. (2009). Strong boards, CEO power and bank risk-taking. *Journal of Banking & Finance*, *33*(7), 1340—1350.

Pathan, S., & Faff, R. (2013). Does board structure in banks really affect their performance? *Journal of Banking & Finance*, *37*(5), 1573—1589.

Ray, S. C., & Das, A. (2010). Distribution of cost and profit efficiency: Evidence from Indian banking. *European Journal of Operational Research*, *201*(1), 297—307.

Rosenstein, S., & Wyatt, J. G. (1990). Outside directors, board independence, and shareholder wealth. *Journal of Financial Economics*, *26*(2), 175—191.

Sathye, S., & Sathye, M. (2017). Do ATMs Increase Technical Efficiency of Banks in a Developing Country? Evidence from Indian Banks. *Australian Accounting Review*, *27*(1), 101—111.

Saw, A. T. W., Kamarudin, F., & Latiff, A. S. A. (2020). Price Efficiency of Islamic and Conventional Banks: Evidence from Panel Data. *International Journal of Economics* *and Management*, *14*(2), 301-310.

Scornet, E., Biau, G., & Vert, J.P. (2015). Consistency of random forests. *The Annals of Statistics*, *43*(4), 1716—1741.

Sealey Jr, C. W., & Lindley, J. T. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. *The Journal of Finance*, *32*(4), 1251—1266.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(25), 1—21.

Sufian, F., & Kamarudin, F. (2014). The impact of ownership structure on bank productivity and efficiency: Evidence from semi-parametric Malmquist Productivity Index. *Cogent Economics & Finance*, *2*(1), 932700.

Sufian, F., & Majid, M. Z. A. (2009). Bank efficiency and share prices in China: empirical evidence from a three-stage banking model. *International Journal of Computational Economics and Econometrics*, *1*(1), 23—47.

Tone, K. (2002). A strange case of the cost and allocative efficiencies in DEA. *Journal of the Operational Research Society*, *53*(11), 1225—1231.

Tsolas, I. E., Charles, V., & Gherman, T. (2020). Supporting Better Practice Benchmarking: A DEA-ANN Approach to Bank Branch Performance Assessment. *Expert Systems with Applications*, *160*, 113599.

Table A1: List of Banks

DMU	Status	Name of the Banks	DMU	Status	Name of the Banks
DMU1	Public	Allahabad Bank	DMU20	Public	Vijaya Bank
DMU2	Public	Andhra Bank	DMU21	Private	Axis Bank
DMU3	Public	Bank of Baroda	DMU22	Private	Catholic Syrian Bank Ltd
DMU4	Public	Bank of India	DMU23	Private	City Union Bank Limited
DMU5	Public	Bank of Maharashtra	DMU24	Private	DCB Bank Limited
DMU6	Public	Canara Bank	DMU25	Private	Dhanlaxmi Bank
DMU7	Public	Central Bank of India	DMU26	Private	Federal Bank
DMU8	Public	Corporation Bank	DMU27	Private	HDFC Bank
DMU9	Public	Dena Bank	DMU28	Private	ICICI Bank
DMU10	Public	IDBI Bank Limited	DMU29	Private	Indusind Bank
DMU11	Public	Indian Bank	DMU30	Private	Jammu & Kashmir Bank Ltd
DMU12	Public	Indian Overseas Bank	DMU31	Private	Karnataka Bank Ltd
DMU13	Public	Oriental Bank of Commerce	DMU32	Private	Karur Vysya Bank
DMU14	Public	Punjab & Sind Bank	DMU33	Private	Kotak Mahindra Bank Ltd
DMU15	Public	Punjab National Bank	DMU34	Private	Lakshmi Vilas Bank
DMU16	Public	Syndicate Bank	DMU35	Private	Ratnakar Bank Ltd
DMU17	Public	UCO Bank	DMU36	Private	South Indian Bank
DMU18	Public	Union Bank of India	DMU37	Private	Yes Bank Ltd.
DMU19	Public	United Bank of India	DMU38	Public	State Bank of India

Table A2: Public Sector Banks: Key Financial Data

Stats	2008	09	10	11	12	13	14	15	16	17	2018	$Trend$	$CAGR$	Avg
II	91.24	117.94	132.87	160.81	213.19	243.22	272.55	298.27	304.53	300.80	314.46	11.91	3.45	222.72
OI	14.13	18.61	21.65	21.21	22.35	25.26	29.03	33.59	36.33	50.71	54.65	13.08	3.87	29.78
TI	105.37	136.55	154.52	182.02	235.54	268.48	301.59	331.86	340.86	351.51	369.10	12.07	3.50	252.49
IE	63.34	83.14	91.92	101.43	144.27	169.87	191.92	212.17	217.26	211.72	216.77	11.83	3.42	154.89
EE	12.45	15.23	18.20	24.27	25.36	28.21	32.96	35.64	38.44	39.56	43.70	12.09	3.51	28.55
RNP	11.58	15.05	17.14	19.67	21.85	22.34	16.30	16.34	-9.35	0.23	-40.65	-212.09	-3.51	8.23
SHF	77.20	92.32	106.30	129.00	158.23	181.98	207.37	223.14	236.76	259.99	280.47	12.44	3.63	177.52
Dep	1055.90	1356.09	1613.36	1933.58	2209.83	2538.09	2930.94	3204.99	3321.63	3588.54	3934.44	12.70	3.73	2517.04
Bor	68.33	114.21	140.60	178.21	206.26	249.71	283.71	289.59	362.82	335.69	403.35	17.52	5.90	239.32
Inv	345.49	443.33	535.20	593.32	667.69	779.52	880.93	919.27	998.69	1132.47	1329.46	13.03	3.85	784.12
$Advs$	771.64	982.03	1178.57	1460.48	1710.90	1971.07	2260.19	2429.48	2479.69	2504.45	2713.02	12.11	3.52	1860.14
TA	1301.19	1642.92	1943.80	2342.89	2668.50	3077.53	3543.55	3865.91	4077.17	4341.58	4778.52	12.55	3.67	3053.05
ROA	0.88	0.86	0.86	0.84	0.77	0.65	0.35	0.33	-0.31	-0.18	-1.19	-202.80	-135.51	0.35
ROE	15.95	15.70	16.56	14.93	13.33	11.34	6.02	5.92	-5.74	-3.39	-22.19	-203.05	-139.12	6.22
EQA	5.56	5.41	5.18	5.57	5.77	5.71	5.55	5.58	5.67	5.66	5.72	0.25	102.77	5.58

Interest Income (II), Other Income (OI), Total Income (TI), Interest Expended (IE), Employee Expenses (EE), Other exp, Reported Net Profit (RNP), ShareHolders Fund (SHF), Deposits (Dep), Borrowings (Bor), Investments (Inv), Advances (Advs), Total Assets (TA), Return on Assets (ROA %), Return on Equity (ROE %), EQ to Total Assets (EQA %), Trend in % CARG (%) and Average (Avg).

Table A3: Private Sector Banks: Key Financial Data

Stats	2008	09	10	11	12	13	14	15	16	17	2018	$Trend$	$CAGR$	Avg
II	38.30	47.19	45.78	54.33	75.57	93.39	106.35	120.59	140.57	154.58	169.21	14.46	4.42	95.08
OI	9.22	10.02	11.46	11.77	14.19	16.93	20.21	23.93	28.68	36.21	37.92	13.72	4.11	20.05
TI	47.51	57.20	57.24	66.10	89.76	110.32	126.56	144.52	169.25	190.79	207.13	14.32	4.36	115.13
IE	26.15	31.51	28.16	32.04	48.65	59.97	66.64	74.19	84.60	90.81	95.80	12.53	3.66	58.05
EE	3.64	4.57	4.96	6.75	8.14	9.47	10.62	12.35	14.61	16.62	17.91	15.58	4.92	9.97
RNP	5.27	6.10	7.49	10.05	12.87	16.41	19.25	22.22	23.60	23.36	23.12	14.38	4.38	15.43
SHF	50.42	56.25	67.76	78.89	90.18	109.35	125.53	151.97	181.56	211.20	244.05	15.42	4.84	124.29
Dep	358.20	402.36	451.31	562.02	658.12	782.68	895.99	1033.39	1230.40	1448.45	1701.10	15.22	4.75	865.82
Bor	50.35	79.10	85.65	106.93	148.39	172.49	186.04	221.73	278.39	254.22	370.65	19.90	7.36	177.63
Inv	151.43	167.74	195.53	238.92	298.38	353.69	367.10	376.77	444.00	462.06	547.16	12.39	3.61	327.52
$Advs$	278.50	319.28	350.48	447.79	542.35	642.98	757.37	895.31	1091.73	1251.80	1502.20	16.56	5.39	734.53
TA	506.06	567.21	636.74	787.72	954.02	1121.74	1274.50	1471.38	1766.12	2008.44	2401.21	15.21	4.74	1226.83
ROA	0.96	0.94	0.85	0.99	1.03	1.09	0.98	0.94	0.77	0.78	0.60	-4.10	63.07	0.90
ROE	11.82	11.70	10.17	11.98	12.49	13.42	10.63	9.02	6.79	7.70	5.13	-7.30	43.42	10.08
EQA	8.80	8.80	8.79	9.64	8.24	8.19	8.25	8.63	8.51	8.85	8.90	0.11	101.21	8.69

Interest Income (II), Other Income (OI), Total Income (TI), Interest Expended (IE), Employee Expenses (EE), Other exp, Reported Net Profit (RNP), ShareHolders Fund (SHF), Deposits (Dep), Borrowings (Bor), Investments (Inv), Advances (Advs), Total Assets (TA), Return on Assets (ROA %), Return on Equity (ROE %), EQ to Total Assets (EQA %), Trend in % CARG (%) and Average (Avg).

Table A4: Period 2008 - 2013 Descriptive Statistics: Variables of Interest

Year	Stats	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
2008	Mean	20,833	8.51	9.42	15.92	804.06	46.71	258.67	17.02	551.03	67.55	11.93
	SD	30,104	13.10	11.04	23.70	1,060.10	62.26	346.56	22.28	753.50	91.05	19.29
	Min	544	0.17	0.16	0.30	11.01	0.52	3.61	0.24	5.86	1.07	0.09
	Max	179,205	77.86	41.09	126.09	5,891.31	319.29	1,895.01	119.44	4,167.68	489.50	88.11
2009	Mean	21,893	10.46	10.76	18.82	1,054.36	60.04	320.04	20.18	685.54	86.29	14.77
	SD	33,989	16.23	11.15	27.65	1,429.98	77.28	465.93	27.44	945.02	113.41	23.51
	Min	566	0.21	0.18	0.33	13.11	0.74	4.04	0.32	8.01	1.38	0.16
	Max	205,896	97.47	38.38	156.49	8,261.31	429.15	2,759.54	155.74	5,425.03	637.88	126.91
2010	Mean	22,492	12.28	11.00	21.68	1,209.51	63.40	383.24	23.05	808.11	93.91	17.09
	SD	34,899	20.75	11.28	34.00	1,565.42	81.49	508.28	30.42	1,081.61	121.93	26.72
	Min	707	0.23	0.22	0.39	15.89	0.85	5.07	0.33	11.70	1.44	0.13
	Max	212,277	127.55	44.13	203.19	9,071.28	473.22	2,957.85	177.36	6,319.14	709.94	149.68
2011	Mean	22,954	16.43	12.29	27.12	1,466.31	70.39	434.77	27.61	1,007.43	113.17	16.99
	SD	36,315	25.19	12.77	38.99	1,839.10	84.90	522.97	34.49	1,306.87	140.26	27.74
	Min	907	0.72	0.43	0.94	20.50	0.94	8.92	0.44	19.05	1.89	0.19
	Max	222,933	152.12	47.64	230.15	10,535.02	488.68	2,956.01	198.26	7,567.19	813.94	158.25
2012	Mean	24,867	17.66	12.93	30.54	1,696.01	101.49	502.47	34.95	1,188.13	151.62	18.70
	SD	36,094	27.83	13.31	44.47	2,067.34	112.68	575.24	42.25	1,508.30	183.82	27.09
	Min	1,328	0.84	0.59	1.39	59.38	2.78	23.34	1.09	41.32	4.65	0.67
	Max	215,481	169.74	54.67	260.69	11,706.53	632.30	3,121.98	239.49	8,675.79	1,065.22	143.51
2013	Mean	26,082	19.83	14.29	34.92	1,967.94	120.70	589.01	40.91	1,376.92	176.19	21.54
	SD	38,034	30.33	14.76	50.33	2,396.16	132.52	646.01	48.23	1,784.24	207.53	30.34
	Min	1,859	1.25	0.94	2.27	98.89	6.22	33.01	1.96	63.76	8.79	0.95
	Max	228,296	183.81	70.05	292.84	13,719.22	753.26	3,508.78	271.99	10,456.17	1,196.55	160.37

Note. (1) Number of Employees; (2) Employee Expenses; (3) Fixed Assets; (4) Operating Expenses (adjusting employee expenses); (5) Loanable Funds; (6) Interest Paid on Funds; (7) Investments; (8) Income from Investments; (9) Advances; (10) Interest Income from Advances; (11) Non-Interest Income. Price of the Non-Interest Income is equal to unity. All figures are in Rs. Billions.

Table A5: Period 2014 - 2018 Descriptive Statistics: Variables of Interest

Year	Stats	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
2014	Mean	27,433	22.97	17.66	40.92	2,260.58	135.87	651.06	46.98	1,587.88	198.20	25.09
	SD	37,431	36.82	18.99	60.38	2,780.87	151.10	713.46	55.59	2,084.78	234.94	35.39
	Min	2,430	1.57	1.34	3.19	111.85	7.60	36.34	2.45	79.36	11.28	0.73
	Max	222,033	225.04	80.02	357.26	15,775.39	870.69	3,988.00	319.42	12,098.29	1,363.51	185.53
2015	Mean	28,191	25.22	18.83	45.82	2,492.72	150.44	676.57	49.73	1,743.14	218.78	29.27
	SD	36,505	38.87	20.37	65.15	3,130.66	169.30	831.22	60.46	2,257.67	263.06	42.18
	Min	2,279	1.96	1.64	3.58	133.23	9.14	39.62	2.62	76.70	12.84	0.85
	Max	213,242	235.37	93.29	380.54	17,819.44	973.82	4,817.59	353.54	13,000.26	1,523.97	225.76
2016	Mean	29,719	27.78	26.80	51.82	2,711.14	157.91	750.54	55.20	1,858.76	231.18	32.91
	SD	36,762	41.06	26.89	71.73	3,532.08	183.49	969.48	71.69	2,499.34	282.91	51.77
	Min	2,185	2.13	1.77	3.81	116.06	9.00	37.92	3.08	69.53	12.04	0.77
	Max	210,985	251.14	103.89	417.82	20,540.67	1,068.04	5,756.52	423.04	14,637.00	1,639.98	278.45
2017	Mean	30,490	29.30	36.15	56.83	2,930.38	157.63	832.55	59.13	1,944.05	235.39	44.22
	SD	36,851	42.92	69.56	79.90	4,009.91	192.45	1,255.41	80.46	2,690.32	301.30	65.46
	Min	2,021	1.95	2.14	3.49	114.39	7.57	41.94	2.77	64.46	10.89	1.11
	Max	209,580	264.89	429.19	464.73	23,624.45	1,136.59	7,659.90	482.05	15,710.78	1,755.18	354.61
2018	Mean	32,312	32.16	35.56	64.97	3,324.03	162.65	979.48	65.21	2,171.34	249.48	47.16
	SD	44,587	53.85	65.13	101.32	5,102.40	238.93	1,714.50	113.52	3,261.33	369.21	77.51
	Min	1,884	1.64	2.04	3.02	113.12	6.68	41.14	2.92	61.10	10.13	1.02
	Max	264,041	331.79	399.92	599.43	30,684.85	1,456.46	10,609.87	703.38	19,348.80	2,204.99	446.01

Note. (1) Number of Employees; (2) Employee Expenses; (3) Fixed Assets; (4) Operating Expenses (adjusting employee expenses); (5) Loanable Funds; (6) Interest Paid on Funds; (7) Investments; (8) Income from Investments; (9) Advances; (10) Interest Income from Advances; (11) Non-Interest Income. Price of the Non-Interest Income is equal to unity. All figures are in Rs. Billions.

Table A6: Bootstrap Regressions

	NTE^1	NCE^1	NPE^1	NTE^2	NCE^2	NPE^2
BoardSize^G	-0.0138	0.0134	-0.1213	0.0065	0.0266	-0.1910
	(0.0016)	(0.0024)	(0.0063)	(0.0687)	(0.0434)	(0.0784)
Ind^G	-0.1244	-0.0950	-0.2544	0.0012	-0.0003	-0.0319
	(0.0003)	(0.0003)	(0.0009)	(0.0101)	(0.0046)	(0.0092)
BoardMeet^G	-0.0139	-0.1507	-0.0840	-0.0369	-0.0452	-0.1030
	(0.0009)	(0.0009)	(0.0027)	(0.0263)	(0.0146)	(0.0328)
GenDivBoard^G	0.0161	0.0663	-0.0580	-0.0062	0.0109	-0.0140
	(0.0005)	(0.0006)	(0.0019)	(0.0167)	(0.0106)	(0.0169)
Duality^G	0.0813	0.1143	0.1086	0.2060	0.1760	0.6460
	(0.0055)	(0.0085)	(0.0317)	(0.2930)	(0.1780)	(0.3410)
ROA^{BC}	0.2824	0.2576	0.3475	0.4220	0.2640	0.7820
	(0.0071)	(0.0076)	(0.0240)	(0.1700)	(0.1140)	(0.2630)
EQA^{BC}	0.0135	0.0374	0.1286	-0.0495	-0.0177	0.2220
	(0.0021)	(0.0032)	(0.0062)	(0.0813)	(0.0589)	(0.0888)
Size^{BC}	0.1724	0.2171	0.3344	0.5380	0.3180	1.0520
	(0.0037)	(0.0050)	(0.0154)	(0.1980)	(0.0998)	(0.2390)
Owner^{OC}	-0.0228	0.0345	0.1934	0.216	0.3060	2.8830
	(0.0145)	(0.0172)	(0.0551)	(0.5910)	(0.3160)	(0.8410)
Year^{OC}	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.1910	0.2275	0.3223	0.0446	0.0284	0.1557
Adj. R^2	0.1518	0.1900	0.2895			
Wald χ^2	49.06	128.6	174.92	178.23	200.96	101.41
MSE	0.0031	0.0048	0.0441	7.4716	1.0407	27.5436
RMSE	0.0555	0.0692	0.2101	2.73343	1.02014	5.2482
N	412	412	412	412	412	412

N - Number of observations, 1 - OLS Bootstrap, 2 - Fractional Logit Bootstrap, G - Governance, BC - Bank Characteristics, and OC - Other Characteristics. Note: R^2 reported for Fractional Logit Bootstrap is pseudo R^2.

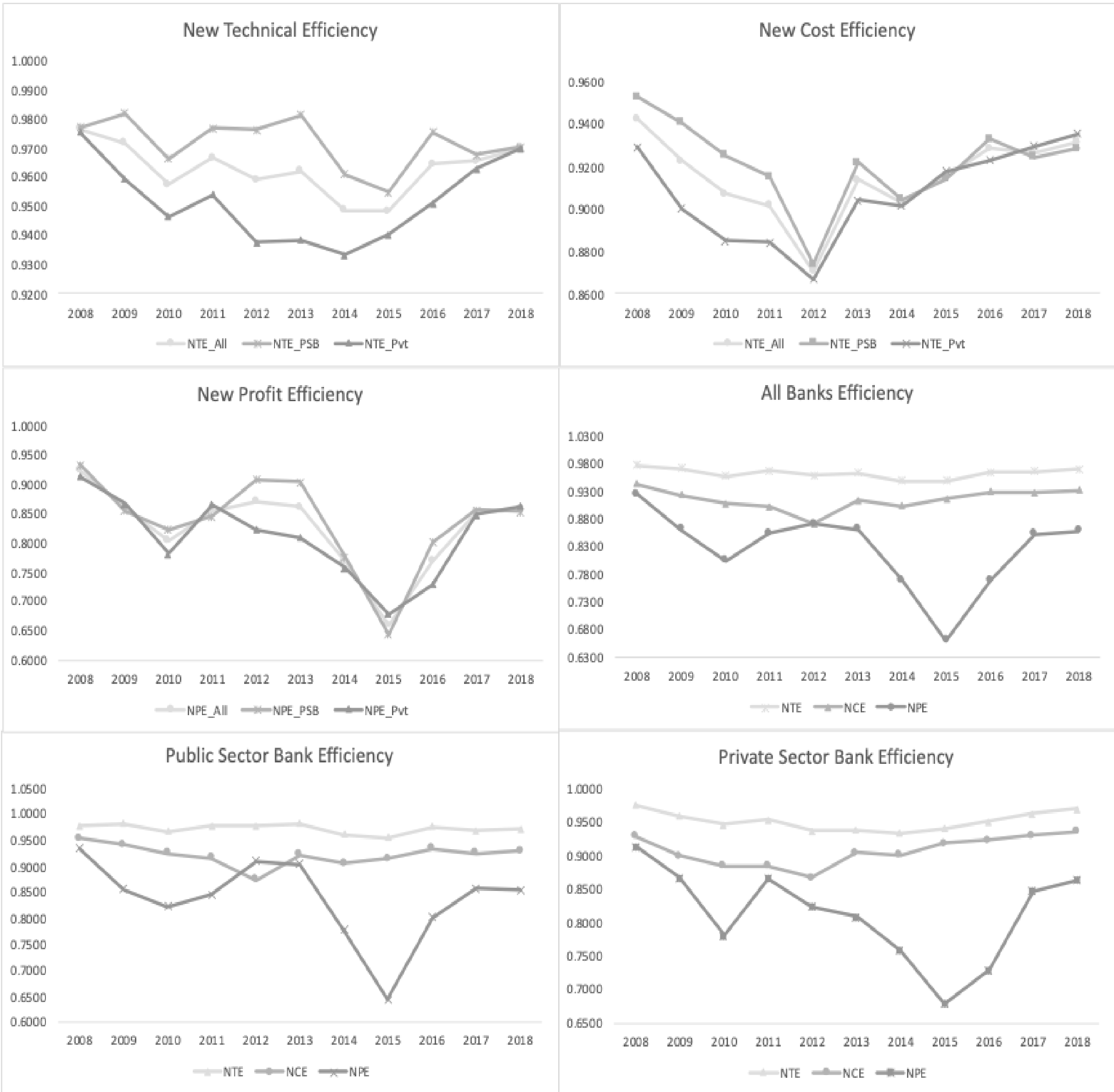
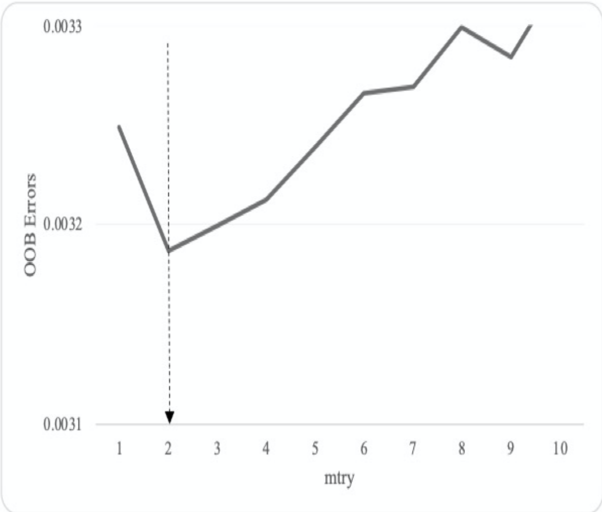
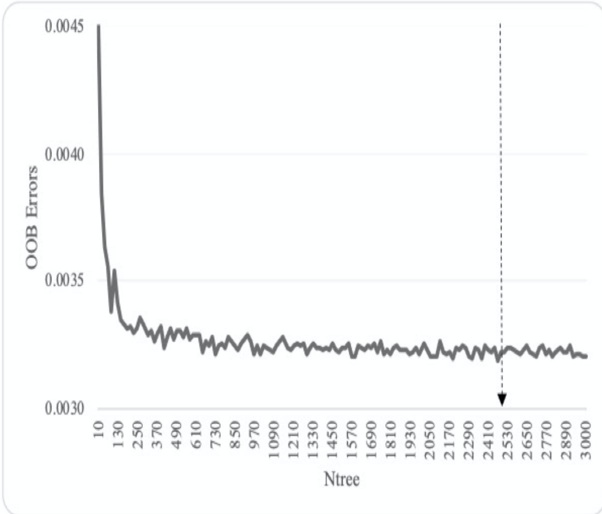


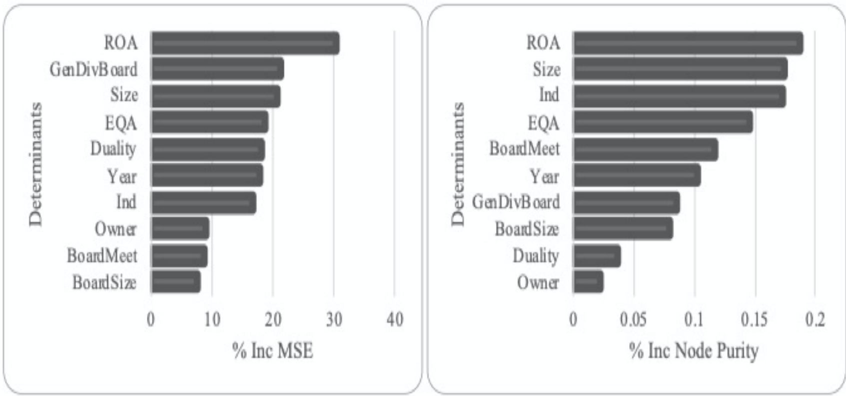
Figure .2: **Average efficiency: 2008 - 2018**



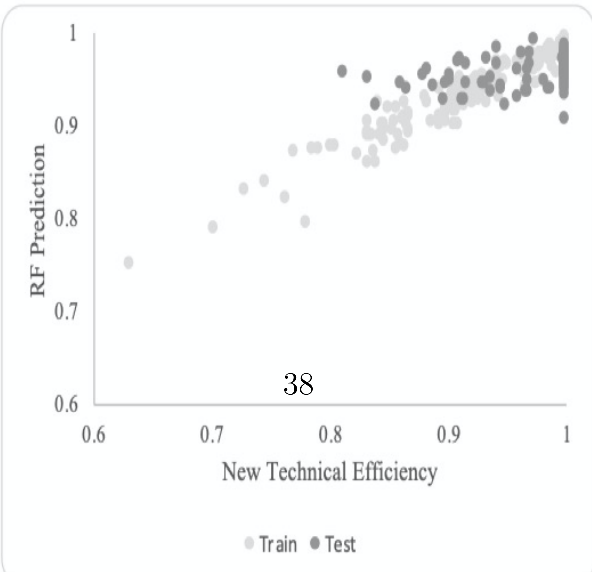
(a) NTE: Evolution of OOB Vs *mtry*



(b) NTE: Evolution of OOB Vs N_{tree}

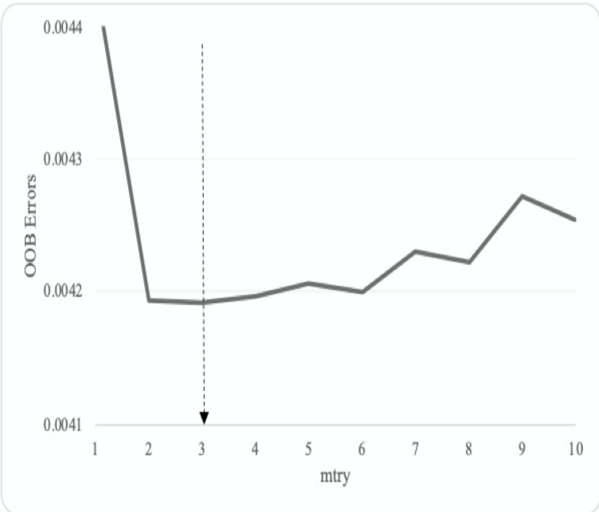


(c) NTE: Reduction of MSE and node purity

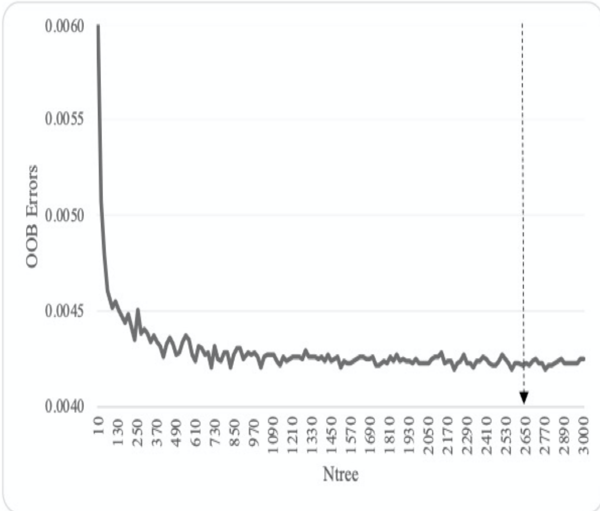


(d) NTE: Scatter Plot

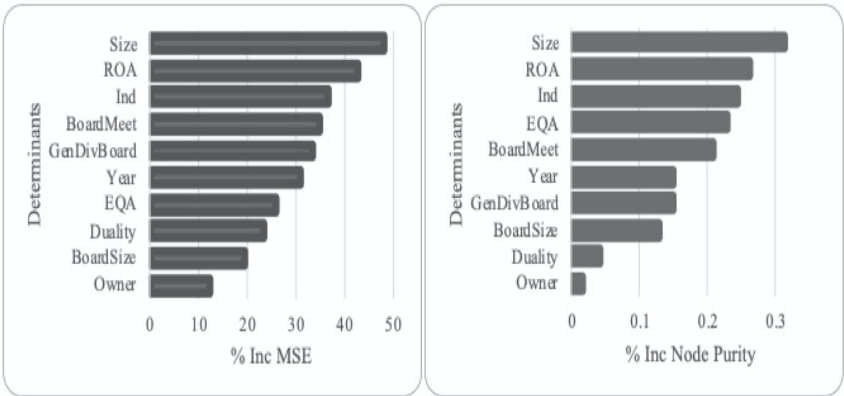
Figure .3: NTE: Random forest outputs



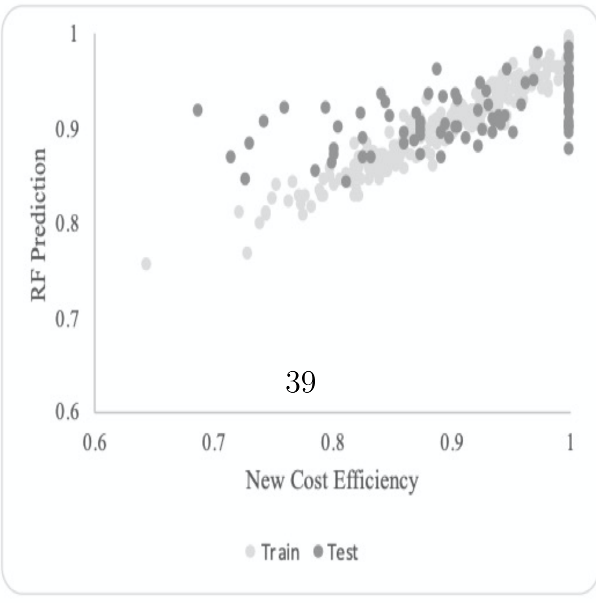
(a) NCE: Evolution of OOB Vs $mtry$



(b) NCE: Evolution of OOB Vs N_{tree}

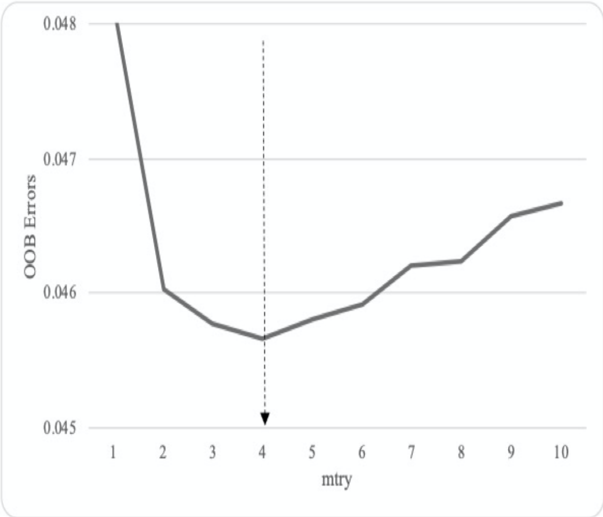


(c) NCE: Reduction of MSE and node purity

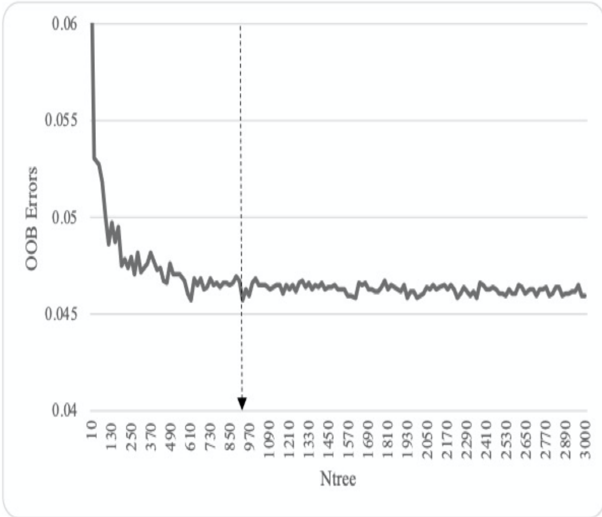


(d) NCE: Scatter Plot

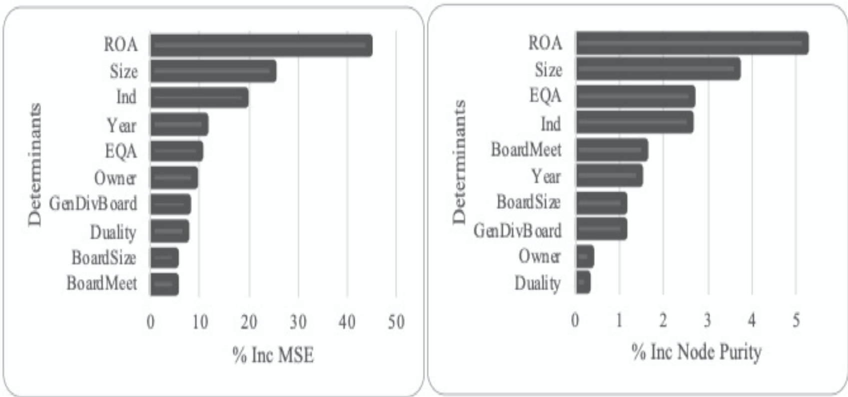
Figure .4: NCE: Random forest outputs



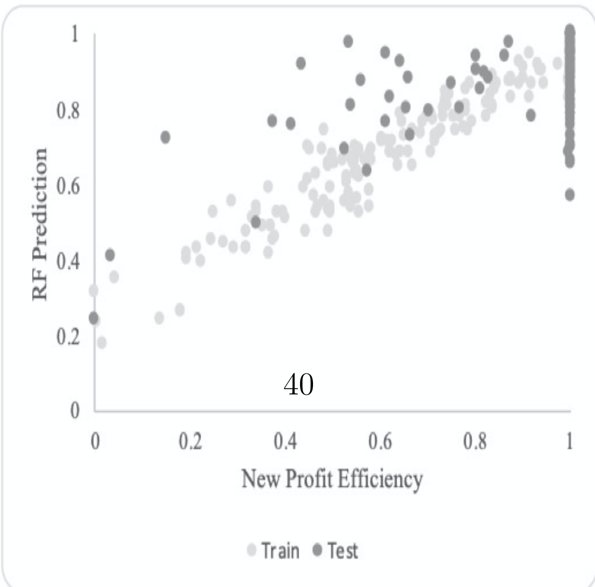
(a) NPE: Evolution of OOB Vs $mtry$



(b) NPE: Evolution of OOB Vs N_{tree}



(c) NPE: Reduction of MSE and node purity



(d) NPE: Scatter Plot

Figure .5: NPE: Random forest outputs